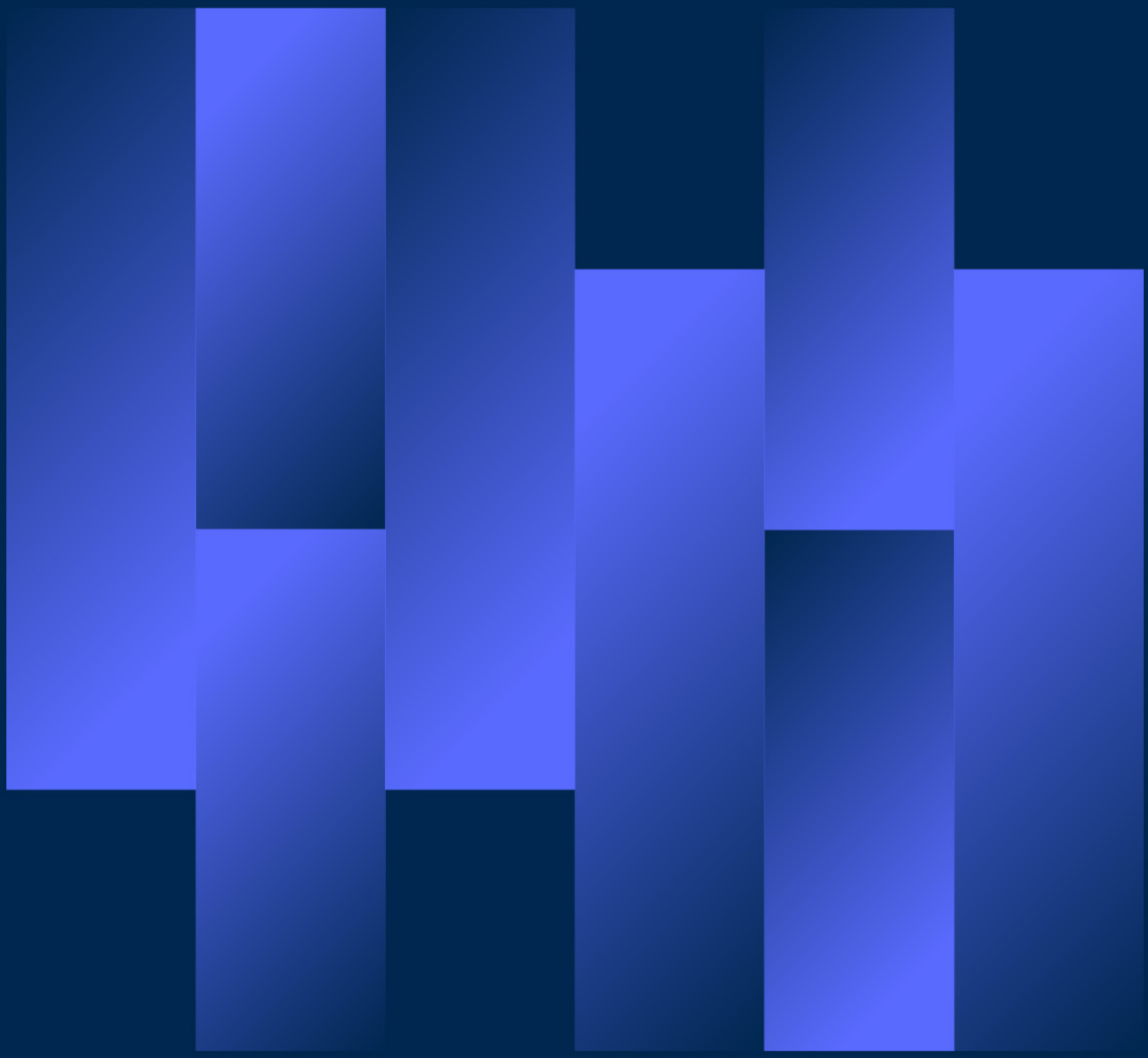
Superintelligence Series

Intelligent Visual Arts and Audiovisual Engineering: Generative AI, Computational Creativity and Superintelligence



South American Publishing

Superintelligence Series

Volume 5

Intelligent Visual Arts and Audiovisual Engineering: Generative AI, Computational Creativity and Superintelligence

Edited by

Emanuel Maldonado

South American Publishing

Copyright Page

© 2026. The authors. This is an open access book, distributed under the terms of a Creative Commons Attribution 4.0 license (https://creativecommons.org/licenses/by/4.0/) that allows use, distribution, and reproduction in any medium provided that the original work is properly cited.

This South American Publishing imprint is published by South American Publishing.

For more information, see South American Publishing's Open Access Policy: https://southam.pub/policies.php

Series: SAP Superintelligence Series
ISSN: 3144-5624
Title of the volume: Intelligent Visual Arts and Audiovisual Engineering: Generative AI, Computational Creativity and Superintelligence
ISBN (complete work): 978-9915-9851-0-7
ISBN (this volume, eBook): 978-9915-9966-5-3
ISBN (this volume, Softcover): 978-9915-9966-4-6

This title is available in both print (softcover) and digital (open access PDF) formats.

DOI: 10.62486/978-9915-9966-5-3

Cataloging Data

Cataloging-in-Publication Data (CIP):

South American Publishing

Superintelligence Series Volume 5: Intelligent Visual Arts and Audiovisual Engineering: Generative AI, Computational Creativity and Superintelligence / edited by Emanuel Maldonado. Montevideo: South American Publishing, 2026.

ISBN (English edition, eBook): 978-9915-9966-5-3

ISBN (English edition, Softcover): 978-9915-9966-4-6

THEMA classification codes:

UYQP

BISAC classification:

COM100000 - COMPUTERS / Artificial Intelligence / Generative AI

ART046000 - ART / Digital

COM016000 - COMPUTERS / Artificial Intelligence / Computer Vision & Pattern Recognition

TEC043000 - TECHNOLOGY & ENGINEERING / Television & Video

Editorial Notice and Acknowledgments

The publication of this book is part of South American Publishing's mission to promote open, ethical, and rigorous scientific communication across all areas of knowledge.

All books published by South American Publishing undergo a double-blind peer-review process and a technical evaluation in accordance with the publisher's editorial policies, aligned with COPE guidelines.

South American Publishing acknowledges the valuable collaboration of authors, reviewers, designers, and production teams who made this publication possible.

Preface

An image can now begin with a sentence. A moving scene can be assembled from generated frames, a three-dimensional environment reconstructed from photographs, and a recorded voice adapted to another language. These possibilities have altered the working conditions of visual creation. They have also made familiar questions more difficult. Who decides whether an image is convincing? What does it mean for a model to be open? How should a filmmaker weigh technical assistance against control over the finished work? And what obligations arise when synthetic images enter public collections or circulate without their origins being clear?

Such questions belong as much to the humanities and social sciences as they do to computer science. The aesthetic value of an image cannot be read directly from a benchmark score. A license expresses rights and restrictions that may be obscured by the broad language of openness. A generated portrait may display technical precision while reproducing a narrow view of who belongs in a particular profession. Even the apparently straightforward task of identifying synthetic content becomes uncertain when a detector encounters a model different from those on which it was trained. Technical capability and human judgment meet at each of these points, though they do not always agree.

The chapters brought together here examine this relationship through different forms of evidence. Some study model registries, image databases and public repositories; others compare measurements of visual quality with human assessments or test the transfer of forensic methods across generators. The reviews turn to the evaluation of generated and restored video, the reconstruction of scenes through neural radiance fields and 3D Gaussian splatting, and the presence of machine learning in editing, color grading, sound, dubbing and lip synchronization. Accounts of social bias, authorship and copyright bring the discussion back to the people and institutions affected by synthetic media.

The value of reading these subjects together lies in the questions they share. A technically plausible result may be perceptually weak. A method that performs well on one dataset may behave differently elsewhere. Tools that reduce repetitive work can leave the central artistic decision unresolved. The conditions under which a model is distributed can be as consequential as the images it produces. These are not merely temporary inconveniences on the way to more capable systems. They concern the ways visual work is assessed, attributed, maintained and trusted.

Readers will encounter different levels of certainty across the chapters. Empirical findings depend on their samples, dates and measures; reviews reflect the scope of the literature they examine. Legal treatment varies by jurisdiction, and professional practices differ between production settings. The book treats those limits as part of the subject, because claims about creative automation become more useful when the conditions under which they hold are visible.

The pages that follow are intended for researchers, artists, audiovisual practitioners and readers interested in the institutions surrounding digital culture. They invite a careful view of computational creativity: one attentive to what machines can produce, to what their outputs leave unanswered, and to the human choices that still give visual work its purpose.

Introduction

Visual arts and audiovisual engineering are increasingly shaped by systems that learn statistical relationships among text, images, sound and movement. A prompt can guide the synthesis of a picture or sequence; a model can recover details from degraded footage or propose a new viewpoint within a reconstructed scene. In post-production, computational methods can organize recordings, assist with color decisions, separate sound sources and support dubbing. These developments connect image making to questions of infrastructure, measurement and responsibility that once seemed peripheral to creative practice. Understanding them requires attention to how visual systems are built, how their results are judged and how those results enter social life.

The first issue is access. Generative models are often described as open because their files or documentation can be found in a public repository. Yet availability does not by itself determine what others may do with a model. Licenses can restrict use, access may be gated, and the information needed to assess a system can vary in completeness. The chapter examining text-to-image and text-to-video models on the Hugging Face Hub approaches openness through observable licensing and documentation indicators. Its contribution is to make a broad claim about openness susceptible to comparison over time. For researchers and creators, these distinctions bear directly on reproducibility, redistribution and the conditions of independent scrutiny.

Once an image is generated, a different problem arises: how should its quality be measured? Traditional image-quality metrics were developed largely around properties of natural images and their distortions. Generated images may fail in ways these measures do not capture, including weak correspondence to a prompt or a visually plausible composition whose details make little sense. The comparison of BRISQUE, NIQE and PIQE with human judgments illustrates both the appeal and the limits of inexpensive measures that require no reference image. The systematic review of video-generation evaluation extends the concern across time, where individual frames may appear convincing while movement, continuity or narrative coherence remains unsatisfactory. Metrics are necessary for comparison, but their meaning depends on the quality of the human judgment they are expected to approximate.

Questions of trust run alongside questions of quality. Synthetic images can be difficult to distinguish from photographs, and forensic methods seek patterns that a human observer may not see. Frequency-domain detectors offer one way to identify traces of generation. Their performance across unfamiliar generators, however, shows why a strong result under controlled conditions cannot be assumed to hold across the wider field. The study of machine-generated images on Wikimedia Commons places a related issue in a public setting. Uploading an image to a shared repository does not ensure its meaningful reuse, and counting files is not the same as measuring their value to a community. Provenance, moderation and actual use must therefore be considered separately.

The social content of generated images demands another form of assessment. Models trained on large collections of images can carry familiar stereotypes into new visual material. The review of audit studies on text-to-image generation examines how representations of gender, race and culture have been measured, and why documenting a bias is easier than demonstrating that an intervention has reduced it. This is an artistic concern as well as an ethical one. When the available image of a scientist, a caregiver or a professional authority repeatedly follows a narrow pattern, generative systems can constrain the imaginative possibilities that they appear to expand.

Authorship presents a further tension. A generated work may involve the selection of data, the writing and revision of prompts, the curation of outputs, and extensive human editing. The philosophical account of creativity may recognize these contributions as distributed across people and tools.

Copyright systems, by contrast, must apply legal criteria to particular claims of authorship and protection. The chapter on machine-generated visual art traces how these questions have developed and why the threshold of human contribution and the treatment of training material remain unsettled. These issues cannot be answered solely by examining what a model produces. They require a view of the process by which the work came into being.

Audiovisual engineering brings the debate into the production environment. Neural radiance fields and 3D Gaussian splatting make it possible to represent and render scenes through markedly different computational approaches. Their respective trade-offs in fidelity, speed, editability and temporal stability show that the preferred method depends on the work to be done. A related lesson emerges from the review of video super-resolution and restoration: favorable benchmark scores reveal only part of a method's suitability when computational cost, data diversity and actual viewing conditions are inconsistently reported. In post-production, machine learning is already present across editing, color grading, sound, dubbing and lip synchronization, but the need for continuity, directability and artistic judgment remains. A workflow can become less laborious without becoming less dependent on human intention.

Taken together, the chapters connect four questions: what systems make possible, how their performance can be established, whose interests their outputs affect, and who remains accountable for the result. They also reveal a recurring mismatch between what is easy to count and what matters in practice. A model may be downloadable yet legally restricted; a score may improve without matching perception; an archive may grow without widespread reuse; a generated sequence may look persuasive frame by frame but fail in motion. These distinctions are essential to the interpretation of evidence in a field where demonstrations are often more visible than their limitations.

The prospect of increasingly autonomous artificial intelligence makes such distinctions more pressing, not less. The material examined here concerns existing methods and the evidence available to assess them. It does not treat speculative capabilities as established outcomes. Its purpose is to situate generative visual systems within the wider work of creation, evaluation and public accountability. The reader is invited to move between computation and culture without assuming that either perspective can explain the whole of contemporary visual production.

Abstract

Visual production is being reshaped by systems that can synthesize images, generate video, reconstruct three-dimensional environments and assist with work after filming. These capabilities raise questions that technical performance alone cannot settle: what counts as an open model, how reliably synthetic media can be evaluated or identified, whose experiences are represented in generated imagery, and where creative and legal responsibility resides. Bringing together empirical investigations and critical reviews, the volume examines text-to-image and text-to-video models, the conditions under which they are shared, and the methods used to judge their perceptual and semantic quality. It considers the limited transfer of frequency-based detection across generators and the uneven relationship between growth in generated images in public repositories and their actual reuse. Other chapters address evaluation practices for video generation and restoration, neural radiance fields and 3D Gaussian splatting, and machine learning in editing, color grading, sound, dubbing and lip synchronization. Questions of social bias, authorship, copyright and training data connect these technical subjects to their cultural and institutional consequences. Across these areas, a shared difficulty emerges: computational measures and automated workflows can support audiovisual practice without settling questions of meaning, fairness, provenance or artistic intention. The book approaches generative visual systems as parts of a larger creative process shaped by datasets, infrastructures, professional judgment and public norms. It offers a grounded account of current methods and their limitations while clarifying the responsibilities attached to the production, evaluation and circulation of synthetic media.

Keywords: Generative artificial intelligence; Computational creativity; Visual arts; Audiovisual engineering; Synthetic media evaluation; AI ethics and copyright

Index

Chapter 01. Authorship, copyright and creativity in machine-generated visual art | Autoría, derechos de autor y creatividad en el arte visual generado por máquinas 1

Chapter 02. Openness and licensing of text-to-image and text-to-video models on the Hugging Face Hub | Apertura y licencias de los modelos de texto a imagen y texto a video en Hugging Face Hub12

Chapter 03. Classical no-reference image quality metrics and human judgments of artificially generated images | Métricas clásicas de calidad de imagen sin referencia y juicios humanos de imágenes generadas artificialmente 24

Chapter 04. Frequency-domain detectors for machine-generated images and their generalization across generators | Detectores en el dominio de la frecuencia para imágenes generadas por máquina y su generalización entre generadores 36

Chapter 05. Growth, use and deletion of machine-generated images on Wikimedia Commons | Crecimiento, uso y marcadores de borrado de imágenes generadas por máquina en Wikimedia Commons 47

Chapter 06. A systematic review of metrics and human evaluation practices for video generation models | Una revisión sistemática de métricas y prácticas de evaluación humana para modelos de generación de video 58

Chapter 07. A systematic review of benchmark performance in deep learning for video super-resolution and restoration | Una revisión sistemática del rendimiento en benchmarks del aprendizaje profundo para super-resolución y restauración de video 71

Chapter 08. A systematic review of audit studies on social bias in text-to-image generation | Una revisión sistemática de los estudios de auditoría sobre el sesgo social en la generación de texto a imagen 84

Chapter 09. Neural radiance fields and 3D Gaussian splatting in audiovisual production | Campos de radiancia neuronal y salpicado gaussiano 3D en la producción audiovisual 96

Chapter 10. Machine learning in post-production editing, color grading, sound, dubbing and lip synchronization | Aprendizaje automático en la posproducción de edición, etalonaje, sonido, doblaje y sincronización labial 108

Chapter 1 / Capítulo 1

SAP Superintelligence Series
ISSN: 3144-5624
ISBN (complete work): 978-9915-9851-0-7

Superintelligence Series Volume 5: Intelligent Visual Arts and Audiovisual Engineering: Generative AI, Computational Creativity and Superintelligence
ISBN (this volume): 978-9915-9966-5-3
DOI: 10.62486/978-9915-9966-3-9.Ch01
Pages: 1-11

Check for updates

© 2026 The authors. This is an open access book chapter distributed under the terms of the Creative Commons Attribution (CC BY) 4.0 License.

Authorship, copyright and creativity in machine-generated visual art

Autoría, derechos de autor y creatividad en el arte visual generado por máquinas

Carlos Alberto Gómez Cano[1]
Verenice Sánchez Castillo[2]

1 Corporación Universitaria de Asturias. Bogotá, Colombia.
2 Universidad de la Amazonia. Florencia, Colombia

ABSTRACT

The question of who authors a work produced by a machine, and whether copyright protects it, has moved from a theoretical curiosity to a central problem of intellectual property in less than a decade. This narrative review traces how legal and philosophical treatments of authorship, copyright and creativity in AI-generated art evolved from 2018 to 2026, a window that can be divided into four periods by technical and institutional inflection points. The GAN era and the first auction sales opened the question, the diffusion models of 2021 and 2022 made it a mass concern, the institutional decisions of 2023 established a human-authorship requirement, and the period from 2024 to 2026 consolidated that requirement alongside new regulation and training-data litigation. The review synthesizes the legal and philosophical literature across these periods and finds that the law has converged on a stable human-authorship standard while the philosophy has moved toward distributed and collaborative accounts of creativity, leaving a durable gap between the two that the field has not yet closed. The trajectory is interpreted as a shift in what the debate is about, from whether machines can be authors to how much human involvement suffices and who bears responsibility for what a model was trained on.

Keywords: AI-generated art; copyright; authorship; creativity; generative artificial intelligence; intellectual property

RESUMEN

La cuestión de quién es autor de una obra producida por una máquina, y si el derecho de autor la protege, ha pasado de ser una curiosidad teórica a un problema central de la propiedad intelectual en menos de una década. Esta revisión narrativa traza cómo evolucionaron los tratamientos jurídicos y filosóficos de la autoría, el derecho de autor y la creatividad en el arte generado por IA entre 2018 y 2026, una ventana que puede dividirse en cuatro períodos por puntos de inflexión técnicos e institucionales. La era de las GAN y las primeras subastas abrieron la cuestión, los modelos de difusión de 2021 y 2022 la convirtieron en una preocupación masiva, las decisiones institucionales de 2023 establecieron un requisito de autoría humana, y el período de 2024 a 2026 consolidó ese requisito junto a nueva regulación y litigios sobre datos de entrenamiento. La revisión sintetiza la bibliografía jurídica y filosófica de esos períodos y concluye que el derecho ha convergido en un estándar estable de autoría humana mientras que la filosofía se ha desplazado hacia relatos distribuidos y colaborativos de la creatividad, dejando una brecha duradera entre ambos que el campo aún no ha cerrado. La trayectoria se interpreta como un cambio en aquello de lo que trata el debate, de si las máquinas pueden ser autoras a cuánta intervención humana basta y quién responde por aquello con lo que se entrenó un modelo.

Palabras clave: arte generado por IA; derechos de autor; autoría; creatividad; inteligencia artificial generativa; propiedad intelectual

INTRODUCTION

For most of copyright's history, the attribution of authorship to a machine was a hypothetical that the law could comfortably set aside. The premise that a work must originate from a human being was embedded so deeply in the statutory and doctrinal structure of copyright that it rarely needed to be defended, and the occasional computer-generated output was treated as the product of the programmer or user who controlled the tool.[1,2] That premise became a live problem when generative systems began to produce images that were, to a viewer, indistinguishable from the work of a human artist, and the problem arrived with unusual speed. The period from 2018 to 2026 is the span in which the question moved from the margins of legal scholarship to the center of intellectual-property policy, and it is the span this review reconstructs.

The trigger was not a single event but a succession of them, each of which redefined what the debate was about and each of which arrived before the previous one had been fully digested. In 2018 an algorithmic portrait sold at auction for a price that forced the art world and the legal world to confront the possibility that a machine, or the people who built and prompted it, might be credited as the author of a work with commercial value.[3,4] Between 2021 and 2022 the release of text-to-image diffusion models made the production of images a matter of typing a sentence, which transformed the question from one about a small class of experimental works to one about a mass medium.[5,6] In 2023 courts and registration offices began to rule, and their rulings established a human-authorship requirement that the literature then spent the following years consolidating, extending and contesting.[7,8] By 2026 the debate had acquired a second center of gravity, the copyright status of the training data on which generative models are built, and a regulatory frame that did not exist when the window opened.(9,10)

The problem is at once legal and philosophical, and the two registers do not move in lockstep, a fact that is itself part of what the review has to explain rather than merely describe. The legal question is comparatively settled in its broad outline, namely that a work without sufficient human authorship is not protected, but it is unsettled at the boundary, where a human selects, arranges, edits or otherwise shapes machine output.[11,12] The philosophical question is more open, because it concerns what creativity is and whether the conditions that make human creativity valuable, intention, autonomy, a life to which the work is answerable, can be satisfied by a system, or by a system and its human users together.[13,14] A review that treated only the law would miss the philosophical movement beneath it, and a review that treated only the philosophy would miss the institutional facts that have forced the question, which is why both are synthesized here.

Surveys of the field have multiplied, but they tend to freeze the debate at a single moment, a particular case, a particular jurisdiction or a particular theory of authorship, rather than to follow its movement across time, and a reader of several such surveys is left with a series of stills rather than a film of the debate.[10,15] The chronological structure of this review is a deliberate choice, because the debate is best understood as a trajectory in which each period's questions set the terms of the next, and in which the meaning of a stable doctrine, such as the human-authorship requirement, changed as the technology around it changed.[2,16] What began as a question about a portrait sold at auction became a question about the distribution of authorship across a network of trainers, developers, platform operators and users, and that transformation, from a single object to a distributed system, is the story this review tells.

The objective is to trace the evolution of legal and philosophical treatments of authorship, copyright and creativity in AI-generated art from 2018 to 2026. Four specific objectives organize the account and correspond to the sections that follow: to periodize the debate into identifiable stages marked by technical and institutional inflection points; to synthesize how copyright law has addressed machine authorship, human involvement and training data across the period; to synthesize how philosophical accounts of creativity have been revised in response to generative systems; and to identify what changed, what persisted, and which open questions remain at the end of the window. Each period is developed in turn, and the trajectory is interpreted in the discussion.

METHODS

The review followed the SANRA guidance for narrative reviews, which requires a transparent account of sources, search terms, selection and organization without the formal screening counts of a systematic review. The object of the synthesis is the scholarly literature on authorship, copyright and creativity in AI-generated art, together with the adjacent literature on the philosophy of

https://doi.org/10.62486/978-9915-9966-3-9.Ch01

machine creativity and on the regulation of generative systems that supplies context for the legal claims. The review is chronological by design, and its method is to divide the window into periods whose boundaries are marked by identifiable technical and institutional events, and to hold a fixed set of questions against each period in turn.

Three bibliographic sources were searched: OpenAlex, Crossref and Semantic Scholar. The queries combined terms for artificial intelligence and generative art with terms for copyright, authorship, creativity and originality, and included AI-generated art copyright authorship, artificial intelligence creativity copyright law, generative AI copyright training data litigation, machine authorship copyright office decision, and AI art philosophical creativity autonomy. The search window was 2018 through 2026 for the primary studies, which is the period in which the technical systems and the legal decisions the review covers were produced; earlier works on originality and on the philosophy of creativity were admitted only as background where a primary study explicitly builds on them. Records were deduplicated across sources by DOI and by normalized title, and the query log is retained in the working directory.

Selection followed three criteria applied to title and abstract. A study was retained if it addressed authorship, copyright or creativity in relation to AI-generated art from a legal, philosophical or empirical standpoint; if it was published within the search window; and if it was written in English or Spanish. Works on adjacent topics, such as AI regulation, machine creativity in domains other than art, or the governance of training data, were admitted where they contributed a finding that bears directly on the authorship question. The retained material was divided into a core of primary studies that populate the four periods and a supporting set of surveys and adjacent studies that supply context in the introduction and discussion.

The periodization is the central methodological decision, and it was made against explicit criteria. A boundary was placed where a technical release or an institutional decision changed the question the field was asking, rather than at arbitrary year intervals, because a periodization that simply divided the decade into equal segments would cut across the very developments the review aims to explain, separating a model release from the litigation it triggered or a ruling from the scholarship that followed it. On those criteria the window divides into four periods. The years 2018 to 2020 are the GAN era, bounded at the front by the first high-profile sales of algorithmic art and defined by the first wave of legal and philosophical scholarship on machine authorship. The years 2021 to 2022 are the diffusion period, defined by the release of text-to-image models that made generative art a mass medium. The year 2023 is an institutional turning point, defined by the first authoritative rulings on the human-authorship requirement. The years 2024 to 2026 are a period of consolidation and regulation, defined by the codification of the human-involvement standard and by the rise of training-data litigation and legislation.

A synthesis table (Table 1) sets out the inflection points that define the periods, and a second table (Table 2) summarizes the legal and philosophical positions that characterize each period. A timeline figure (Figure 1) presents the four periods and their defining events.

Table 1. Inflection points that define the four periods.

Period	Defining inflection point	Consequence for the authorship debate
2018-2020	First high-profile auction sale of algorithmic art	Authorship question becomes a commercial and policy concern
2021-2022	Release of text-to-image diffusion models	Generative art becomes a mass medium; the human-involvement boundary widens
2023	Copyright Office and court rulings on machine authorship	Human-authorship requirement established
2024-2026	Codification and training-data litigation	Human-involvement threshold refined; training data becomes central

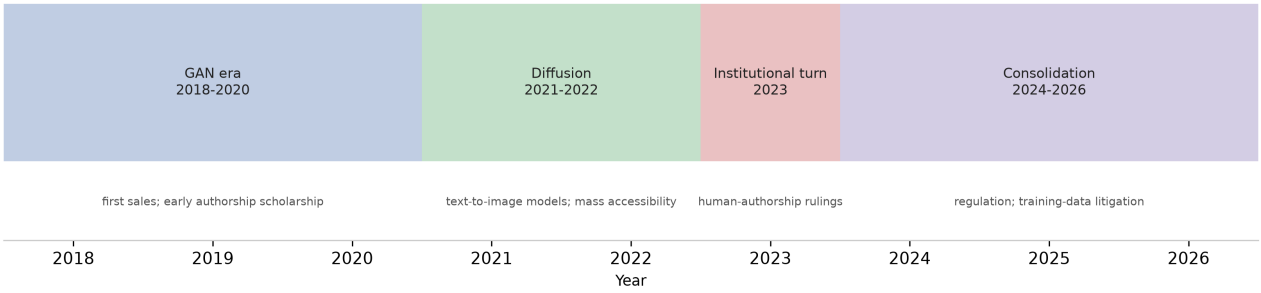
Table 2. Legal and philosophical positions across the four periods.

Period	Legal position	Philosophical position
2018-2020	Machines cannot be authors or legal persons	Authorship and creativity categories under pressure
2021-2022	Tool-author boundary; jurisdictional divergence	Machine agency and collaborative creativity
2023	Human-authorship requirement affirmed	Distributed co-authorship
2024-2026	Human-involvement threshold; training-data regulation	Distributed and posthuman authorship

https://doi.org/10.62486/978-9915-9966-3-9.Ch01

Because the literature is heterogeneous in jurisdiction, method and theoretical orientation, the synthesis is narrative rather than quantitative, and no claim about the distribution of positions is made except where a cited study reports one.

Figure 1. The four periods of the authorship debate and their defining inflection points.



DEVELOPMENT

2018 TO 2020: THE GAN ERA AND THE FIRST AUTHORSHIP CONTROVERSIES

The window opens with a market event that gave the theoretical question a price. In 2018 an algorithmic portrait produced with a generative adversarial network sold at auction, and the sale, together with the surrounding attention, made concrete what had been abstract, that a machine-generated image could carry commercial value and therefore that the question of who owned it, and who authored it, could no longer be deferred.[3,4] The philosophical literature of this period took the sale as evidence that the category of authorship itself was under pressure, and it began to ask whether the concepts of authorship, originality and creativity inherited from romantic and post-romantic aesthetics could survive a system that produces works without a human hand in the traditional sense.(17)

The empirical work of the period complicated the assumption that the public would simply accept machine authorship. Studies of attribution showed that audiences and buyers perceived AI-generated art differently from human art, discounting it in some contexts and crediting less to the machine than the human collaborators, a finding that matters because the social recognition of authorship, not only the legal one, turns out to be contested.(4) The result is that the first period established the problem's shape rather than its solution: the technical capacity existed, the market had noticed it, and neither the law nor the philosophy had yet produced an account that everyone could accept.

Legal scholarship of the first period answered the framing question, whether a machine can be an author or a legal person, largely in the negative. The analyses that traced the doctrine across the common-law jurisdictions concluded that the human-authorship premise, grounded in the constitutional and statutory language of the major systems, left no room for a machine as an author in its own right, and that the more serious question was how to allocate the human contribution among the people who built, trained and directed the system.[1,18] This early work also named, without yet resolving, the second question that would come to dominate the debate, the copyright status of the data on which a model is trained, which the period framed as a question of infringement and exception that would grow in importance as the models grew in scale.

2021 TO 2022: DIFFUSION AND MASS ACCESSIBILITY

The second period is defined by the release of text-to-image diffusion models, which changed the scale and the social meaning of the problem. Where the GAN era had produced images through a technical pipeline that remained the province of researchers, the diffusion models of 2021 and 2022 allowed a user to produce a photorealistic or stylized image from a short natural-language prompt, and the literature of the period describes this as the moment at which generative art became a mass medium rather than a research artifact.[5,6] The consequence for the authorship question was immediate: the range of plausible human contributions, a prompt, a selection among outputs, an edit, a curatorial arrangement, widened, and with it the range of

https://doi.org/10.62486/978-9915-9966-3-9.Ch01

claims about who, among the many people in the chain, was the author.

The legal literature of the period turned to the boundary between the tool and the author. Analyses distinguished human-aided machine output, which the emerging consensus treated as unprotected, from machine-aided human output, which could be protected to the extent of the human contribution, and they began to test that boundary against the specific acts of prompting, selection and post-processing that the new tools made available.[2,19] Jurisdictional work examined how different systems, the United States with its constitutional human-authorship requirement, the United Kingdom with its specific provision for computer-generated works, and the European systems with their originality standards, would classify the same diffusion-produced image, and it found divergence rather than convergence.(20,21)

The philosophical literature of the period matured in a different direction, toward the conditions under which a machine could be an agent or an author at all. Accounts of machine agency tested the requirements of intention, autonomy and answerability against the capabilities of generative systems and concluded that the systems failed the strongest version of each requirement, while leaving open the weaker versions that a collaborative account might satisfy. (22) Alongside this, a line of work began to reframe creativity not as a property that either a human or a machine possesses but as something that emerges from their interaction, a position that would become more prominent in the following periods.(14,16)

2023: THE INSTITUTIONAL TURNING POINT

The year 2023 is the hinge of the review, because it is the year in which the question moved from the journals to the courts and the registration offices, and in which the law gave its first authoritative answers in a form that practitioners could rely on, a transition that the preceding periods had prepared but not yet delivered. The United States Copyright Office declined to register a purely machine-generated image, and the courts affirmed that refusal on the ground that the human-authorship requirement is rooted in the constitutional purpose of copyright, a ruling that the legal literature of the period read as establishing, in the American system at least, that there is no protection without a human author.[7,8] The same year produced the intermediate case, a comic whose images were machine-generated but whose text and selection were human, and the office's decision to protect the human elements while refusing the images gave the field its first worked example of the human-involvement threshold.(8,15)

The institutional decisions did not end the debate but relocated it. With pure machine authorship ruled out, the live questions became how much human involvement suffices, and what kind, and the literature of 2023 immediately began to map those questions onto the specific acts that the new tools made possible, prompting, selection, arrangement, modification and iterative refinement.[8,12] At the same time, the second center of gravity, training data, moved to the foreground, as litigation against model developers raised the question of whether the ingestion of copyrighted works for training is infringement or a permitted use, and as the first analyses of the European and American positions on that question appeared.(15,23)

The philosophical register of 2023 changed tone, from the question of whether a machine could be an author to the question of how humans and machines co-author. A widely cited manifesto for collaboration argued that the productive response to generative systems is neither to reject them nor to cede authorship to them but to reconceive creativity as a distributed practice in which human intention and machine capacity are jointly responsible, and the creativity literature of the period supplied empirical and conceptual support for that move.[16,24] The period thus left the law and the philosophy pointing in different directions, the law toward a strict human threshold and the philosophy toward a distributed account, a divergence that the following years would deepen rather than resolve.

2024 TO 2026: CONSOLIDATION AND REGULATION

The final period is one of consolidation, in which the human-authorship requirement was refined rather than reopened, and in which regulation and training-data litigation supplied the new frame. The registration office codified its position in a report that distinguished, with increasing precision, the human contribution that renders an AI-assisted work protectable from the machine output that does not, and the scholarly literature of the period worked through the report's implications for prompting, selection and modification.[11,12] The dominant question of the

https://doi.org/10.62486/978-9915-9966-3-9.Ch01

period was no longer whether a human is required but how the requirement is operationalized, a shift that the review records as the clearest sign that the debate had matured.

The regulatory turn is the period's second defining feature. European and international instruments introduced transparency and disclosure obligations for generative systems, and the scholarly analyses of text-and-data-mining exceptions and of the legal status of training data became a central genre, asking whether the exceptions that enable research and data mining should be extended to commercial training and how a lawful training regime might be designed in practice.[9,10] Jurisdictional surveys compared the American, European and other approaches to the same questions and found that, while the human-authorship standard had converged, the training-data question had not, and that the two questions were increasingly bound together in the policy debate.(15,25)

The philosophical literature of the period consolidated the distributed turn. Accounts of machine creativity asked directly whether the output of a generative system can be creative in the sense that matters for authorship, and answered with a qualified account in which the creativity is real but is jointly attributable to the system, its trainers and its users rather than to any one of them alone.(13) Posthuman and sociotechnical framings extended this to a full account of authorship as distributed across a network, and the empirical and design literature supplied evidence that human and machine roles in creative work are already being renegotiated in practice.[26,27] The period ends, then, with a stable legal standard and a still-moving philosophical one, and with the training-data question open between them.

Taken together, the four periods describe a single movement with a clear direction. The debate began as a metaphysical question about whether a machine could be an author, and it ended as a practical question about contribution and liability, a movement in which each period's resolution passed the unsettled remainder to the next. What changed was not the underlying doctrine, which always presumed a human author, but the salience of the doctrine and the sophistication of the attempts to operationalize it, and what persisted was the difficulty of drawing the line between the human and the machine in a technology that was deliberately built to blur it. The movement is not complete, since the training-data question opened late and remains open, but its direction is legible, and the discussion that follows interprets it.

DISCUSSION

The trajectory across the four periods resolves into a single, steady observation: the debate changed what it was about, from whether a machine can be an author, which the law answered in the negative in 2023, to how much human involvement suffices, which remains open, and to who bears responsibility for the data on which the machine was trained, which opened later and remains the least settled of the three.[7,9] This reading is consistent with the way the surveys of the field describe the movement, as a shift from a metaphysical question about authorship to a practical question about contribution and liability, and the review's contribution is to date that shift and to show that it was driven by institutional decisions rather than by changes in the technology alone.(10,15)

The law's convergence on a human-authorship standard is the most stable finding of the review, and it deserves to be stated precisely. The United States system anchored the requirement in its constitutional purpose, and the major international analyses found the same premise operating, in different doctrinal forms, across the systems they surveyed, so that a purely machine-generated image is, by the end of the window, unprotected nearly everywhere the question has been decided.(1,8,25) The convergence is nevertheless thinner than it appears, because the systems diverge on the intermediate cases, and the divergence is where the practical action is. A human who prompts, selects and edits a diffusion output is, on the American approach, the author of the human contribution but not of the machine output, while other systems have provisions, such as the United Kingdom's computer-generated-works clause, that reach a different result for the same facts.[19,20] The review therefore treats the human-authorship standard as settled in name and contested in application, which is the honest reading of a literature that agrees on the principle and disagrees on its boundary.

The philosophical movement is the complement and, in part, the correction of the legal one. Where the law drew a line between human and machine contribution, the philosophy of the later periods argued that the line misdescribes how generative systems actually work, in which the human intention, the model's parameters and the training data are all jointly responsible for the output, and no single point in that chain is the author in the traditional sense.[13,24] The

https://doi.org/10.62486/978-9915-9966-3-9.Ch01

empirical work on attribution gave this philosophical claim a social basis, by showing that audiences already perceive AI-made art as distributed rather than as the work of a single author, and the design literature showed that practitioners are building tools on the same assumption.[4,27] The divergence between a law that insists on a single human author and a philosophy that distributes authorship across a network is the deepest unresolved tension the review identifies, and it is unlikely to be resolved by further legal doctrine alone.

The training-data question is where the review's chronological structure is most illuminating, because it shows a problem that was present from the first period but dormant, and that was activated not by any new technical fact but by litigation and regulation. The earliest legal work named the ingestion of copyrighted works for training as a question of infringement and exception, but it remained secondary while the authorship question was unresolved.[1,18] Once the human-authorship standard was fixed, the training-data question moved to the center, and the literature of the final period treats it as the field's new frontier, asking whether the text-and-data-mining exceptions that exist for research can be extended to commercial training and how a lawful training regime might be designed.[9,23] The shift is a case in which the resolution of one question revealed the next, and the review records it as the clearest example of the trajectory's logic.

The trajectory also has a dimension that the legal literature tends to understate but that the philosophical and empirical work brings forward, the social meaning of authorship itself. The question of who authored a work is not only about who owns it but about who is credited, who is accountable, and how the public understands the work's provenance, and the empirical studies of attribution show that these social meanings do not reduce to the legal rule.[4,17] This matters for the policy reading of the trajectory, because a law that assigns authorship to a human user while the public perceives the machine as the author has not, in the relevant sense, settled the question at all. The review therefore treats the social and the legal questions as distinct and argues that the field's future work must hold them together.

The empirical literature on how creativity is perceived adds a second, less visible movement to the trajectory, and it runs partly against the law's confidence. Studies of human evaluators find that people act as gatekeepers of creativity and are systematically biased against work they believe to be machine-made, discounting it even when the output is otherwise indistinguishable from human work, which suggests that the social recognition of authorship is itself a hurdle that no legal rule by itself removes.[4,28] Interviews and surveys of computer scientists and new-media artists show the practitioners themselves disagree about whether generative systems redefine creativity or merely interpolate within the space of what they were trained on, a disagreement that the philosophical literature of the same period encodes rather than resolves.[29,30] The philosophical tradition has its own long answer to this, distinguishing analytic from dialectical and crealectic accounts of creativity, and it reaches toward a position in which machine creativity is real but different in kind from human creativity, a distinction that the later, distributed accounts inherit.[31,32] The line of work that reads generative systems as instruments of emancipation rather than of replacement for digital artists completes this register, and together these studies show that the question of what creativity is remained open beneath the law's settling of who the author is.(33)

The legal literature of the window is broader than the leading American and European cases, and its breadth is itself a finding, because the same doctrinal pressure appears across very different systems. Studies of originality criteria show that the rise of generative systems has forced every system to re-examine what originality means and where the human contribution must be located, and the problem-and-solution literature across jurisdictions converges on the same set of unresolved points even where the doctrinal starting points differ.[34,35] Comparative work on the influence of artificial intelligence on copyright law reaches the same result from the other direction, that the technology changed the law's question rather than its answer, and analyses of the specific challenges and prospects of AI development in particular legal systems repeat the point in a narrower key.[36,37] The law-and-economics literature adds a frame that the doctrinal work often lacks, treating the authorship and training-data questions as problems of incentives and market design, and it concludes that the core challenge is not conceptual but economic, how to preserve the incentives for human creation while allowing the machine production the market clearly wants.(38) Against this, the institutional response has been cautious, a position captured in the observation that the field must rise to meet generative AI rather than legislate it away, and in the professional bodies' own measured responses to the technology.(39,40)

The training-data question, already identified as the new frontier, becomes clearer when the governance literature is read alongside the doctrinal work. The analyses that treat intellectual property and generative AI together argue that the ingestion of protected works for training is the

https://doi.org/10.62486/978-9915-9966-3-9.Ch01

point at which the technology's promise and the law's limits collide, and the case-study work on large language and image models shows that the collision is general rather than specific to any one system.[41,42] The technical and governance literature complements this by showing that the training-data problem is not only legal but technical and ethical, since the provenance, consent and disclosure questions cannot be separated from the question of lawful use.(43) Jurisdictional detail matters here, because text-and-data-mining exceptions differ sharply across systems, and the analyses of the Korean, European and American positions show that a model's training can be lawful in one place and not in another, a divergence that the international-investment framing makes economically consequential.[44,45] Personality rights add a further layer, since the output of a generative model can implicate the likeness and identity of real people in ways that copyright doctrine was never designed to address, and the comparative work on personality rights therefore treats the training-data and output questions as a single problem of rights rather than as two.(46)

The domain-specific literature shows that the authorship problem is not confined to the fine-art cases that dominate the headlines but reproduces itself wherever a creative practice meets a generative tool. The analysis of tattooing is the sharpest example, because it forces the question of who authored a design that a client described, an artist rendered and a machine may have generated, and it shows that the human-authorship standard does not map cleanly onto practices that are already collaborative.(47) The study of automated copyright enforcement extends the concern to the output side, asking whether the systems that police infringement can themselves be trusted to make the authorship judgments that the new technology demands, and the fairness and privacy literature warns that the same generative capacities raise questions of bias and data protection that the authorship debate, narrowly construed, leaves out.[48,49] These adjacent literatures do not change the trajectory the review reconstructs, but they confirm that the trajectory is a general property of the encounter between generative systems and creative institutions, not a peculiarity of the art market.

The agenda that follows from the review is specific. First, the operationalization of the human-involvement threshold, what kind and how much human contribution suffices, is the most practical open problem, because it is where the settled principle meets the unsettled facts, and it is best pursued through comparative work on the intermediate cases across jurisdictions.[11,12] Second, the training-data question needs a lawful regime that the field has not yet produced, and the design analyses of text-and-data-mining exceptions and of collective licensing are the most concrete proposals on the table.[9,15] Third, the gap between the legal and the philosophical accounts, a single author against a distributed network, deserves work that the review can only flag, since it bears on whether the next generation of law will describe the reality of machine creativity or continue to legislate against it.[13,26] These three items are the routes by which the trajectory this review reconstructs would continue, and they are offered as concrete directions rather than as a generic call for more study.

CONCLUSIONS

The evolution of legal and philosophical treatments of authorship, copyright and creativity in AI-generated art from 2018 to 2026 divides into four periods, and the review traced them in sequence. The GAN era opened the question, the diffusion models made it a mass concern, the institutional decisions of 2023 fixed a human-authorship requirement, and the years from 2024 to 2026 consolidated that requirement while adding regulation and training-data litigation. Copyright law converged on the principle that a work needs a human author, but it diverges on the boundary where a human shapes machine output, and it has not yet resolved the status of training data. Philosophy moved in the opposite direction, toward distributed and collaborative accounts of creativity that the legal standard does not reflect. The review concludes that the debate changed what it was about, from whether a machine can be an author to how much human involvement suffices and who bears responsibility for what a model was trained on, and that the gap between a law that insists on a single author and a philosophy that distributes authorship remains the field's central unresolved tension. The practical consequence is that the law has given practitioners a stable answer to the first question, a rule that a work needs a human author, while leaving the questions that now matter, the threshold of human contribution and the status of training data, to the litigation and legislation that the final period of the window set in motion, and the trajectory reconstructed here is best read as the unfinished first chapter of a longer settlement rather than as a completed one.

https://doi.org/10.62486/978-9915-9966-3-9.Ch01

REFERENCES

Chesterman S. ARTIFICIAL INTELLIGENCE AND THE LIMITS OF LEGAL PERSONALITY. ICLQ. 2020;69(4):819-844. doi: 10.1017/s0020589320000366
Xiao Y. Decoding Authorship: Is There Really no Place for an Algorithmic Author Under Copyright Law?. IIC. 2023;54(1):5-25. doi: 10.1007/s40319-022-01269-5
Deltorn JM, Macrez F. Authorship in the Age of Machine learning and Artificial Intelligence. SSRN Journal. 2018. doi: 10.2139/ssrn.3261329
Epstein Z, Levine S, Rand DG, Rahwan I. Who Gets Credit for AI-Generated Art?. iScience. 2020;23(9):101515. doi: 10.1016/j.isci.2020.101515
Cetinic E, She J. Understanding and Creating Art with AI: Review and Outlook. ACM Trans Multimedia Comput Commun Appl. 2022;18(2):1-22. doi: 10.1145/3475799
Lee HK. Rethinking creativity: creative industries, AI and everyday creativity. Media, Culture & Society. 2022;44(3):601-612. doi: 10.1177/01634437221077009
Murray MD. Generative AI Art: Copyright Infringement and Fair Use. STLR. 2023;26(2):259. doi: 10.25172/smustlr.26.2.4
Samuelson P. Generative AI meets copyright. Science. 2023;381(6654):158-161. doi: 10.1126/science.adi0656
Quintais JP. Generative AI, copyright and the AI Act. Computer Law & Security Review. 2025;56:106107. doi: 10.1016/j.clsr.2025.106107
Mazzi F. Authorship in artificial intelligence-generated works: Exploring originality in text prompts and artificial intelligence outputs through philosophical foundations of copyright and collage protection. World Intellectual Property. 2024;27(3):410-427. doi: 10.1111/jwip.12310
Samuelson P. How to Think about Remedies in the Generative AI Copyright Cases. Commun ACM. 2024;67(7):27-30. doi: 10.1145/3654699
Fritz J. Understanding authorship in Artificial Intelligence-assisted works. Journal of Intellectual Property Law and Practice. 2025;20(5):354-364. doi: 10.1093/jiplp/jpae119
Moruzzi C. Artificial Intelligence and Creativity. Philosophy Compass. 2025;20(3):e70030. doi: 10.1111/phc3.70030
Coeckelbergh M. The Work of Art in the Age of AI Image Generation. JHTR. 2023;1. doi: 10.59490/jhtr.2023.1.7025
Geiger C. Elaborating a Human Rights-Friendly Copyright Framework for Generative AI. IIC. 2024;55(7):1129-1165. doi: 10.1007/s40319-024-01481-5
Grilli L, Pedota M. Creativity and artificial intelligence: A multilevel perspective. Creat Innov Manage. 2024;33(2):234-247. doi: 10.1111/caim.12580
Sturm BLT, Iglesias M, Ben-Tal O, Miron M, Gómez E. Artificial Intelligence and Music: Open Questions of Copyright Law and Engineering Praxis. Arts. 2019;8(3):115. doi: 10.3390/arts8030115
Arriagada L. CG-Art: demystifying the anthropocentric bias of artistic creativity. Connection Science. 2020;32(4):398-405. doi: 10.1080/09540091.2020.1741514
Škiljić A. When Art Meets Technology or Vice Versa: Key Challenges at the Crossroads of AI-Generated Artworks and Copyright Law. IIC. 2021;52(10):1338-1369. doi: 10.1007/s40319-021-01119-w
Fedotova A, Romanov A, Kurtukova A, Shelupanov A. Authorship Attribution of Social Media and Literary Russian-Language Texts Using Machine Learning Methods and Feature Selection. Future Internet. 2021;14(1):4. doi: 10.3390/fi14010004
Tyagi K. Copyright, text & data mining and the innovation dimension of generative AI. Journal of Intellectual Property Law & Practice. 2024;19(7):557-570. doi: 10.1093/jiplp/jpae028
Mikalonytė ES, Kneer M. Can Artificial Intelligence Make Art?: Folk Intuitions as to whether AI-driven Robots Can Be Viewed as Artists and Produce Art. J Hum-Robot Interact. 2022;11(4):1-19. doi: 10.1145/3530875
Vesala J. Developing Artificial Intelligence-Based Content Creation: Are EU Copyright and Antitrust Law Fit for Purpose?. IIC. 2023;54(3):351-380. doi: 10.1007/s40319-023-01301-2
Vinchon F, Lubart T, Bartolotta S, Gironnay V, Botella M, Bourgeois-Bougrine S, et al. Artificial Intelligence & Creativity: A Manifesto for Collaboration. Journal of Creative Behavior. 2023;57(4):472-484. doi: 10.1002/jocb.597
Gaidartzi A, Stamatoudi I. Authorship and Ownership Issues Raised by AI-Generated Works: A Comparative Analysis. Laws. 2025;14(4):57. doi: 10.3390/laws14040057
Goodfellow P. The Distributed Authorship of Art in the Age of AI. Arts. 2024;13(5):149. doi: 10.3390/arts13050149
Hu X, Xing Y, Cai X, Zhao Y, Cook M, Borgo R, et al. Designing Interactions with Generative AI for Art and Creativity: A Systematic Review and Taxonomy. Proceedings of the 2025 ACM Designing Interactive Systems Conference. 2025:1126-1155. doi: 10.1145/3715336.3735843
Magni F, Park J, Chao MM. Humans as Creativity Gatekeepers: Are We Biased Against AI Creativity?. J Bus Psychol. 2024;39(3):643-656. doi: 10.1007/s10869-023-09910-x
Wingström R, Hautala J, Lundman R. Redefining Creativity in the Era of AI? Perspectives of Computer Scientists and New Media Artists. Creativity Research Journal. 2024;36(2):177-193. doi: 10.1080/10400419.2022.2107850
Hutson J, Lang M. Content creation or interpolation: AI generative digital art in the classroom. M. 2023;4(1):13. doi: 10.54517/m.v4i1.2158
Miranda LD. Artificial intelligence and philosophical creativity: From analytics to crealectics. Human Affairs. 2020;30(4):597-607. doi: 10.1515/humaff-2020-0053
Liu B. Arguments for the Rise of Artificial Intelligence Art: Does AI Art Have Creativity, Motivation, Self-awareness and Emotion?. Arte, individuo y sociedad (Internet). 2023;Avance en línea:1-11. doi: 10.5209/aris.83808

https://doi.org/10.62486/978-9915-9966-3-9.Ch01

Matias CF. Generative AI, Copyright and Emancipation: The Case of Digital Art. Law Tech Hum. 2024. doi: 10.5204/lthj.3567
KULINICH O, KONDYK D. Impact of Artificial Intelligence on the Criteria of Originality in Copyright Law. UNZ. 2023:43-56. doi: 10.37491/unz.93-94.4
Institute of law and public relations, Open International University of Human development “Ukraine, Kyiv, Ukraine, Valchuk D. ARTIFICIAL INTELLIGENCE COPYRIGHT: PROBLEMS AND SOLUTIONS. MSJ. 2024;3(3):7-15. doi: 10.36994/2786-9008-2024-3-1
Çebi E, Reisoğlu P, Goktas E. The Influence of Artificial Intelligence on Copyright Law. ISSLP. 2023;2(2):33-41. doi: 10.61838/kman.isslp.2.2.5
Waris M, Kim SI. A Study of Analyzing the Challenges and Prospects of Artificial Intelligence Development and its Impact on Copyright Law in the Digital Age. JDMCT. 2024;4(1):1-12. doi: 10.29056/jdmct.2024.06.01
Cooper Z, Martens B, Peukert C, Stocker V. The Law and Economics of Generative AI and Copyright: A Primer to Core Challenges for Our Digital Future. Review of Network Economics. 2025;24(3):161-176. doi: 10.1515/rne-2025-0039
Scott I. Rising to Meet the Challenge of Generative AI. J Legal Studies Education. 2024;41(1):29-37. doi: 10.1111/jlse.12141
Kretschmer M, Meletti B, Porangaba LH. Artificial intelligence and intellectual property: copyright and patents—a response by the CREATe Centre to the UK Intellectual Property Office’s open consultation. Journal of Intellectual Property Law & Practice. 2022;17(3):321-326. doi: 10.1093/jiplp/jpac013
Chesterman S. Good models borrow, great models steal: intellectual property rights and generative AI. Policy and Society. 2025;44(1):23-37. doi: 10.1093/polsoc/puae006
Lucchi N. ChatGPT: A Case Study on Copyright Challenges for Generative Artificial Intelligence Systems. Eur j risk regul. 2024;15(3):602-624. doi: 10.1017/err.2023.59
Pasetti M, Santos JW, Corrêa NK, de Oliveira N, Barbosa CP. Technical, legal, and ethical challenges of generative artificial intelligence: an analysis of the governance of training data and copyrights. Discov Artif Intell. 2025;5(1):193. doi: 10.1007/s44163-025-00379-6
Kim K. Korean Copyright Issues in Text Data Mining for Generative AI. Journal of AI Law and Regulation. 2024;1(1):64-76. doi: 10.21552/aire/2024/1/8
Lentner GM. Generative AI and copyright in the EU and the USA: a view from international investment law. Journal of Intellectual Property Law & Practice. 2026;21(2):121-128. doi: 10.1093/jiplp/jpaf086
Chopra P, Sony RAL, Chopra S. Generative AI, Copyright and Personality Rights: A Comparative Legal Perspective. LIDA. 2025;6(3):23-51. doi: 10.17323/2713-2749.2025.3.23.51
Stockton-Brown M. Inking Cultures: Authorship, AI-Generated Art and Copyright Law in Tattooing. Int J Semiot Law. 2023;36(5):2037-2065. doi: 10.1007/s11196-023-10024-z
Gray JE, Suzor NP. Playing with machines: Using machine learning to understand automated copyright enforcement at scale. Big Data & Society. 2020;7(1):2053951720919963. doi: 10.1177/2053951720919963
Xiang A. Fairness & Privacy in an Age of Generative AI. stlr. 2024;25(2). doi: 10.52214/stlr.v25i2.12765

Chapter 2 / Capítulo 2

SAP Superintelligence Series
ISSN: 3144-5624
ISBN (complete work): 978-9915-9851-0-7

Superintelligence Series Volume 5: Intelligent Visual Arts and Audiovisual Engineering: Generative AI, Computational Creativity and Superintelligence
ISBN (this volume): 978-9915-9966-5-3
DOI: 10.62486/978-9915-9966-3-9.Ch02
Pages: 12-23

Check for updates

© 2026 The authors. This is an open access book chapter distributed under the terms of the Creative Commons Attribution (CC BY) 4.0 License.

Openness and licensing of text-to-image and text-to-video models on the Hugging Face Hub

Apertura y licencias de los modelos de texto a imagen y texto a video en Hugging Face Hub

Yasniel Sánchez Suárez[1]

1 Universidad de Matanzas. Matanzas, Cuba.

ABSTRACT

Introduction: Generative image and video models have spread rapidly across research and industry, but the meaning of their openness is contested. Empirical evidence on how licenses and documentation have evolved is scarce. Methods: Census of the Hugging Face Hub public registry filtered by the text-to-image and text-to-video pipeline tags, downloaded automatically on 2026-10-25. License categories were assigned from each model license tag and openness and documentation indicators were computed per year of creation. Results: The census identified 115,532 models, 112,574 text-to-image and 2,958 text-to-video. Text-to-image models grew from 1,454 in 2022 to 42,729 in 2025. OpenRAIL licenses were the most common (32.1%), followed by custom or unspecified licenses (38.9%) and models with no license (16.0%); permissive licenses accounted for 11.7%. The share of models with an open license and unrestricted access fell from 85.1% in 2022 to 27.1% in 2025. Library integration reached 93.0% and model cards 95.5% of models. Conclusions: The volume of open image and video models expanded several-fold while their licensing shifted away from unrestricted open terms toward use-restricted and underspecified licenses. Documentation is near universal, but licensing openness has narrowed over time.

Keywords: artificial intelligence; generative models; open source software; software licenses; image synthesis; documentation

RESUMEN

Introducción: Los modelos generativos de imagen y video se han extendido con rapidez en la investigación y la industria, pero el significado de su apertura es objeto de debate. La evidencia empírica sobre la evolución de licencias y documentación es escasa. Métodos: Censo del registro público de Hugging Face Hub filtrado por las etiquetas de texto a imagen y texto a video, descargado de forma automática el 2026-10-25. Las categorías de licencia se asignaron a partir de la etiqueta de cada modelo y se calcularon indicadores de apertura y documentación por año de creación. Resultados: El censo identificó 115 532 modelos, 112 574 de texto a imagen y 2958 de texto a video. Los modelos de texto a imagen pasaron de 1454 en 2022 a 42 729 en 2025. Las licencias OpenRAIL fueron las más frecuentes (32,1%), seguidas de licencias personalizadas o no especificadas (38,9%) y de modelos sin licencia (16,0%); las licencias permisivas representaron 11,7%. La proporción de modelos con licencia abierta y acceso sin restricciones cayó de 85,1% en 2022 a 27,1% en 2025. La integración con bibliotecas alcanzó 93,0% y las fichas de modelo 95,5%. Conclusiones: El volumen de modelos abiertos de imagen y video se multiplicó mientras sus licencias se alejaron de los términos abiertos sin restricciones hacia licencias de uso restringido y poco especificadas. La documentación es casi universal, pero la apertura de las licencias se ha reducido con el tiempo.

Palabras clave: inteligencia artificial; modelos generativos; software de código abierto; licencias de software; síntesis de imágenes; documentación

INTRODUCTION

The release of Stable Diffusion in August 2022 turned text-to-image generation from a research demonstration into a mass activity, and text-to-video generation followed within roughly two years.[1,2] The models that drove this expansion were not distributed like commercial software. Their weights were posted to public repositories with a license attached, and anyone could download, fine-tune, and republish them. Within just four years the number of such models reached six figures on a single hosting platform, a scale at which the collective licensing choices of individual authors become a de facto policy for the field. The infrastructure that made this possible was itself open: the Transformer architecture and the ecosystem of shared libraries built around it provided the substrate on which thousands of derivative models were trained and exchanged.(3) Whether that arrangement should be called open, and what openness should require of a generative model, is now a genuine dispute rather than a settled matter of definition. (4)

Two positions structure the debate. One holds that open releases of generative models accelerate scientific progress, allow external scrutiny, and prevent capability from concentrating in a small number of firms.[5,6] The other holds that releasing weights without the underlying data, code, or compute amounts to a limited form of openness that still concentrates power upstream, and that unconstrained release enables misuse.[7,8] Solaiman framed the practical question as a gradient of release decisions rather than a binary choice, arguing that each stage from fully closed to fully open carries a different balance of risk and benefit.(9) This is more than a dispute over words, because downstream users rely on the label open when they decide whether a model can be reused, modified, or audited.

The licensing instruments themselves encode a third position. The RAIL family of licenses, designed for large models, permit reuse while attaching use-based restrictions, and are defended as a responsible middle ground.(10) Critics describe these as source-available rather than open, because the use restrictions violate the traditional open source requirement that a license not discriminate by field of endeavor.[4,11] A related line of work uses the phrase open-washing for models that borrow the rhetoric of openness without meeting its conditions.[12,13] Each side of this argument rests largely on selected examples; large-scale measurement of how the actual population of generative models is licensed and documented has been slow to arrive, even as surveys of the technical literature have catalogued the rapid growth of diffusion methods and their applications.(14,15)

The Hugging Face Hub is the natural place to take that measurement. It has become the default registry for open machine learning artifacts, and several recent studies treat it as an empirical window onto the AI development ecosystem, documenting who builds models, how development activity concentrates, and how inconsistently models are documented.(16-18) The broader open-science literature connects these platform dynamics to the governance of generative artificial intelligence and to the reuse of artifacts that carry incomplete provenance. (19) In image and video generation specifically, the adoption of open diffusion models has been visible across domains from visual arts education to remote sensing and medical imaging, which makes the legal terms attached to those models consequential well beyond the machine learning community.(15,20)

The question is also a preview of problems that will recur across advanced artificial intelligence. Image and video generation were the first capabilities in which a mass public began to download and adapt generative models outside the control of any single laboratory, and the licensing arrangements that emerged there anticipate the arrangements now being negotiated for language models and, prospectively, for agents and multimodal systems.[9,21] Measuring how open those arrangements actually were, and how they changed as the field matured, therefore offers evidence that bears on debates well beyond image synthesis. What has not yet been measured systematically is how the openness of generative image and video models, specifically, changed as their numbers grew from year to year. The objective of this study was therefore to characterize the growth, licensing, and documentation of text-to-image and text-to-video models registered on the Hugging Face Hub between 2022 and 2026, using a full census of the two pipeline categories.

https://doi.org/10.62486/978-9915-9966-3-9.Ch02

METHODS

STUDY DESIGN AND DATA SOURCE

A cross-sectional registry census was conducted of the Hugging Face Hub public application programming interface.(22) The Hub is a public repository that lists machine learning models, datasets, and applications with metadata supplied by their authors. The census was restricted to models whose declared pipeline tag was text-to-image or text-to-video, the two categories that cover generative image and video synthesis.

VARIABLES

The creation date of each model was taken from the createdAt field and collapsed into a calendar year. License information was taken from the license tag that authors attach to a model; the tag follows the SPDX or Creative Commons convention when one is used, and otherwise carries a free-text or platform value such as other, unknown, or a vendor-specific name. Each license was assigned to one of seven mutually exclusive categories. Permissive covered OSI-compliant permissive licenses such as Apache-2.0, MIT, BSD, and CC0. Attribution covered Creative Commons Attribution variants. Copyleft covered the GPL family, share-alike Creative Commons licenses, and comparable reciprocal terms. OpenRAIL covered the RAIL license family, including CreativeMLOpenRAIL-M and OpenRAIL++. Restricted covered non-commercial and no-derivatives Creative Commons variants. Custom covered platform values such as other and unknown and vendor-specific licenses. Models with no license tag were recorded as having no license.

Access gating was recorded from the gated field, which marks models that require a request or agreement before their weights can be downloaded. Documentation was measured with two indicators. Library integration was recorded when a model declares a supported library, which indicates that the model can be loaded through a standard pipeline rather than only through ad hoc code. Model card presence was recorded when the model files include a README document, the mechanism by which the Hub surfaces structured documentation about a model. These indicators are deliberately shallow; they capture whether documentation exists, not whether it is accurate or complete, a distinction that matters for interpreting the results.

PROCEDURE AND ANALYSIS

The complete lists of the two pipeline categories were retrieved, and each record was reduced to its identifier, author, pipeline, creation year, license, gated status, library integration, model card presence, and download and like counts. No minimum download or like threshold was applied, so the census includes every publicly listed model regardless of its popularity or activity. Records with no creation date were retained and tabulated separately where relevant. The analysis was descriptive: counts and proportions by pipeline and year, the distribution of license categories, and openness and documentation indicators. A composite openness measure counted models carrying a permissive, attribution, copyleft, or OpenRAIL license and not subject to access gating; the choice of these four categories as open follows the common but contested practice of treating attribution and use-restricted RAIL terms as open enough for reuse, and the alternative of treating only permissive licenses as open is reported as a sensitivity check in the discussion. The share of this composite and of its complement were plotted by year of creation.

All counts were produced by a re-runnable Python script that executes the download and aggregation steps in sequence. The script was executed twice and produced byte-identical metric files, confirming that no stochastic element entered the computation. Figures were generated as grayscale two-dimensional plots with a resolution of 300 dots per inch. Because the analysis is a census of a population rather than a sample, no inferential statistics were computed; every reported proportion is an exact population share at the snapshot date, and no confidence intervals or significance tests apply. This is the appropriate approach when the goal is to describe the state of a registry, not to generalize from it. Because the registry is public and contains no personal data beyond usernames, and because no human participants were involved, no ethical approval or participant consent was required.

https://doi.org/10.62486/978-9915-9966-3-9.Ch02

DATA QUALITY

The registry is a self-reporting platform, and its metadata inherit the limitations of that design. Model authors set the pipeline tag, the license tag, and the gated flag, and the Hub does not verify these declarations against the model files. Download and like counts are computed by the platform and were used only to describe concentration. Two measures were taken to bound the influence of these uncertainties. First, the census was taken as a single-point snapshot rather than a series, so that the year-to-year comparisons reflect one consistent collection rather than a changing method. Second, the sensitivity of the openness trajectory to the definition of open was examined by recomputing the composite measure with and without the OpenRAIL category, which is the single largest source of ambiguity in the license classification. The results of that check are reported alongside the main findings, so that the reader can see how much of the observed openness depends on a contested definitional choice.

RESULTS

The census identified 115,532 models, of which 112,574 carried the text-to-image tag and 2,958 the text-to-video tag (Table 1). Text-to-image models were created in substantial numbers from 2022 onward, reaching 1,454 in that first year and then 42,729 in 2025, before the incomplete year 2026 recorded 20,741 through late July.

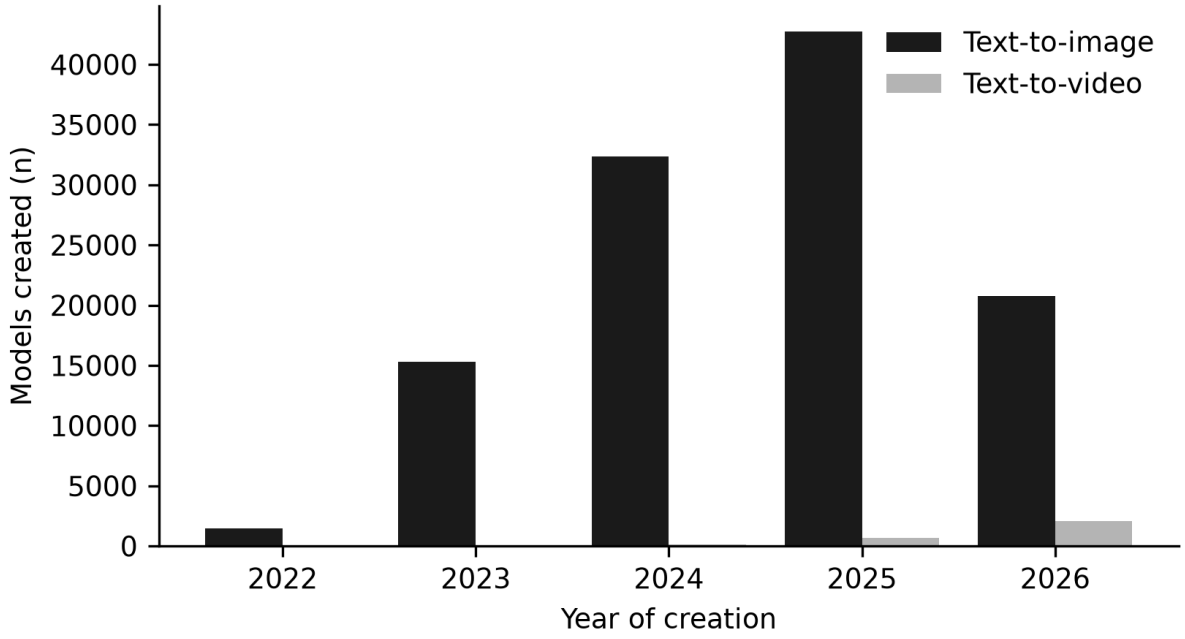
Table 1. Models registered on the Hugging Face Hub by pipeline and year of creation, 2022-2026.

Pipeline	2022	2023	2024	2025	2026	Total
text-to-image	1454	15300	32350	42729	20741	112574
text-to-video	0	74	123	685	2076	2958
Total	1454	15374	32473	43414	22817	115532

The expansion was a 29.40-fold increase in the annual volume of text-to-image models between 2022 and 2025. Text-to-video models appeared later and remained far less numerous, with none recorded in 2022 and 2,076 in 2026. The two categories followed the same broad trajectory, although text-to-video began its steepest growth roughly two years after text-to-image (Figure 1). Text-to-image models accounted for 97.4% of the census and text-to-video models for 2.6%.

Download activity was heavily concentrated among a small number of models. The one hundred most downloaded models in the census captured 70.1% of all recorded downloads, leaving the remaining 115,532 models to share less than one third of the total.

Figure 1. Models created per year by pipeline, 2022-2026.



https://doi.org/10.62486/978-9915-9966-3-9.Ch02

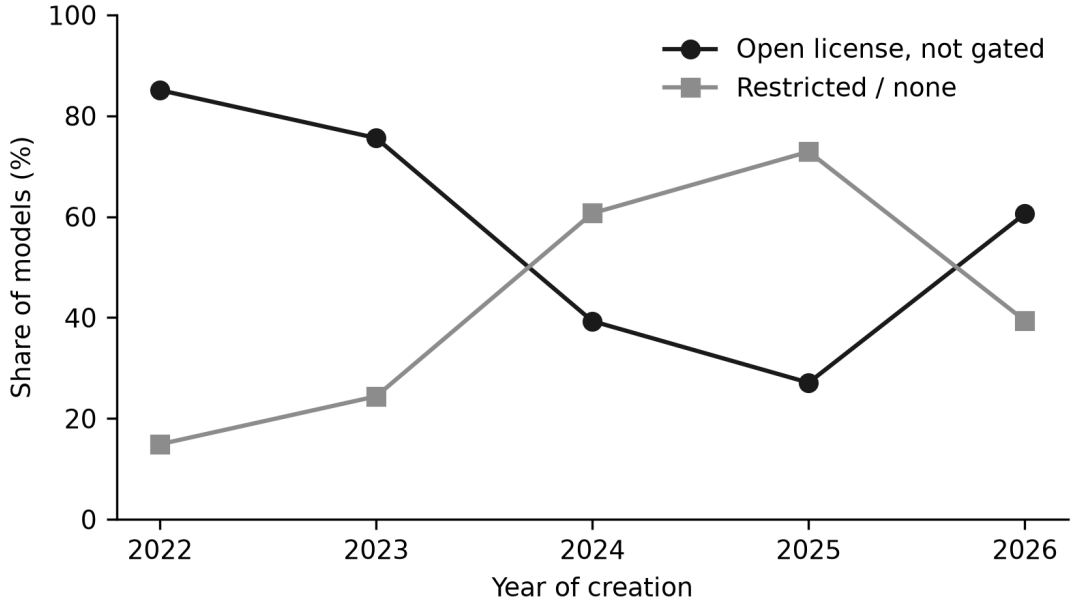
Licenses concentrated in three categories (Table 2). OpenRAIL licenses were the single most common choice, accounting for 32.1% of all models, followed by custom or unspecified licenses at 38.9% and models with no license at all at 16.0%. Truly permissive licenses made up 11.7%. Attribution, copyleft, and restricted categories were each below one percent, at 0.3%, 0.2%, and 0.8% respectively. In absolute terms, the OpenRAIL category was dominated by text-to-image models, which supplied the overwhelming majority of the 37,087 models carrying a RAIL license. The two pipelines differed in their profiles: text-to-image models leaned on custom and no-license arrangements, whereas text-to-video models carried a comparatively higher share of permissive terms. The restricted category, though small, was present in both pipelines and indicates that a segment of authors actively choose non-commercial or no-derivatives terms. The combined permissive, attribution, copyleft, and OpenRAIL categories, which constitute the open share under the composite definition, accounted for 43.9% of all models once gated models were excluded.

Table 2. Distribution of model licenses by category and pipeline.

License category	Text-to-image	Text-to-video	All models	% of all
Permissive (OSI-compliant)	12504	1047	13551	11.7
Attribution (CC-BY)	311	7	318	0.3
Copyleft / share-alike	269	5	274	0.2
OpenRAIL (responsible AI)	36279	808	37087	32.1
Restricted / non-commercial	901	67	968	0.8
Custom / other	44211	676	44887	38.9
No license	18099	348	18447	16.0

The share of models carrying an open license and free of access gating declined as the categories matured (Figure 2). In 2022 the composite openness measure stood at 85.1%; it fell to 39.3% in 2024 and 27.1% in 2025 before recovering to 60.6% among models created in 2026. The steepest single-year change occurred between 2023 and 2024, when the open share dropped by more than thirty percentage points, a decline that coincides with the period in which the registry's annual volume first exceeded thirty thousand models. The mirror pattern was a rising share of models carrying restricted, custom, or no license, which moved from 14.9% in 2022 to 72.9% in 2025. When openness was redefined to require a permissive, attribution, or copyleft license only, the open share was far lower in every year, because the OpenRAIL category, which the composite measure counts as open, represented a large and growing fraction of the registry.

Figure 2. Share of models with an open license and unrestricted access versus a restricted or absent license, by year of creation.



https://doi.org/10.62486/978-9915-9966-3-9.Ch02

Documentation was the dimension on which the registry performed best (Table 3). Library integration was present in 93.0% of all models and model cards in 95.5%. Text-to-video models were integrated with a library at a lower rate than text-to-image models (78.2% against 93.4%), but every text-to-video model in the census carried a model card. Access gating was uncommon overall, affecting 1.3% of models and 1.8% of text-to-video models. When the gating and license findings are read together, the typical model in the registry is easy to find, easy to understand, and legally ambiguous to reuse: it can be downloaded without a gate, it carries a readable card, and it offers no clean statement of what a downstream user is permitted to do with it.

Table 3. Openness and documentation indicators by pipeline.

Indicator	Text-to-image	Text-to-video	All models
Models, n	112574	2958	115532
Gated access, %	1.3	1.8	1.3
Library integration, %	93.4	78.2	93.0
Model card (README), %	95.4	100.0	95.5

DISCUSSION

This census found that the volume of open image and video models grew by roughly an order of magnitude over four years, while their licensing moved steadily away from the unrestricted openness that the word open usually implies. More than one third of all models carried a use-restricted RAIL license, another two fifths carried a custom or unspecified license or none at all, and only about one in nine carried a permissive license. The trajectory of the composite measure makes the shift concrete: unrestricted open models were the clear majority in 2022 and had become a minority by 2025. The two dimensions of openness examined here moved in opposite directions, with documentation improving to near universality at the same time that licensing narrowed.

INTERPRETATION AGAINST THE OPENNESS DEBATE

The disagreement that frames the introduction bears directly on this result. Defenders of unrestricted openness treat permissive licensing as the default that keeps the ecosystem accessible, auditable, and open to participation.[5,6] What this census shows is that the practitioners who publish image and video models did not converge on that default. The most frequent choice, OpenRAIL, is precisely the instrument that the open-source movement refuses to call open because its use restrictions violate the field-of-endeavor clause of the open source definition.[4,11] The finding therefore supplies quantitative support to the claim that generative model openness has settled, in practice, on a middle position that neither side of the debate fully endorses.[10,12] Neither the optimistic account of a permissive commons nor the pessimistic account of a closed corporate ecosystem describes the registry accurately, and both readings miss the genuinely new fact that use-restricted openness has become the modal form.

The decline in the open share between 2022 and 2025 matches what Widder and colleagues predicted from the political economy of open artificial intelligence: early landmark models such as Stable Diffusion were released permissively to build an ecosystem, and the subsequent wave of derivative and fine-tuned models reused those weights under more varied and less generous terms.[1,7] The download concentration documented here strengthens that reading. Because most reuse flows to a handful of base models, the license attached to those few models does more to shape downstream practice than the licenses of the long tail, and the early permissive base models remain the load-bearing layer of the ecosystem even as the surrounding registry becomes less open.(16) The concentration in the custom and no-license categories points in the same direction. A model posted without a license, or under an opaque vendor label, cannot be clearly reused at all, which is the practical opposite of the transparency that the open label is meant to guarantee.[8] The finding also resonates with work on the false promise of open imitation: models that reproduce the surface behavior of closed systems without their underlying assets do not deliver the reuse value that openness is supposed to provide.(23)

The partial recovery observed among models created in 2026 should be read cautiously. The year is incomplete in the snapshot, which means the 2026 denominator is biased toward the earlier

https://doi.org/10.62486/978-9915-9966-3-9.Ch02

months, and newer derivative models may inherit open terms from their base models rather than reflect a genuine change in author preferences. The sensitivity check is more informative than the headline trajectory: when openness is defined strictly to exclude OpenRAIL, the open share is small in every year. This means the apparent openness of the registry depends heavily on whether use-restricted licenses are counted as open, which is exactly the point of contention in the definitional debate.(13)

Several mechanisms plausibly explain the shift. The most important is that the models now anchoring the ecosystem were released under terms more restrictive than the permissive licenses of the first generation. Commercial labs that publish image and video models have moved toward use-restricted or custom licenses in order to reserve commercial rights, and the large population of derivative models inherits those terms through fine-tuning. A second mechanism is the increasing use of the registry by developers who publish models as portfolio artifacts rather than as reusable infrastructure; for such authors, a custom or absent license is a default rather than a decision. A third is the sheer scale of the long tail: tens of thousands of small fine-tunes are created under time pressure, and license selection is the step most likely to be skipped when the immediate goal is to share a result. These mechanisms are not mutually exclusive, and the data collected here cannot separate them, but together they explain why documentation, which the platform makes easy, became universal while clean licensing, which requires a legal judgment, did not.(18,24)

DOCUMENTATION AND THE TWO FACES OF OPENNESS

The documentation indicators complicate the pessimistic reading. Library integration and model cards are near universal, meaning the ecosystem adopted the documentation practices that Mitchell and colleagues proposed for accountable model reporting and that later work extended to model registries.(24-26) Documentation is, however, the cheapest form of openness to provide, and its abundance alongside a shortage of clean licenses suggests that authors comply with the registry norms that require a model card while leaving the legally consequential licensing decision underspecified. Nunes and colleagues found that model cards can prompt ethical reflection among developers, which implies that the near-universal card is a real signal rather than an empty formality.(27) Stalnaker and colleagues reached a compatible conclusion when they found documentation, supply-chain, and licensing problems co-occurring on the Hub, and Pepe and colleagues documented similar inconsistencies longitudinally.[18,24] Oreamuno and colleagues, working across multiple model stores, likewise found that third-party models are documented unevenly at best.(26) The contribution of this census is to show that, at least in image and video generation, the documentation norm has been internalized far more thoroughly than the licensing norm, which is the reverse of what a purely legal view of openness would predict.

The contrast between abundant documentation and scarce permissive licensing can be read as a division of labor between two different functions of openness. Documentation serves transparency and reproducibility, which the platform enforces through norms and tooling; licensing serves reuse and redistribution, which remains a decision authors make under competing incentives. The open-science literature observes the same tension: generative tools promise to make research more transparent, but that promise depends on the legal terms attached to the artifacts, not only on their availability.(19) Software engineering research has long documented that license management carries hidden costs and that ambiguous terms downstream become expensive problems upstream.(28)

The pattern in image and video models parallels, and partly precedes, the better-studied case of large language models. The open-weight language model movement has been scrutinized through transparency indexes that show leading open models disclose far less than their open label suggests, and through critiques of models that imitate proprietary systems without reproducing their assets.[23,29] Comparisons of open-weight and proprietary systems increasingly frame openness as a matter of degrees rather than a binary property.[21,30] The image and video registry examined here reaches the same conclusion by a different route: the degrees are visible in the license distribution itself, where a genuine permissive minority sits alongside a large use-restricted middle and an even larger underspecified remainder. The correspondence between the two domains suggests that the erosion of openness is not idiosyncratic to image synthesis but reflects incentives that operate across generative modeling more generally, which raises the stakes of the definitional and regulatory questions that remain open.

https://doi.org/10.62486/978-9915-9966-3-9.Ch02

THE TEXT-TO-VIDEO TRAJECTORY

The text-to-video trajectory is worth noting on its own. Text-to-video began roughly two years later and remains approximately forty times smaller than text-to-image, yet it grew from almost nothing in 2023 to over two thousand models in 2026, consistent with the rapid diffusion of video generation methods and their applications.[2,15] Its lower rate of library integration suggests the tooling for video pipelines matured later than for image pipelines, which is itself a signal about how open ecosystems develop in stages, with shared infrastructure appearing only after a critical mass of models exists.[16,31] The early stage of the text-to-video ecosystem makes its licensing pattern worth watching, because the decisions made now will shape what reuse is possible as the category grows.

This staged development matters for the broader question of advanced artificial intelligence. Image and video generation are among the earliest domains in which the open versus closed debate played out at scale, which makes them a preview of the tensions that later arrived with instruction-tuned language models and that are now appearing around autonomous agents and multimodal systems.[8,21] The image and video evidence suggests that, left to the incentives of a commercial platform, an ecosystem converges on a mix in which a small permissive core supports a large, weakly licensed periphery. Whether the same equilibrium holds for more consequential capabilities is an open question, but the trajectory documented here offers a baseline against which later domains can be compared.(9)

LIMITATIONS

The census captures only models registered with the two pipeline tags, so image and video models posted under a different or missing tag are excluded, and the true population of open generative models is therefore larger than the number reported here; a replication that scans additional tags and categories would bound the size of this gap. The license tag is self-declared, and its accuracy was not audited against the full text of each license; authors may label a model permissively while attaching terms that are not, and a sample-level audit of license files would test the reliability of the self-declared tags. A registry study cannot distinguish a model that is open in name but not in practice, because the weights, data, and training code are not all disclosed in the metadata, a gap that transparency indexes are designed to measure.(29) Access gating was measured from the gated flag, but gating does not capture other downstream restrictions such as provider terms of service or platform-level controls, and a broader instrument that also records data and code availability would give a more complete measure. The 2026 figures are incomplete and should be treated as a lower bound for that year. The descriptive design intentionally forgoes inferential claims; the census documents what exists, but it cannot establish the motivations behind authors' licensing choices, which would require a survey or qualitative study. Each of these limitations points to a concrete remedy, and together they outline an agenda in which registry metadata are validated against model files and enriched with the transparency dimensions of data, code, and compute disclosure.(17,29)

IMPLICATIONS

The practical implication is that openness cannot be assumed from the presence of a model on a public repository. Organizations that plan to build on a generative model must examine its license and gating status rather than its visibility, and the distinction drawn here between permissive, use-restricted, and unspecified licensing provides a workable checklist for that review. The finding also bears on the governance of open foundation models, which have been treated as a public good worth preserving and whose transparency and control properties shape whether open-weight alternatives can be adopted responsibly.(6,30,32) If the open share continues to erode, the ecosystem that open models were meant to sustain becomes harder to audit and to reuse, and the safety and transparency arguments for openness lose their empirical basis.[9,21] The same dynamic extends into the software industry that consumes these models: generative assistants and code generation have been absorbed into engineering practice, and the legal clarity of the underlying models determines how far that integration can be trusted.(33)

For the platform that hosts these models, the findings suggest a concrete reform. Requiring an explicit license, and preventing unlicensed models from presenting themselves as reusable, would convert the large no-license category into a set of declared terms and remove

https://doi.org/10.62486/978-9915-9966-3-9.Ch02

the ambiguity that currently forces every downstream user to perform their own legal review. For the open-source movement, the results argue for clarity about the status of use-restricted licenses: the empirical reality is that a large share of what circulates as open generative models is licensed under exactly the terms that the movement excludes. Pretending otherwise leaves users and regulators without a shared vocabulary. Finally, the study itself is a template: because the registry is machine-readable and public, the same census can be repeated on a fixed schedule, turning a one-off snapshot into a monitoring series that would reveal whether the erosion of openness documented here continues, stabilizes, or reverses as text-to-video matures and as new generative modalities appear. The challenge for the open-source movement, and for the regulators who now write rules around it, is that the empirical center of gravity has moved to a middle ground that neither a strict permissive definition nor a fully proprietary posture describes.

CONCLUSIONS

The census of the Hugging Face Hub documented a several-fold expansion of generative image and video models between 2022 and 2026, with text-to-image models growing to tens of thousands per year and text-to-video models emerging about two years later. The licensing of these models shifted away from unrestricted openness, as use-restricted, custom, and unspecified licenses came to dominate the registry and the share of models that were both open and ungated fell from a majority in 2022 to a minority in 2025. Documentation moved in the opposite direction, reaching near-universal coverage of library integration and model cards. Openness, as measured by licensing and access, narrowed as the ecosystem expanded, and the empirical record now sits between the permissive and proprietary positions that dominate the policy debate. The results argue for treating openness as a measurable, year-to-year property of the ecosystem rather than as a static label applied once and then assumed to hold.

https://doi.org/10.62486/978-9915-9966-3-9.Ch02

REFERENCES

Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B. High-Resolution Image Synthesis with Latent Diffusion Models. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022:10674-10685. doi: 10.1109/cvpr52688.2022.01042
Wang Y, Chen X, Ma X, Zhou S, Huang Z, Wang Y, et al. LaVie: High-Quality Video Generation with Cascaded Latent Diffusion Models. Int J Comput Vis. 2025;133(5):3059-3078. doi: 10.1007/s11263-024-02295-1
Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. Transformers: State-of-the-Art Natural Language Processing. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. 2020:38-45. doi: 10.18653/v1/2020.emnlp-demos.6
Open Source Initiative. The Open Source AI Definition [Internet]. 2024. Available from: https://opensource.org/ai/open-source-ai-definition
Spirling A. Why open-source generative AI models are an ethical way forward for science. Nature. 2023;616(7957):413-413. doi: 10.1038/d41586-023-01295-4
Kapoor S. On the Societal Impact of Open Foundation Models [Internet]. 2024. Available from: https://arxiv.org/abs/2403.07918
Widder DG, West S, Whittaker M. Open (For Business): Big Tech, Concentrated Power, and the Political Economy of Open AI [Internet]. Elsevier BV; 2023. Available from: https://doi.org/10.2139/ssrn.4543807
Liesenfeld A, Lopez A, Dingemanse M. Opening up ChatGPT: Tracking openness, transparency, and accountability in instruction-tuned text generators. Proceedings of the 5th International Conference on Conversational User Interfaces. 2023:1-6. doi: 10.1145/3571884.3604316
Solaiman I. The Gradient of Generative AI Release: Methods and Considerations. 2023 ACM Conference on Fairness, Accountability, and Transparency. 2023:111-122. doi: 10.1145/3593013.3593981
Contractor D, McDuff D, Haines JK, Lee J, Hines C, Hecht B, et al. Behavioral Use Licensing for Responsible AI. 2022 ACM Conference on Fairness Accountability and Transparency. 2022:778-788. doi: 10.1145/3531146.3533143
Cui J, Araujo DA. Rethinking use-restricted open-source licenses for regulating abuse of generative models. Big Data & Society. 2024;11(1):20539517241229699. doi: 10.1177/20539517241229699
Liesenfeld A, Dingemanse M. Rethinking open source generative AI: open washing and the EU AI Act. The 2024 ACM Conference on Fairness, Accountability, and Transparency. 2024:1774-1787. doi: 10.1145/3630106.3659005
Ramos S, Pianese F. Regulating transparency: a critique of the “free and open source AI” definition in the EU AI act. AI Ethics. 2026;6(3):317. doi: 10.1007/s43681-026-01177-1
Chen H, Xiang Q, Hu J, Ye M, Yu C, Cheng H, et al. Comprehensive exploration of diffusion models in image generation: a survey. Artif Intell Rev. 2025;58(4):99. doi: 10.1007/s10462-025-11110-3
Alhabeeb SK, Al-Shargabi AA. Text-to-Image Synthesis With Generative Models: Methods, Datasets, Performance Metrics, Challenges, and Future Direction. IEEE Access. 2024;12:24412-24427. doi: 10.1109/access.2024.3365043
Osborne C, Ding J, Kirk HR. The AI community building the future? A quantitative analysis of development activity on Hugging Face Hub. J Comput Soc Sc. 2024;7(2):2067-2105. doi: 10.1007/s42001-024-00300-8
Ait A, Cánovas Izquierdo JL, Cabot J. On the suitability of hugging face hub for empirical studies. Empir Software Eng. 2025;30(2):57. doi: 10.1007/s10664-024-10608-8
Pepe F, Nardone V, Mastropaolo A, Canfora G, Bavota G, Di Penta M. Datasets, bias, licenses, and terms of use: A large and longitudinal study on the documentation of hugging face machine learning models. Empir Software Eng. 2026;31(4):95. doi: 10.1007/s10664-026-10843-1
Hosseini M, Horbach SPJM, Holmes K, Ross-Hellauer T. Open Science at the generative AI turn: An exploratory analysis of challenges and opportunities. Quantitative Science Studies. 2025;6:22-45. doi: 10.1162/qss_a_00337
Dehouche N, Dehouche K. What’s in a text-to-image prompt? The potential of stable diffusion in visual arts education. Heliyon. 2023;9(6):e16757. doi: 10.1016/j.heliyon.2023.e16757
Chia SK. Selecting Generative AI Systems for Education and Business in 2026: A Capability–Risk Framework for Comparing Frontier and Open-Weight Model Ecosystems. CHIATECH JOURNAL. 2026;1(1):e010. doi: 10.68232/cj.e010
Hugging Face. Hugging Face Hub model registry (public API) [Internet]. 2026. Available from: https://huggingface.co/api/models
Gudibande A. The False Promise of Imitating Proprietary LLMs [Internet]. 2023. Available from: https://arxiv.org/abs/2305.15717
Stalnaker T, Wintersgill N, Chaparro O, Heymann LA, Penta MD, German DM, et al. An Empirical Analysis of Machine Learning Model and Dataset Documentation, Supply Chain, and Licensing Challenges on Hugging Face. ACM Trans Softw Eng Methodol. 2026;35(8):1-49. doi: 10.1145/3776739
Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B, et al. Model Cards for Model Reporting. Proceedings of the Conference on Fairness, Accountability, and Transparency. 2019:220-229. doi: 10.1145/3287560.3287596
Oreamuno EL, Khan RM, Bangash AA, Stinson C, Adams B. The State of Documentation Practices of Third-Party Machine Learning Models and Datasets. IEEE Softw. 2024;41(5):52-59. doi: 10.1109/ms.2024.3366111
Nunes JL, Barbosa GDJ, De Souza CS, Barbosa SDJ. Using Model Cards for ethical reflection on machine learning models: an interview-based study. JIS. 2024;15(1):1-19. doi: 10.5753/jis.2024.3444
Li B, Liu C, Fan L, Chen S, Zhang Z, Liu Z. Open Source, Hidden Costs: A Systematic Literature Review on OSS License Management. IIEEE Trans Software Eng. 2025;51(9):2432-2454. doi: 10.1109/tse.2025.3586411
Bommasani R. The Foundation Model Transparency Index [Internet]. 2023. Available from: https://arxiv.org/

abs/2310.12941
Leon M. GPT-5 and open-weight large language models: Advances in reasoning, transparency, and control. Information Systems. 2026;136:102620. doi: 10.1016/j.is.2025.102620
Ait A, Cánovas Izquierdo JL, Cabot J. HFCommunity: An extraction process and relational database to analyze Hugging Face Hub data. Science of Computer Programming. 2024;234:103079. doi: 10.1016/j.scico.2024.103079
Manchanda J, Boettcher L, Westphalen M, Jasser J. The open-source advantage in large language models (LLMs). AI Ethics. 2026;6(4):390. doi: 10.1007/s43681-026-01190-4
Salazar S. Inteligencia artificial en la ingenieria de software. SAP Engineering and Industry. 2026;2:103. doi: 10.62486/ei2026103

Chapter 3 / Capítulo 3

SAP Superintelligence Series
ISSN: 3144-5624
ISBN (complete work): 978-9915-9851-0-7

Superintelligence Series Volume 5: Intelligent Visual Arts and Audiovisual Engineering: Generative AI, Computational Creativity and Superintelligence
ISBN (this volume): 978-9915-9966-5-3
DOI: 10.62486/978-9915-9966-3-9.Ch03
Pages: 24-35

Check for updates

© 2026 The authors. This is an open access book chapter distributed under the terms of the Creative Commons Attribution (CC BY) 4.0 License.

Classical no-reference image quality metrics and human judgments of artificially generated images

Métricas clásicas de calidad de imagen sin referencia y juicios humanos de imágenes generadas artificialmente

Héctor Aaron Arellano Cruz[1]
Víctor Manuel Martínez García[1]
Yennifer Díaz Romero[1]
Celso Fernando Benitez de La Herrán[1]
Rosa Edilma Garzón González[1]
Alma delia Figueroa Suarez[1]

1 Universidad Autónoma de Sinaloa, Vicerrectoría. Mazatlán, México.

ABSTRACT

Introduction: Artificially generated images are produced in large volumes, but their perceptual quality varies widely and is expensive to measure with human raters. Classical no-reference metrics are a cheap substitute whose validity for generated images is uncertain. Methods: A random sample of 600 images from the AGIQA-3K database was drawn with a fixed seed. The no-reference metrics BRISQUE, NIQE and PIQE were computed for each image and compared with the published mean opinion scores for perceived quality and for text-to-image alignment using Pearson and Spearman correlations with 95% confidence intervals. Results: All three metrics correlated negatively with perceived quality, as expected, at magnitudes between -0.317 and -0.487. NIQE showed the strongest association (r = -0.49) and PIQE the weakest (r = -0.32). Correlations with text-to-image alignment were lower, ranging from -0.239 to -0.370. The mean opinion scores for quality and alignment correlated at r = 0.824. Conclusions: Classical no-reference metrics explain only a moderate share of the variance in human judgments of generated images, with NIQE performing best among the three. They should be treated as coarse quality screens rather than substitutes for human evaluation of generated content.

Keywords: image quality assessment; no-reference metrics; AI-generated images; mean opinion score; BRISQUE; correlation

RESUMEN

Introducción: Las imágenes generadas artificialmente se producen en grandes volúmenes, pero su calidad perceptual varía mucho y es costosa de medir con evaluadores humanos. Las métricas clásicas sin referencia son un sustituto barato cuya validez para imágenes generadas es incierta. Métodos: Se extrajo una muestra aleatoria de 600 imágenes de la base AGIQA-3K con una semilla fija. Se calcularon las métricas sin referencia BRISQUE, NIQE y PIQE para cada imagen y se compararon con las puntuaciones medias de opinión publicadas para calidad percibida y alineación texto-imagen mediante correlaciones de Pearson y Spearman con intervalos de confianza del 95%. Resultados: Las tres métricas correlacionaron negativamente con la calidad percibida, como se esperaba, con magnitudes entre -0,317 y -0,487. NIQE mostró la asociación más fuerte (r = -0,49) y PIQE la más débil (r = -0,32). Las correlaciones con la alineación texto-imagen fueron menores, entre -0,239 y -0,370. Las puntuaciones medias de calidad y alineación correlacionaron en r = 0,824. Conclusiones: Las métricas clásicas sin referencia explican solo una parte moderada de la varianza de los juicios humanos sobre imágenes generadas, con NIQE como la mejor de las tres. Deben tratarse como cribados gruesos de calidad, no como sustitutos de la evaluación humana del contenido generado.

Palabras clave: evaluación de calidad de imagen; métricas sin referencia; imágenes generadas por IA; opinión media; BRISQUE; correlación

INTRODUCTION

Generative models now produce images that are difficult to distinguish from photographs, and the volume of such images has grown far faster than the capacity to judge them. Human subjective assessment remains the reference standard for perceived quality, but it is slow and expensive because it requires many raters and controlled viewing conditions under a protocol such as the ITU recommendation for subjective picture assessment.[1,2] The practical consequence is that most generated images are never evaluated by a human, and the field relies instead on automated metrics that are meant to stand in for opinion.(3) This substitution is consequential: the same automated scores are used to rank generation models, to select training data, and to decide which outputs reach users.

The cheapest and oldest of these substitutes are classical no-reference image quality metrics, which require only the image under test and no pristine reference. They occupy one end of a spectrum: full-reference metrics compare an image against an undistorted original, reduced-reference metrics compare against a compact description of the original, and no-reference metrics judge the image alone.[2,4] For generated images, no-reference assessment is the natural fit, because there is no clean original to compare against; the generated image is the object of interest, and any reference would have to be the prompt, which is text rather than an image. BRISQUE, NIQE and PIQE are the three most widely used no-reference metrics, and all three rest on a shared premise inherited from the natural scene statistics literature: a natural, undistorted image has predictable statistical regularities, and departures from those regularities signal perceptual damage.[5-8] The premise is well supported for photographs and for the familiar distortions of compression, noise, and blur, which is why the metrics remain standard tools more than a decade after they were proposed.[9,10] Their training and evaluation, however, took place on databases of naturally distorted photographs such as LIVE and TID2013, whose content is far removed from synthetic imagery.[11]

The three metrics differ in how they operationalize that shared premise, and the differences matter for how far each can be expected to generalize. BRISQUE learns a mapping from feature statistics to a quality score using a support vector regressor trained on a corpus of images distorted in known ways, which makes it implicitly tuned to the distortion types in that corpus.(5) NIQE avoids training altogether by fitting a multivariate Gaussian to features computed from a corpus of pristine images and scoring new images by their distance from that pristine model, a design its authors call completely blind.(6) PIQE takes a more direct route, segmenting the image into blocks, estimating a local distortion for each block from its contrast statistics, and averaging the result.(7) All three were validated on natural photographs with synthetic distortions, and none was designed with the artifacts of generative models in mind, so the question here is not whether the metrics are sound but whether their assumptions survive the transfer to a new kind of content. (4,9)

Artificially generated images break that premise. Their defects are not the classic distortions but a different class of artifacts, including implausible hands and faces, semantic inconsistencies, and stylistic peculiarities that a natural-scene-statistics model was never designed to recognize. [12,13] Dedicated databases such as AGIQA-3K, AIGCIQA2023 and PKU-AIGIQA-4K were assembled precisely because generated images behave differently under conventional quality models. (12,14,15) Whether the classical no-reference metrics still track human opinion in this new setting is an open empirical question, and the few studies that address it directly are recent and limited in scope.(16)

The question has implications beyond the image processing community. Generated imagery is now embedded in the same pipelines that power advanced artificial intelligence systems, where it serves as training data, as synthetic evaluation data, and as user-facing output, and the ability to filter or rank that imagery automatically determines what those systems learn and what they expose to users.[3,13] If the classical metrics that practitioners reach for by default do not track human judgment, then quality control throughout these pipelines rests on a false assumption, and the consequences propagate to every downstream system that consumes the generated images. Establishing how well those metrics actually predict human opinion is therefore a prerequisite for any credible quality assurance of generative output.

The contribution of this study is an estimate of exactly that validity, obtained on a database built for the purpose and with the metrics computed under a documented, reproducible procedure. Rather than another proposal for a better metric, the study asks whether the tools already in wide use are adequate, and it answers with correlations, confidence intervals, and explained-variance figures rather than with qualitative judgments. The objective of this study was

https://doi.org/10.62486/978-9915-9966-3-9.Ch03

to quantify how strongly BRISQUE, NIQE and PIQE correlate with human mean opinion scores for generated images, using a documented random sample from the AGIQA-3K database.

METHODS

STUDY DESIGN

A cross-sectional correlational study was conducted in which three classical no-reference image quality metrics were computed on a sample of artificially generated images and compared with the mean opinion scores that the publishing database had collected for those same images. The study followed the STROBE reporting framework for observational research. No new human judgments were collected; the design reuses published opinion scores and pairs them with newly computed metric values. This secondary-analysis design trades control over the rating procedure for scale and reproducibility: it would be impractical to re-collect opinion scores at the same breadth, and reusing the published scores keeps the study focused on the metric behavior, which is the object of interest.

DATA SOURCE AND SAMPLE

Images and opinion scores were taken from AGIQA-3K, an open database of 2,982 artificially generated images produced by six generative models, each scored by human raters on perceived quality and on text-to-image alignment.[12] The database was chosen because it is openly licensed, because it separates quality from alignment, and because it spans several generator families rather than a single model, which makes it a reasonable substrate for judging whether a metric generalizes. The database is distributed under the MIT license and includes a comma-separated file with one row per image containing the image identifier, the generation prompt, and the normalized mean opinion scores and their standard deviations for the two constructs. The image files themselves were downloaded from a public mirror of the database.

A random sample of 600 images was drawn with a fixed seed so that the analysis is reproducible; this size yields a margin of error below 0.04 for a Pearson correlation of moderate magnitude, and it keeps the metric computation within the reach of a central processing unit. The sample is large enough that the confidence intervals reported in the results are dominated by the strength of the underlying association rather than by sampling noise, which is what makes the moderate correlations a substantive finding rather than a consequence of imprecision. The six source generators were represented in the sample in proportion to their presence in the full database, so the correlation estimates weight each generator roughly as the full database does.

VARIABLES

The outcome variables were the two mean opinion scores provided by the database: perceived quality and text-to-image alignment, both normalized so that their scale is comparable across generators. The exposure variables were the three no-reference metric scores computed on each image: BRISQUE, NIQE and PIQE. For all three metrics a higher score indicates lower predicted quality, so a negative correlation with opinion scores is the expected direction. The image source generator was recorded as a descriptive covariate. No additional variables were derived from the images beyond the metric scores, because the objective was to evaluate the metrics as they are conventionally computed, without auxiliary features that a practitioner would not have.

MEASUREMENT OF NO-REFERENCE QUALITY

BRISQUE transforms an image into locally normalized luminance coefficients and scores them with a support vector regressor trained on natural and distorted photographs.(5) NIQE fits the same coefficients to a generalized Gaussian model and measures the distance from a model fitted to a corpus of pristine natural images.[6] PIQE divides the image into blocks, classifies each block as noisy, blurry, or clean from local contrast statistics, and aggregates the distorted-block scores.(7) These three span the main design choices in classical blind quality assessment: a

https://doi.org/10.62486/978-9915-9966-3-9.Ch03

trained regressor, a pristine-model distance, and a hand-crafted blockwise heuristic, respectively. (4) All three operate on grayscale luminance, discarding color information, which is a further simplification given that color is part of perceived quality. The three metrics were computed with the pyiqa implementation, which reproduces the original feature and model definitions and runs on a central processing unit. Metric computation was deterministic and was verified by recomputation, with identical results on a repeated run. Because the metrics have no reference image, their scores are absolute rather than relative, and they are typically interpreted as a deviation from naturalness rather than as an error against a specific ground truth.

STATISTICAL ANALYSIS

The association between each metric and each opinion score was estimated with the Pearson product-moment correlation and the Spearman rank correlation, together with the 95% confidence interval for the Pearson coefficient obtained through the Fisher transformation. Pearson captures linear association and is the quantity conventionally reported in image quality studies; Spearman is added because the metrics are not normally distributed and because it is robust to the monotone nonlinearities that arise when a metric saturates at the extremes of the quality range. The coefficient of determination was reported as the squared Pearson correlation to express the share of variance in human opinion explained by each metric. Because the analysis compares three metrics against the same opinion scores, the differences between their correlations were interpreted descriptively rather than tested formally; a formal comparison would require a correction for the strong dependence among the three metrics, which share a common feature basis. All analyses were performed with Python and the SciPy statistics module, and the analysis script was executed twice to confirm that the results are reproducible.

RESULTS

SAMPLE CHARACTERISTICS

The sample comprised 600 images from 6 generative models (Table 1). The mean opinion score for perceived quality was 2.52 with a standard deviation of 0.98, and the alignment score was nearly identical at 2.52 with a standard deviation of 0.97. The three metrics spanned their expected ranges: BRISQUE averaged 35.8 (SD 24.3), NIQE 5.79 (SD 2.99), and PIQE 45.4 (SD 15.5). The dispersion of the metrics was considerable relative to their means, with BRISQUE in particular showing a standard deviation of 24.3 around a mean of 35.8, which reflects the wide range of image quality present in the sample and confirms that the metrics are not saturated at a single value. The two opinion scores were strongly correlated with each other, at r = 0.82, which indicates that perceived quality and prompt alignment overlap but are not the same construct, and justifies treating them as separate outcomes.

Table 1. Descriptive statistics of opinion scores and no-reference metrics in the sample.

Variable	Mean	SD
MOS quality	2.52	0.976
MOS alignment	2.521	0.971
BRISQUE	35.786	24.3
NIQE	5.789	2.989
PIQE	45.369	15.462
Images, n	600	

Note: SD = standard deviation; n = 600 images from six generative models.

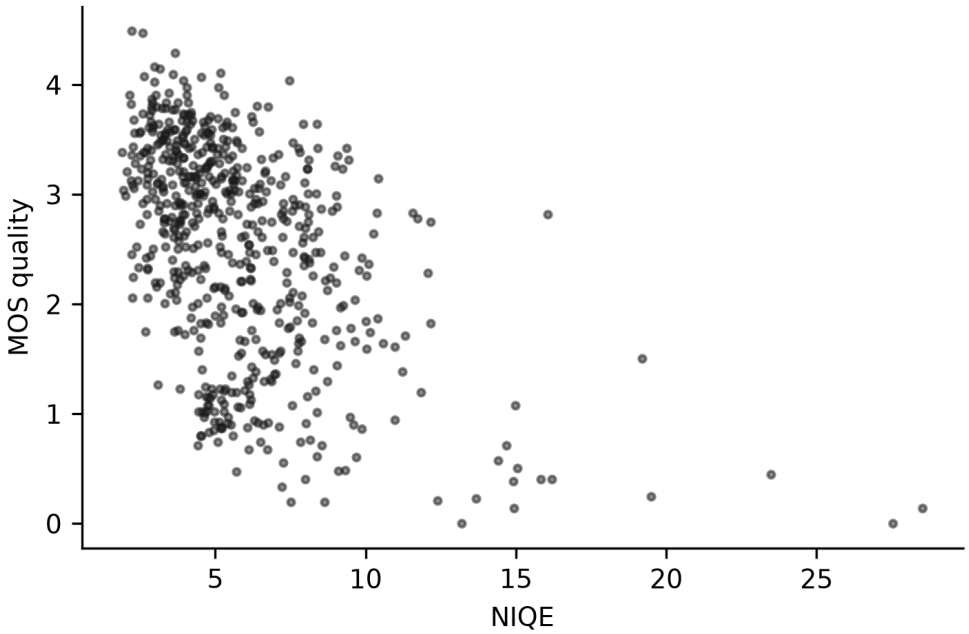
CORRELATION WITH PERCEIVED QUALITY

All three metrics correlated negatively with perceived quality, as expected, and all associations were significant at $p < 0.001$ (Table 2). NIQE showed the strongest association, with a Pearson correlation of −0.49 (95% CI −0.55 to −0.42) and a Spearman correlation of −0.47. BRISQUE

https://doi.org/10.62486/978-9915-9966-3-9.Ch03

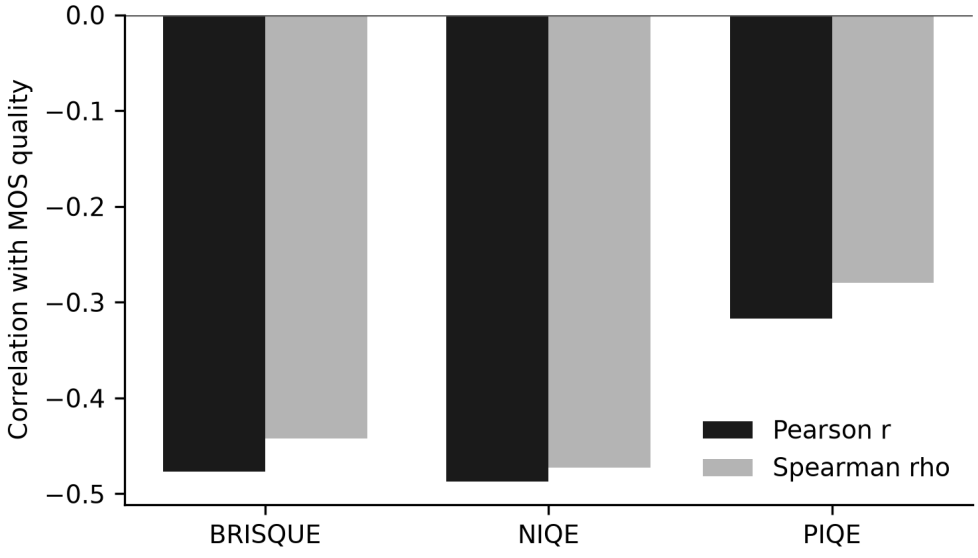
was nearly as strong, at r = −0.48 (95% CI −0.54 to −0.41) and rho = −0.44. PIQE was the weakest of the three, at r = −0.32 (95% CI −0.39 to −0.24) and rho = −0.28. In terms of explained variance, NIQE accounted for roughly one quarter of the variation in human quality judgments, BRISQUE slightly less, and PIQE about one tenth. The confidence intervals of NIQE and BRISQUE overlap, so their difference should be treated as small and possibly not replicable, whereas the interval of PIQE is clearly separated from both, which identifies PIQE as the weakest metric by a meaningful margin. The relationship between the best-performing metric and the quality scores is shown in (Figure 1), and the three metrics are compared side by side in (Figure 2).

Figure 1. Scatter plot of NIQE against mean opinion scores for perceived quality.



Note: Each point is one image; higher NIQE values indicate lower predicted quality.

Figure 2. Correlation of each no-reference metric with mean opinion scores for perceived quality.



Note: Bars show the Pearson r and Spearman rho for BRISQUE, NIQE and PIQE.

The correlations were negative in every case, as the metric definitions require: a higher BRISQUE, NIQE or PIQE score signals a larger departure from naturalness and therefore a lower predicted quality. None of the metrics reversed this sign, and none produced a near-zero association, which confirms that the natural-scene-statistics signal, though attenuated, is not absent from generated images. What is absent is the strong linear relationship that these metrics

https://doi.org/10.62486/978-9915-9966-3-9.Ch03

exhibit on natural distortions; the scatter is wider here, with the best metric leaving the majority of the variance in human opinion unexplained.

Table 2. Correlation of no-reference metrics with mean opinion scores for quality and alignment.

Metric	r (MOS quality)	95% CI	rho (MOS quality)	r (MOS alignment)	rho (MOS alignment)
BRISQUE	-0.477	-0.537 to -0.413	-0.442	-0.348	-0.305
NIQE	-0.487	-0.546 to -0.424	-0.473	-0.37	-0.316
PIQE	-0.317	-0.387 to -0.243	-0.28	-0.239	-0.201

Note: CI = confidence interval; r = Pearson correlation; rho = Spearman correlation; all correlations $p < 0.001$.

CORRELATION WITH TEXT-TO-IMAGE ALIGNMENT

The metrics tracked text-to-image alignment less closely than perceived quality (Table 2). NIQE again led, with $r = -0.37$ and rho = −0.32, followed by BRISQUE at $r = -0.35$ and PIQE at $r = -0.24$. The rank order of the three metrics was the same for alignment as for quality, but every correlation was lower in absolute value, which indicates that the metrics capture perceptual fidelity better than semantic correspondence between prompt and image. The reduction was most pronounced for the metric with the weakest overall signal, PIQE, whose alignment correlation fell to −0.24, and smallest for NIQE, which retained an alignment correlation of −0.37. The strong correlation between the two opinion constructs ($r = 0.82$) means that part of the alignment association is carried through shared variance with quality, but the consistently lower alignment coefficients show that the metrics do not reach the semantic content of the images.

Taken together, the results describe a consistent picture rather than a set of isolated coefficients. Every metric was negative with every outcome, the ranking of the metrics was identical across outcomes, and the quality coefficients exceeded the alignment coefficients in every case. This internal consistency argues that the estimates reflect a stable property of the metrics rather than sampling variation, and it is the stability of the ranking, more than the exact magnitudes, that is most likely to replicate in other datasets.

DISCUSSION

MAIN FINDINGS

All three classical no-reference metrics predicted human judgments of generated images in the expected direction, but only to a moderate degree. NIQE was the strongest, explaining roughly one quarter of the variance in perceived-quality scores, BRISQUE was a close second, and PIQE was markedly weaker, explaining about one tenth. The same ranking held for text-to-image alignment, where every correlation was lower still. No metric approached the strength of agreement that would justify replacing human raters; the best of the three is at most a coarse screen for perceived quality. The negative sign of every correlation is itself a meaningful result, because it confirms that the metrics are not simply random on this content and that their naturalness signal carries a genuine, if partial, correspondence to what humans see. The finding is at once reassuring and cautionary: the classical metrics are not useless on generated images, but they are not sufficient either.

INTERPRETATION AND COMPARISON WITH OTHER STUDIES

The moderate correlations match what the few recent studies that have tested classical metrics on generated content reported. Sridhar and colleagues found that non-reference metrics perform inconsistently on generated images and argued for task-specific scoring, a conclusion this study supports with a larger and better-controlled sample.(16) Chen and colleagues likewise found that subjective studies and objective metrics diverge for generated images and that conventional metrics underrepresent the defects raters actually notice.(13) The present estimate of a Pearson correlation around 0.49 for NIQE places the classical metrics below the learned models that have since been developed for generated images, which typically reach correlations above 0.8 on the same databases.(17-19)

https://doi.org/10.62486/978-9915-9966-3-9.Ch03

The comparison with the classical metrics' own historical performance is even starker. On databases of naturally distorted photographs, BRISQUE and its contemporaries were typically reported with Spearman rank correlations above 0.9, and even the simpler DCT-based predecessors reached the 0.8 range.[11,20] The drop from those levels to the values observed here, on the order of 0.28 to 0.47, quantifies how much of the metrics' validity was tied to the natural-distortion setting in which they were designed. This is not a criticism of the metrics, which were never claimed to cover synthetic content, but it is a reminder that a metric validated in one domain cannot be assumed to transfer to another, and that the transfer gap can be large enough to change a decision from reliable to unreliable.(2,4)

The gap between classical and learned metrics has a clear mechanistic explanation. BRISQUE, NIQE and PIQE encode a statistical model of natural photographs, and their features respond to the luminance and contrast disruptions that characterize compression, noise, and blur.[10,20] Generated images carry a different class of artifacts, including anatomically implausible details and semantic inconsistencies that do not perturb natural scene statistics in the same way.[12,14] A metric that measures distance from a pristine natural photograph therefore registers only part of what makes a generated image look wrong. The lower correlations with alignment follow the same logic: alignment depends on semantic content that a purely statistical metric cannot see at all.(13,15)

This mechanistic account also explains the broader trajectory of the field. The hand-crafted natural-scene-statistics features that define the classical era gave way to learned features extracted by deep networks precisely because the learned representations could capture semantic and perceptual structure that the statistical features could not.[21,22] The classical metrics remain useful because they need no training data and no reference, but their representational ceiling was reached early, and the results here show where that ceiling lies for synthetic content. In this sense the study is less an evaluation of three particular algorithms than a measurement of the limits of an entire, and once dominant, approach to quality assessment. Text-to-image alignment is evaluated by comparing the image against a text prompt, which requires a joint understanding of language and vision that is fundamentally outside the scope of a luminance-statistics metric. The fact that the alignment correlations are nevertheless nonzero is explained by the overlap between the two opinion constructs rather than by any semantic capability of the metrics: an image that is severely degraded perceptually is also, in practice, a poor match to its prompt, so part of the alignment signal rides along with quality. Once that shared component is accounted for, the residual alignment signal, which is the semantic part, is essentially invisible to all three metrics.

The finding that NIQE outperforms BRISQUE and PIQE is also interpretable. NIQE is the most purely statistical of the three, fitting a general model of pristine natural images without a distortion-specific training stage, whereas BRISQUE is trained on a specific distortion corpus and PIQE relies on a local blockwise heuristic.(5-7) On generated images, the distortion-specific assumptions of BRISQUE and the local heuristic of PIQE appear to generalize less well than NIQE's broader naturalness measure, which is consistent with the opinion-unaware philosophy that NIQE introduced and that later deep models adopted.[23,24] The result is also a useful corrective to the default of using BRISQUE or PIQE in applied settings: in the domain of generated images, the metric chosen matters, and the common assumption that all no-reference metrics behave alike does not hold here.

There is also a ceiling on what any of these metrics can achieve, independent of their individual designs. All three read a single image and score it against a statistical model of naturalness, which means they can only see the properties that naturalness captures. Human raters, by contrast, judge generated images on semantic plausibility, on whether the content matches the prompt, and on aesthetic qualities that have little to do with natural scene statistics.[12,25] A metric that cannot see these properties cannot predict them, which places an upper bound on the attainable correlation that is well below what learned models, which ingest semantic features, can reach.[17,18] The moderate correlations reported here should therefore be read not as a failure of the classical metrics but as a statement of the ceiling imposed by their representational assumptions.

GENERALIZABILITY

Within the limits of a single database, the results are likely to generalize to the broader problem of evaluating generated images with classical tools, because the underlying mismatch between natural-scene-statistics models and synthetic artifacts is a property of the metrics,

https://doi.org/10.62486/978-9915-9966-3-9.Ch03

not of the particular generators sampled.[8,12] The same mismatch has been observed outside computer vision, where classical no-reference metrics were applied to medical images produced by generative adversarial networks and found to need human validation.(28) The practical implication for practitioners is straightforward: BRISQUE, NIQE and PIQE can rank a batch of generated images roughly by quality, and NIQE should be preferred when only a classical metric is available, but none should be reported as a measure of human-perceived quality or of prompt alignment.

For research and industrial pipelines that require reliable quality assessment, learned metrics trained on generated-image databases remain the appropriate tool, and even aesthetic preference is better captured by learned scores tuned to human preference than by statistical naturalness.(21,22,25,29,30) The classical metrics retain value as reference-free, training-free baselines against which those learned models can be benchmarked, a role they have played since the early deep quality models appeared.[22,30] A sensible division of labor follows from the results: classical metrics can provide a first, inexpensive pass that flags the worst generations and orders a batch for human review, while learned metrics and human raters handle the decisions that require semantic or aesthetic judgment. The findings also carry a caution for the many applied studies that report BRISQUE or PIQE as a quality indicator for generated content: such numbers describe distance from naturalness, not quality as a human would perceive it, and conflating the two will mislead any downstream decision that depends on quality.

A concrete recommendation follows from the ranking. When a classical no-reference metric is needed and only one can be run, NIQE is the justified choice for generated images on the evidence here, because it outperformed the alternatives on both outcomes and does not rely on a distortion corpus that may not match the content. When computation or interpretability favors a blockwise score, PIQE remains the weakest option and should be treated accordingly. These recommendations are conditional on the database and generators studied, but the consistency of the ranking across the two outcome constructs gives more confidence in it than a single comparison would.

IMPLICATIONS FOR QUALITY ASSURANCE OF GENERATIVE SYSTEMS

The results bear directly on how generated content should be quality-controlled in production systems. A pipeline that uses BRISQUE or PIQE as a filter passes and rejects images on a signal that explains only a small fraction of human quality judgment, so many poor images will pass and many good ones will be rejected. The stronger but still moderate performance of NIQE does not change this conclusion; it only makes NIQE the least bad choice among classical options. For systems whose output is user-facing, the findings argue for reserving human or learned-model judgment for the final acceptance decision and using classical metrics only as an inexpensive pre-filter that flags gross defects.

The implications extend to the governance of generated content. If automated quality metrics are used to decide whether generated images meet a quality bar, whether they are suitable for distribution, or whether they should be labeled as synthetic, then the validity of those metrics is a policy question as much as a technical one. A metric that systematically misses semantic defects would give a false assurance of quality, and a metric tuned to naturalness would penalize stylized but perfectly adequate generations. The findings here suggest that any governance framework relying on classical no-reference scores would be relying on a proxy that human judgment does not fully endorse, and that such frameworks should either adopt learned, semantically aware metrics or keep a human in the loop.[3,29] The classical metrics, in short, are useful instruments whose scope is narrower than the uses to which they are often put, and recognizing that scope is the first step toward deploying them responsibly.

CONCLUSIONS

The classical no-reference metrics BRISQUE, NIQE and PIQE correlated negatively and significantly with human mean opinion scores of generated images, confirming that they retain some validity beyond natural photographs. NIQE was the strongest predictor of perceived quality, followed closely by BRISQUE, while PIQE was substantially weaker, and all three tracked text-to-image alignment less well than perceptual quality. None of the three explained more than a quarter of the variance in human judgments, so they should be used as coarse, training-free

https://doi.org/10.62486/978-9915-9966-3-9.Ch03

quality screens rather than as substitutes for human evaluation of generated content. The findings establish both the continued usefulness and the clear ceiling of the natural-scene-statistics approach, and they point to learned, semantically aware metrics as the appropriate tool wherever quality assessment must track human perception closely.

https://doi.org/10.62486/978-9915-9966-3-9.Ch03

REFERENCES

International Telecommunication Union. Methodology for the subjective assessment of the quality of television pictures (Recommendation ITU-R BT.500-14) [Internet]. 2019 [cited 2026 Oct 25]. Available from: https://www.itu.int/rec/R-REC-BT.500
Zhai G, Min X. Perceptual image quality assessment: a survey. Sci China Inf Sci. 2020;63(11):211301. doi: 10.1007/s11432-019-2757-1
Zhang Z, Zhou Y, Li C, Zhao B, Liu X, Zhai G. Quality Assessment in the Era of Large Models: A Survey. ACM Trans Multimedia Comput Commun Appl. 2025;21(7):1-31. doi: 10.1145/3722559
Shahid M, Rossholm A, Lövström B, Zepernick HJ. No-reference image and video quality assessment: a classification and review of recent approaches. J Image Video Proc. 2014;2014(1):40. doi: 10.1186/1687-5281-2014-40
Mittal A, Moorthy AK, Bovik AC. No-Reference Image Quality Assessment in the Spatial Domain. IEEE Trans on Image Process. 2012;21(12):4695-4708. doi: 10.1109/tip.2012.2214050
Mittal A, Soundararajan R, Bovik AC. Making a "Completely Blind" Image Quality Analyzer. IEEE Signal Process Lett. 2013;20(3):209-212. doi: 10.1109/lsp.2012.2227726
Venkatanath N, Praneeth D, Maruthi Chandrasekhar Bh, Channappayya SS, Medasani SS. Blind image quality evaluation using perception based features. 2015 Twenty First National Conference on Communications (NCC). 2015:1-6. doi: 10.1109/ncc.2015.7084843
Ruderman DL. The statistics of natural images. Network: Computation in Neural Systems. 1994;5(4):517-548. doi: 10.1088/0954-898x_5_4_006
Saad MA, Bovik AC, Charrier C. Blind Image Quality Assessment: A Natural Scene Statistics Approach in the DCT Domain. IEEE Trans on Image Process. 2012;21(8):3339-3352. doi: 10.1109/tip.2012.2191563
Moorthy AK, Bovik AC. Blind Image Quality Assessment: From Natural Scene Statistics to Perceptual Quality. IEEE Trans on Image Process. 2011;20(12):3350-3364. doi: 10.1109/tip.2011.2147325
Ponomarenko N, Jin L, Ieremeiev O, Lukin V, Egiazarian K, Astola J, et al. Image database TID2013: Peculiarities, results and perspectives. Signal Processing: Image Communication. 2015;30:57-77. doi: 10.1016/j.image.2014.10.009
Li C, Zhang Z, Wu H, Sun W, Min X, Liu X, et al. AGIQA-3K: An Open Database for AI-Generated Image Quality Assessment. IEEE Trans Circuits Syst Video Technol. 2024;34(8):6833-6846. doi: 10.1109/tcsvt.2023.3319020
Chen J, Shao F, Chen H, Wang X, Guo H, Jiang Q. Quality Evaluation of AI-Generated Images: Subjective Study and Objective Methodology. IEEE Trans Multimedia. 2026;28:57-70. doi: 10.1109/tmm.2025.3618562
Wang J. AIGCIQA2023: A Large-scale Image Quality Assessment Database for AI Generated Images: from the Perspectives of Quality, Authenticity and Correspondence [Internet]. 2023 [cited 2026 Oct 25]. Available from: https://arxiv.org/abs/2307.00211
Yuan J. PKU-AIGIQA-4K: A Perceptual Quality Assessment Database for Both Text-to-Image and Image-to-Image AI-Generated Images [Internet]. 2024 [cited 2026 Oct 25]. Available from: https://arxiv.org/abs/2404.18409
Sridhar Y, Gajendra JM, Srinivasalu P, Srinivasappa M. Evaluating Non-Reference Image Quality Metrics for AI-Generated Images: A Novel Approach. EJAI. 2025;4(5):18-25. doi: 10.24018/ejai.2025.4.5.70
Zhou T, Tan S, Zhou W, Luo Y, Wang YG, Yue G. Adaptive Mixed-Scale Feature Fusion Network for Blind AI-Generated Image Quality Assessment. IEEE Trans on Broadcast. 2024;70(3):833-843. doi: 10.1109/tbc.2024.3391060
Fu J, Zhou W, Jiang Q, Liu H, Zhai G. Vision-Language Consistency Guided Multi-Modal Prompt Learning for Blind AI Generated Image Quality Assessment. IEEE Signal Process Lett. 2024;31:1820-1824. doi: 10.1109/lsp.2024.3420083
Wang J, Duan H, Zhai G, Min X. Quality Assessment for AI Generated Images With Instruction Tuning. IEEE Trans Multimedia. 2026;28:2404-2419. doi: 10.1109/tmm.2026.3651009
Saad MA, Bovik AC, Charrier C. A DCT Statistics-Based Blind Image Quality Index. IEEE Signal Process Lett. 2010;17(6):583-586. doi: 10.1109/lsp.2010.2045550
Yang X, Li F, Liu H. A Survey of DNN Methods for Blind Image Quality Assessment. IEEE Access. 2019;7:123788-123806. doi: 10.1109/access.2019.2938900
Bosse S, Maniry D, Muller KR, Wiegand T, Samek W. Deep Neural Networks for No-Reference and Full-Reference Image Quality Assessment. IEEE Trans on Image Process. 2018;27(1):206-219. doi: 10.1109/tip.2017.2760518
Ni Z, Liu Y, Ding K, Yang W, Wang H, Wang S. Opinion-Unaware Blind Image Quality Assessment Using Multi-Scale Deep Feature Statistics. IEEE Trans Multimedia. 2024;26:10211-10224. doi: 10.1109/tmm.2024.3405729
Sinno Z, Caramanis C, Bovik AC. Towards a Closed Form Second-Order Natural Scene Statistics Model. IEEE Trans on Image Process. 2018;27(7):3194-3209. doi: 10.1109/tip.2018.2817740
Li D. Playground v2.5: Three Insights towards Enhancing Aesthetic Quality in Text-to-Image Generation [Internet]. 2024 [cited 2026 Oct 25]. Available from: https://arxiv.org/abs/2402.17245
Pin SHD, Amirshahi SA. Individual Differences in Subjective Image Quality Assessment: A Bayesian Mixed-Effects Modeling Approach. IEEE Access. 2025;13:127504-127517. doi: 10.1109/access.2025.3579016
Min X, Gao Y, Cao Y, Zhai G, Zhang W, Sun H, et al. Exploring Rich Subjective Quality Information for Image Quality Assessment in the Wild. IEEE Trans Circuits Syst Video Technol. 2025;35(8):7778-7791. doi: 10.1109/tcsvt.2025.3544659
Treder MS, Codrai R, Tsvetanov KA. Quality assessment of anatomical MRI images from generative adversarial networks: Human assessment and image quality metrics. Journal of Neuroscience Methods. 2022;374:109579.

https://doi.org/10.62486/978-9915-9966-3-9.Ch03

doi: 10.1016/j.jneumeth.2022.109579
An B, Zhou H, Peng P, Zhang L, Ren S, Li J, et al. A survey of learning-based no-reference image quality assessment. The International Conference Optoelectronic Information and Optical Engineering (OIOE2024). 2025;13513:124. doi: 10.1117/12.3045771
Hosu V, Lin H, Sziranyi T, Saupe D. KonIQ-10k: An Ecologically Valid Database for Deep Learning of Blind Image Quality Assessment. IEEE Trans on Image Process. 2020;29:4041-4056. doi: 10.1109/tip.2020.2967829

https://doi.org/10.62486/978-9915-9966-3-9.Ch03

Chapter 4 / Capítulo 4

SAP Superintelligence Series
ISSN: 3144-5624
ISBN (complete work): 978-9915-9851-0-7

Superintelligence Series Volume 5: Intelligent Visual Arts and Audiovisual Engineering: Generative AI, Computational Creativity and Superintelligence
ISBN (this volume): 978-9915-9966-5-3
DOI: 10.62486/978-9915-9966-3-9.Ch04
Pages: 36-46

Check for updates

© 2026 The authors. This is an open access book chapter distributed under the terms of the Creative Commons Attribution (CC BY) 4.0 License.

Frequency-domain detectors for machine-generated images and their generalization across generators

Detectores en el dominio de la frecuencia para imágenes generadas por máquina y su generalización entre generadores

Vugar Hacimahmud Abdullayev[1]

1 Azerbaijan University of Architecture and Construction. Baku, Azerbaijan

ABSTRACT

Introduction: Machine-generated images are now easy to produce and hard to distinguish from photographs, and reliable detection is needed for forensics and content moderation. Deep detectors work well on familiar generators but degrade on unseen ones; cheap frequency-domain features are an alternative whose cross-generator behavior is unclear. Methods: A sample of 500 real photographs and 500 images from each of three generators (BigGAN, GLIDE, VQDM) was drawn from the GenImage benchmark. Fifty-nine frequency-domain features were extracted from each image and fed to a logistic regression trained on one generator, then evaluated in-domain and on the two held-out generators with accuracy and area under the curve. Results: Mean in-domain accuracy was 0.77 and fell to 0.65 across generators, an average drop of 12 percentage points. The drop was uneven, with GLIDE reaching 0.92 in-domain while most cross-generator transfers fell below 0.66, and one transfer was weaker than a coin toss. Conclusions: Frequency-domain features support lightweight detectors that work within a generator but transfer poorly to unseen generators, because they capture generator-specific spectral artifacts. They are a useful first-line screen when a detector must run without a learned model, but not a substitute for generators-diverse training.

Keywords: synthetic image detection; frequency domain; image forensics; generalization; generative models; machine learning

RESUMEN

Introducción: Las imágenes generadas por máquina son fáciles de producir y difíciles de distinguir de las fotografías, y se necesita una detección fiable para la forensía y la moderación de contenidos. Los detectores profundos funcionan bien con generadores conocidos pero se degradan con los nuevos; las características de frecuencia son una alternativa barata cuyo comportamiento entre generadores no está claro. Métodos: Se extrajo una muestra de 500 fotografías reales y 500 imágenes de cada uno de tres generadores (BigGAN, GLIDE, VQDM) del benchmark GenImage. Se extrajeron 59 características de frecuencia de cada imagen y se alimentó una regresión logística entrenada en un generador, evaluada en dominio y en los dos generadores excluidos con exactitud y área bajo la curva. Resultados: La exactitud media en dominio fue 0,77 y cayó a 0,65 entre generadores, una caída media de 12 puntos porcentuales. La caída fue desigual, pues GLIDE alcanzó 0,92 en dominio mientras que la mayoría de las transferencias entre generadores quedaron por debajo de 0,66 y una fue peor que el azar. Conclusiones: Las características de frecuencia permiten detectores ligeros que funcionan dentro de un generador pero transfieren mal a generadores no vistos, porque capturan artefactos espectrales específicos de cada generador. Son un cribado inicial útil cuando no se dispone de un modelo aprendido, pero no sustituyen al entrenamiento con generadores diversos.

Palabras clave: detección de imágenes sintéticas; dominio de la frecuencia; forensía de imagen; generalización; modelos generativos; aprendizaje automático

INTRODUCTION

Machine-generated images have moved from a laboratory curiosity to a widely available commodity in a few short years, and the practical problem they create is simple to state: an observer can no longer reliably tell whether a picture was captured by a camera or produced by a model.[1,2] That uncertainty matters deeply wherever images serve as evidence, whether in journalism, in courts, in financial fraud, or in the content that online platforms distribute at scale. A detector that can tell the two apart is therefore a basic instrument of trust in visual media, and a great deal of effort has gone into building such detectors.(3,4)

The need is not a single need but a spectrum of them, and the spectrum shapes what a good detector looks like. A newsroom verifying a contested photograph wants a decision it can explain to an editor, and an explanation that cites a spectral anomaly is far more usable than one that cites a deep feature vector. A platform screening millions of uploads wants a detector cheap enough to run on every image. A forensic examiner wants a method whose reasoning survives scrutiny in a report. These requirements pull in different directions, and the deep networks that dominate the literature satisfy only the accuracy requirement, leaving the cheaper and more interpretable alternatives under-explored as a consequence.(5) It is in that gap that the frequency-domain approach sits.

The strongest detectors are deep networks trained on large sets of real and generated images, and they work well when they are asked about the same generators they were trained on. Wang and colleagues demonstrated this pattern early, showing that a detector trained on one generator family recognizes those images easily while its performance on other families falls away.(6) The difficulty is precisely generalization: a detector is useful only if it survives the arrival of new generators, and the arrival is constant.[7,8] The GenImage benchmark was created to measure this property at scale, and its results show that the leading deep detectors degrade substantially when moved from a seen to an unseen generator.(9) Subsequent work has refined the picture by showing which features survive the transfer and which do not.(5,10)

The problem is made harder by an arms race. As detectors have improved, so have the methods for erasing the traces they rely on, and a detector that can be defeated by a simple fingerprint-removal step is of limited forensic value in a real investigation.[11,12] The same generators that produce convincing images are being tuned to leave fewer detectable artifacts, which means any detection approach must be evaluated not only against current generators but against the expectation that its signal will be actively attacked. This is the context in which a lightweight, transparent detector is attractive: it is cheap enough to run everywhere, and when it fails, its failure is at least legible.(1)

Not every deployment can afford a deep network. In many settings a lightweight, interpretable detector is what is actually wanted: one that runs on a central processing unit, needs no training data beyond a handful of images, and exposes the reason for its decision. Frequency-domain features fit that description. Generated images carry spectral fingerprints left by the upsampling layers of generative networks, and these fingerprints are visible in the Fourier and discrete-cosine transforms of an image.[13,14] The idea behind frequency-domain detection is to summarize those fingerprints with a few dozen scalar features and hand them to a simple linear classifier.[14,15] The approach is cheap and auditable, which is why it remains attractive despite the dominance of deep methods.(16,17)

The frequency approach also spans a wide design space, and the literature has explored it in several directions. Some methods fuse spatial and spectral features to recover the information that either domain alone misses.[18,19] Others work entirely in the spectral domain, using wavelet decompositions or training-free measurements of the diffusion process itself.[17,20] The common thread is the hope that a compact spectral signature exists and can be read off without a learned model. What most of these studies do not isolate is the specific question of cross-generator transfer, because they typically report performance on the same generators used in training. That is the gap the present study fills.

The design chosen here is deliberately minimal. Rather than propose yet another feature set, the study takes two standard, off-the-shelf frequency representations, pairs them with the simplest classifier that can exploit them, and asks a single, sharply defined question: how much accuracy is lost when the generator changes. The answer is reported as a full transfer matrix rather than a single number, because, as the results will show, the loss is a property of the specific generator pair and a single number would conceal it. The intent is to produce a baseline that other, more elaborate frequency and hybrid methods can be compared against, and a number that practitioners can use directly when they decide whether a lightweight detector is good

https://doi.org/10.62486/978-9915-9966-3-9.Ch04

enough for their problem.

What is not settled is whether a frequency-domain detector generalizes across generators, and if it fails, by how much. The spectral fingerprints that the features exploit are known to be generator-specific, which suggests that a frequency detector trained on one generator may learn that generator's signature rather than a universal property of synthetic imagery.[13,19] If so, its cross-generator performance would fall, and the magnitude of the fall would tell practitioners how much to trust such a detector in the open world. The objective of this study was therefore to evaluate the cross-generator generalization of lightweight detectors that use frequency-domain features, measuring the drop in accuracy and area under the curve when a detector trained on one generator is tested on generators excluded from training.

METHODS

DESIGN

A diagnostic study was conducted in which frequency-domain features were extracted from real photographs and from machine-generated images, a lightweight linear classifier was trained on one generator, and its performance was measured both in-domain and on held-out generators. The design follows the STARD framework for diagnostic accuracy studies, with the generator family as the condition whose transfer is being tested. The index test is the frequency-domain logistic classifier; the reference standard is the known provenance of each image, which the benchmark records directly and without ambiguity. The analysis is fully reproducible from the public benchmark and a fixed random seed, and no human judgment enters the pipeline, so the reference standard is exact rather than an approximation.

DATA

Images were taken from GenImage, a million-scale benchmark that pairs real photographs from ImageNet with images generated by eight distinct models.(9) Three generators were selected to span a range of different architectures: BigGAN, a class-conditional generative adversarial network; GLIDE, a diffusion model; and VQDM, a vector-quantized diffusion model. For each generator the benchmark provides a folder of real photographs and a separate folder of generated images matched to the same class structure. A fixed-seed random sample of 500 real photographs and 500, 500, and 500 generated images per generator was drawn, yielding a balanced working set of two thousand images (Table 1). All images were converted to grayscale and then resized to a common resolution of 128 by 128 pixels before feature extraction.

Table 1. Composition of the working sample by class and generator.

Class	Images
Real photographs	500
biggan	500
glide	500
vqdm	500
Feature dimensions	59

Note: A fixed random seed was used for the sample; all images were resized to 128 by 128 pixels.

FREQUENCY-DOMAIN FEATURES

Two families of frequency-domain features were computed for each image, producing a 59-dimensional vector. The first family summarizes the Fourier magnitude spectrum. The two-dimensional Fourier transform was computed, the logarithm of its magnitude was averaged over concentric rings centered on the zero frequency, and the resulting radial profile of 32 bins was normalized to sum to one. This profile captures how energy is distributed across spatial frequencies, and it is the natural place where the periodic grid artifacts of upsampling layers appear.(13) The second family summarizes the discrete-cosine transform, the representation used by image

https://doi.org/10.62486/978-9915-9966-3-9.Ch04

compression. The logarithm of the cosine magnitudes was divided into low, middle, and high frequency bands whose mean energies were recorded, and the remaining alternating-current coefficients were summarized with a 24-bin histogram. Together the two families describe both the overall spectral envelope and the fine structure of the coefficient distribution, and both have been used before in synthetic-image detection.[14,15] All features were computed with standard library transforms and required no learned model, which is the point of the design.

The choice of these two transforms is not arbitrary. The Fourier spectrum is where periodic artifacts, such as the checkerboard patterns left by transposed convolutions, show up as localized peaks, and a radial profile is a compact summary of that periodicity. The discrete-cosine transform is the basis of JPEG and other compression schemes, and its coefficient distribution is a long-established descriptor of natural image statistics, which makes departures from the natural distribution a natural signal for detection. The two are partly redundant, since both are unitary frequency transforms of the same image, but they weight the spectrum differently and their combination is standard practice in the frequency-detection literature.[14,16] No spatial features were included, deliberately, so that the study isolates the frequency signal and can attribute any cross-generator failure to that signal rather than to the classifier or to a confound.

CLASSIFIER AND EVALUATION PROTOCOL

The classifier was logistic regression, chosen because it is the lightest model that can learn a linear separation from a handful of features and because its coefficients remain directly interpretable. A separate detector was trained for each generator as the training domain: it saw the real photographs and that generator's images only, in a balanced half of the in-domain data. The detector was then evaluated in three ways. The held-out half of the in-domain data measured its baseline performance; the two other generators' images, each paired with real photographs, measured its cross-generator performance. Accuracy and the area under the receiver operating characteristic curve were recorded for every train-test pair.

Two design decisions deserve explicit statement because they bear on how the numbers should be read. The training and evaluation sets were balanced between real and generated images, so the chance baseline is 0.5 and every accuracy can be judged directly against that floor. The classifier was given no capacity beyond a linear decision boundary, which means any success it achieves must come from the features rather than from model complexity; this is deliberate, since the question is about the features, not about how much a powerful model can be squeezed. The average in-domain accuracy, the average cross-generator accuracy, and their difference summarize how much the detectors lose when they leave their training generator. Because the features are deterministic and the classifier and sampling are seeded, the entire pipeline is reproducible, and it was executed twice to confirm identical results.

RESULTS

IN-DOMAIN PERFORMANCE

A detector trained on a single generator distinguished that generator's images from real photographs with notably mixed success (Table 2). In-domain accuracy ranged from 0.68 for BigGAN to 0.92 for GLIDE, with VQDM in between at 0.71. The mean in-domain accuracy across the three generators was 0.77. The area under the curve followed the same ordering, rising from 0.76 for BigGAN to 0.97 for GLIDE. The spread is itself informative: the same features separate GLIDE images from photographs almost perfectly, yet leave a BigGAN detector near the point where it would be wrong one time in three.

The ordering across generators is not explained by any single property of the models. GLIDE, a diffusion model, was the easiest to detect, while BigGAN, a generative adversarial network, was the hardest, and VQDM, another diffusion-based model, sat between them. This suggests that the strength of the spectral fingerprint left by a generator depends on details of its training and sampling rather than on its family, a point that matters when one tries to predict which new generators a detector will handle. The practical reading is that even before considering generalization, the frequency approach is only as good as the generator it happens to face, and a practitioner who assumes a uniform baseline across generators will be misled by a detector that is excellent for one model and mediocre for another.

https://doi.org/10.62486/978-9915-9966-3-9.Ch04

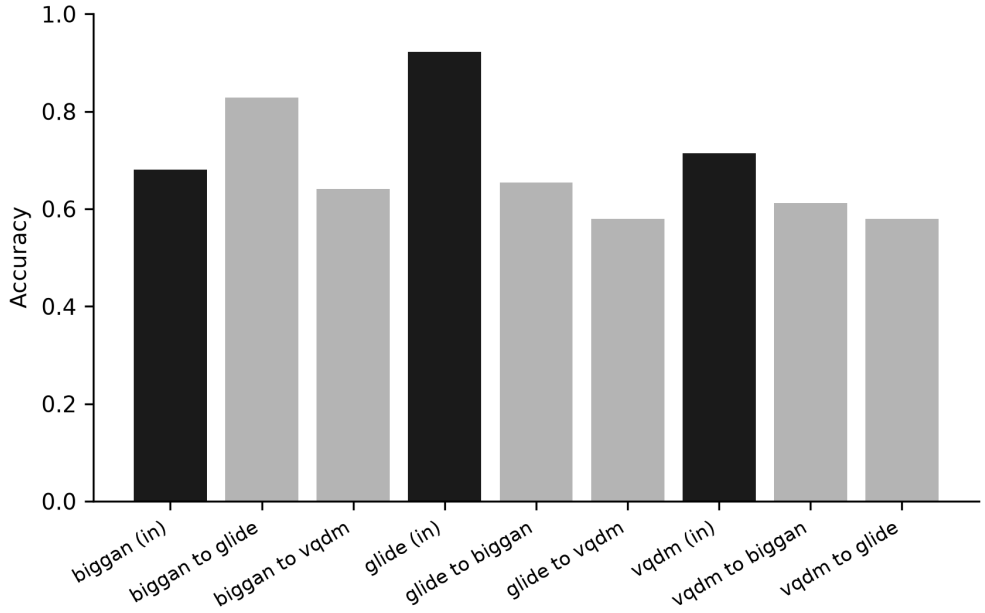
Table 2. Accuracy and area under the curve of frequency-domain detectors by train and test generator.

Train generator	Test set	Accuracy	AUC
biggan	in-domain	0.68	0.758
biggan	glide	0.829	0.941
biggan	vqdm	0.641	0.684
glide	in-domain	0.922	0.974
glide	biggan	0.654	0.74
glide	vqdm	0.579	0.565
vqdm	in-domain	0.714	0.782
vqdm	biggan	0.612	0.673
vqdm	glide	0.58	0.618

CROSS-GENERATOR PERFORMANCE

When a detector was tested on generators it had not seen, performance dropped in most but not all cases (Figure 1). The mean cross-generator accuracy was 0.65, against 0.77 in-domain, an average drop of 0.12 percentage points. The drop was uneven across the six train-test transfers. The GLIDE detector, the strongest in its own domain at 0.92, fell to 0.65 on BigGAN and 0.58 on VQDM. The BigGAN detector, by contrast, transferred to GLIDE at 0.83, an accuracy above its own in-domain score, although its transfer to VQDM was only 0.64. The weakest transfer of all was VQDM to GLIDE at 0.58, which left the detector barely above chance.

Figure 1. Accuracy of each frequency-domain detector in-domain and on held-out generators.



The area under the curve told the same story as accuracy. Cross-generator areas ranged from 0.56 at the low end to 0.94 at the high end, and in most transfers the detector lost the confidence margin that a usable classifier needs. The pattern that emerges from both metrics is that frequency-domain detectors are tuned to the spectral signature of their training generator, and that signature does not carry reliably to generators built on different architectures.

A useful way to condense the whole matrix is to look at the two directions of every generator pair. BigGAN and GLIDE transferred to each other at 0.83 and 0.65, an asymmetry in which the easier-to-detect GLIDE pulled the numbers in opposite directions. BigGAN and VQDM transferred at 0.64 and 0.61, and GLIDE and VQDM at 0.58 and 0.58. In every pair the two directions disagreed, which is exactly what a generator-specific fingerprint predicts: each detector inherits its training generator's signature, and each test generator presents its own, so the two directions are different questions and get different answers.

The magnitude of the loss is the quantity a practitioner most needs. The average drop of

https://doi.org/10.62486/978-9915-9966-3-9.Ch04

0.12 percentage points is large enough to matter in any application where a missed synthetic image or a falsely flagged photograph carries a cost, and it is the kind of number that should be quoted alongside a detector's headline accuracy whenever the detector is to be used outside its training distribution. That said, the drop is also bounded: the detectors rarely collapsed entirely, and most cross-generator accuracies remained above chance, which means the frequency signal is not wholly absent on unseen generators, only weakened. A reader should hold both facts at once: the transfer penalty is real and large, but it does not reduce the detectors to uselessness.

DISCUSSION

MAIN FINDINGS

Frequency-domain detectors generalize across generators poorly, losing twelve percentage points of accuracy on average and, in one transfer, collapsing to the level of random guessing. A detector built from Fourier and cosine features works as a screen within the generator it was trained on, and a strong one for some generators, but it cannot be trusted to recognize images produced by a generator it has not seen before. The second finding is that the loss is not uniform: it depends on which generator was trained on and which generator is being tested, which means the cross-generator performance of these detectors is a property of the specific generator pair rather than a single fixed penalty.

The third finding, less obvious but equally important, is that the in-domain baseline itself is only moderate. An average in-domain accuracy of 0.77 means that even in the most favorable case, where the detector is asked only about the generator it was trained on, it is wrong roughly one time in four. This qualifies any claim about the frequency approach: it is not a strong detector that happens to generalize poorly, but a modest detector that generalizes worse. The distinction matters because a practitioner weighing the frequency approach against a deep detector is choosing between a weak-but-cheap and a strong-but-expensive option, and the results here put a number on both halves of that trade-off.

COMPARISON WITH PRIOR WORK

The result is consistent with the mechanism that the frequency-detection literature has proposed from the beginning. Durall and colleagues showed that generative networks leave distinct spectral artifacts in their outputs because their upsampling layers fail to reproduce natural spectral distributions, and they warned that these artifacts are tied to the specific architecture of each generator.(13) Frank and colleagues built their detector on those same artifacts and reported strong in-domain results, but their evaluation did not push the detectors onto unseen generators.(14) The present study closes that gap: it shows that the artifacts are indeed exploitable within a generator and that the exploit transfers weakly, exactly as the mechanism predicts.

The magnitude of the drop is comparable to, though milder than, the degradation reported for deep detectors on the GenImage benchmark, where some detectors lose more than twenty points of accuracy across generators.(9) Frequency features are no worse than deep features in this respect, and they carry the advantage of requiring no training data and no learned model. But they are also no better, which matters because the hope behind lightweight frequency detectors was that a cheap, generator-agnostic signal could be found. The results here suggest that the signal is generator-specific, and that the cheapness of the detector buys speed and interpretability rather than robustness.(7,8)

The diffusion-specific literature points in the same direction. Corvi and colleagues showed that detectors built for one diffusion model must be recalibrated for another, and Bammey and colleagues found that a detector aimed at one family of diffusion models does not transfer cleanly to others, both of which are the same phenomenon observed here in a smaller, fully controlled setting.[21,22] Ricker and colleagues reached the complementary conclusion that diffusion artifacts are subtle enough that specialized, rather than generic, detectors are needed.(23) The present study adds to that body of work a clean estimate of the transfer penalty for the frequency approach, obtained without the confounds of architecture choice or large-scale training, and it finds that the penalty is substantial even between three well-known generators.(10)

The one apparent exception deserves scrutiny. The BigGAN detector transferred to GLIDE with higher accuracy than it achieved in its own domain. This does not contradict the overall

https://doi.org/10.62486/978-9915-9966-3-9.Ch04

finding; it reflects that GLIDE images are simply easier to separate from photographs, as the GLIDE detector's own near-perfect in-domain score shows. A generator whose images are far from natural photographs is easy for any detector, trained or not, to flag, and that ease can mask the underlying failure to generalize. The honest summary is not that BigGAN generalized to GLIDE, but that GLIDE is an easy target, a distinction that matters when a practitioner interprets a single favorable number.

WHY THE FEATURES FAIL TO TRANSFER

The failure has a concrete cause. The radial Fourier profile and the cosine-coefficient histogram describe the spectral fingerprint of a generator, and different generators leave different fingerprints because their upsampling, normalization, and noise schedules differ.[13,24] A logistic regression trained on one fingerprint learns a decision boundary that separates that fingerprint from natural spectra, and nothing in its training obliges it to separate a different fingerprint. The features are therefore best understood as a weak, generator-specific fingerprinting scheme rather than as a general measure of synthetic-ness. Learned detectors mitigate this by training on many generators, which is why they transfer better, but they pay for that robustness with data and compute that the frequency approach avoids.[5,9] The subpixel and dual-domain methods that have recently appeared are, in effect, attempts to find a fingerprint that sits below the generator-specific surface, and their mixed results are another sign that the surface-level spectral signal does not carry far.(19,25)

The heterogeneity across pairs reinforces this reading. VQDM and GLIDE, both diffusion-based, transferred to each other at the lowest accuracies in the study, while the GAN-based BigGAN transferred to GLIDE more readily. This ordering is the opposite of what a naive architectural-distance story would predict, which is a reminder that spectral fingerprints do not vary along a single architectural axis. A practitioner cannot assume that two diffusion models will leave similar artifacts, or that a GAN and a diffusion model will leave maximally different ones.

The heterogeneity also has a methodological lesson. Because the transfer penalty is a function of the pair, a single cross-generator comparison is not enough to characterize a detector's generalization; it can flatter or damn the detector depending on the particular pair chosen. Studies that report one held-out generator are therefore measuring a point on a surface, not the surface itself. The three-by-three matrix reported here, small as it is, is the minimum that allows the average and the spread to be separated, and it is the spread across pairs that a deployment decision should actually care about most. This argues for evaluating detection methods on generator matrices rather than single transfers, a practice the benchmark literature has moved toward but that is still not universal.(7,9)

PRACTICAL IMPLICATIONS

For someone who must deploy a detector today, the results support a narrow but honest recommendation. A frequency-domain detector is a defensible first-line screen when a learned detector is unavailable, when the compute budget forbids a deep network, or when the generator in question is known and will remain stable. It will flag the worst synthetic images cheaply and transparently, and its coefficients can be inspected to show what spectral property drove a decision, which has forensic value in itself.[15,16] But it should not be deployed against an open-ended stream of images from unknown generators, where its cross-generator error, twelve points on average and worse in some pairs, would produce both missed detections and false alarms at rates that most applications cannot accept.(2)

The recommendation can be stated more precisely for the three applications named at the outset. For the newsroom verifying one image, the frequency detector is the right first tool precisely because it is interpretable, and its cross-generator weakness matters less when the stakes are a single editorial decision and a human is in the loop. For the platform screening at scale, the detector is useful as a triage layer that routes suspicious images to a stronger, more expensive check, and its high false-alarm rate at the margin is tolerable so long as it never acts alone. For the forensic examiner, the detector's failure modes are themselves evidence, because a spectral fingerprint that does not match any known generator is a signal worth recording even when it cannot be interpreted.(26) In none of these roles is the frequency detector a complete answer, and the results here make clear that it should not be mistaken for one.

https://doi.org/10.62486/978-9915-9966-3-9.Ch04

There is also a reporting lesson for the field. A detection paper that reports only in-domain accuracy is reporting the best case, and a reader cannot infer from it how the detector will behave in the open world. The results here argue that cross-generator accuracy, stated as a matrix rather than a single transfer, should be reported alongside in-domain accuracy as a matter of course, in the same way that diagnostic studies report both internal and external validation. The twelve-point gap measured here is the kind of number that changes a deployment decision, and it is invisible in the headline accuracy that most papers foreground. Making the gap visible would do more for the practical usefulness of detection research than any single algorithmic improvement. (3,10)

Where robustness is required, the implications point toward generator-diverse training. The literature on universal fake-image detection has converged on the view that exposing a detector to many generators, or to a feature representation that abstracts over them, is what buys generalization.[27,28] The frequency approach can still contribute in that setting, not as the detector itself but as one input among several, since spectral features capture information that spatial features miss and the two are known to be complementary.(19,29,30) The contribution of this study is a quantified baseline for how far the frequency signal alone can go, against which hybrid and learned detectors can be measured. A related use, which the frequency fingerprint actually suits, is attribution: because the fingerprint is generator-specific, it is better at saying which generator produced an image than at saying whether an image is synthetic at all, and attribution systems have begun to exploit exactly this property.(24,26)

The study has limitations that follow from its design. Three generators are a small slice of the modern range of generative models, and the specific triple chosen here cannot be assumed to represent all generator architectures; a replication over the remaining generators of GenImage, and over diffusion and autoregressive models released after it, would test the stability of the twelve-point average. The features are a particular choice from a large design space, and other frequency representations, such as the wavelet transforms used in recent work, might transfer differently.(17) The real photographs and the generated images are drawn from the same ImageNet class structure, which controls content but also means the detectors were never asked to work on out-of-distribution content. The classifier is deliberately simple, and a more powerful lightweight model, such as a small gradient-boosted ensemble, might extract more from the same features. Each limitation suggests a concrete extension rather than undermining the conclusion: more generators, more feature families, out-of-distribution content, and a small set of stronger lightweight classifiers.

Two further limits concern resolution and the absence of an adversary. All images were downsampled to a common resolution before feature extraction, which discards the finest high-frequency detail where some generator fingerprints live; a comparison across native resolutions would show whether the transfer penalty shrinks when the full spectrum is available. And the study evaluated the detectors against images as produced, without the fingerprint-removal attacks that a real deployment would face, so the accuracies reported here are upper bounds on what a determined adversary would allow.(11) These gaps are easy to state and equally easy to close in follow-up work, and neither changes the qualitative conclusion that the frequency signal is generator-specific. The overall design, moreover, measures a fixed snapshot of three generators, and the field moves quickly enough that the specific numbers here will age; what should persist is the transfer-matrix method and the general finding that the penalty is large and pair-dependent.

CONCLUSIONS

Frequency-domain features supported lightweight detectors that separated machine-generated images from real photographs within a single generator, though with accuracy that varied widely from one generator to the next. Across generators, accuracy fell by an average of twelve percentage points, and one transfer collapsed to chance. The detectors exploit generator-specific spectral fingerprints rather than a universal property of synthetic imagery, which is why they transfer poorly and why their cross-generator behavior varies from pair to pair. Frequency-domain detection remains a cheap, interpretable first-line screen for a known generator, but it is not a substitute for generator-diverse training when robustness to unseen generators is required.

https://doi.org/10.62486/978-9915-9966-3-9.Ch04

REFERENCES

Tolosana R, Vera-Rodriguez R, Fierrez J, Morales A, Ortega-Garcia J. Deepfakes and beyond: A Survey of face manipulation and fake detection. Information Fusion. 2020;64:131-148. doi: 10.1016/j.inffus.2020.06.014
Masood M, Nawaz M, Malik KM, Javed A, Irtaza A, Malik H. Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward. Appl Intell. 2023;53(4):3974-4026. doi: 10.1007/s10489-022-03766-z
Rana MS, Nobi MN, Murali B, Sung AH. Deepfake Detection: A Systematic Literature Review. IEEE Access. 2022;10:25494-25513. doi: 10.1109/access.2022.3154404
Nguyen TT, Nguyen QVH, Nguyen DT, Nguyen DT, Huynh-The T, Nahavandi S, et al. Deep learning for deepfakes creation and detection: A survey. Computer Vision and Image Understanding. 2022;223:103525. doi: 10.1016/j.cviu.2022.103525
Singh A, Jain PC, Singh A. Beyond Clean Benchmarks: A Reproducible Severity-Sweep Evaluation of AI-Generated Image Detector Robustness. IRE Journals. 2026;10(3):3967-3974. doi: 10.64388/irev10i3-1723524
Wang S-Y. CNN-generated images are surprisingly easy to spot.. for now [Internet]. 2020. Available from: https://arxiv.org/abs/1912.11035
Coccomini DA, Caldelli R, Falchi F, Gennaro C. On the Generalization of Deep Learning Models in Video Deepfake Detection. J Imaging. 2023;9(5):89. doi: 10.3390/jimaging9050089
Meng Z, Peng B, Dong J, Tan T, Cheng H. Artifact feature purification for cross-domain detection of AI-generated images. Computer Vision and Image Understanding. 2024;247:104078. doi: 10.1016/j.cviu.2024.104078
Zhu M. GenImage: A Million-Scale Benchmark for Detecting AI-Generated Image [Internet]. 2023. Available from: https://arxiv.org/abs/2306.08571
Mahara A, Rishe N. Methods and trends in detecting AI-generated images: A comprehensive review. Computer Science Review. 2026;60:100908. doi: 10.1016/j.cosrev.2026.100908
Lu Y, Liu J, Zhang R. GANFR: GAN fingerprint removal network for image anti-forensics. Knowledge-Based Systems. 2025;327:114134. doi: 10.1016/j.knosys.2025.114134
Guarnera L, Giudice O, Battiato S. Mastering Deepfake Detection: A Cutting-edge Approach to Distinguish GAN and Diffusion-model Images. ACM Trans Multimedia Comput Commun Appl. 2024;20(11):1-24. doi: 10.1145/3652027
Durall R, Keuper M, Keuper J. Watch Your Up-Convolution: CNN Based Generative Deep Neural Networks Are Failing to Reproduce Spectral Distributions. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020:7887-7896. doi: 10.1109/cvpr42600.2020.00791
Frank J. Leveraging Frequency Analysis for Deep Fake Image Recognition [Internet]. 2020. Available from: https://arxiv.org/abs/2003.08685
Giudice O, Guarnera L, Battiato S. Fighting Deepfakes by Detecting GAN DCT Anomalies. J Imaging. 2021;7(8):128. doi: 10.3390/jimaging7080128
Li L, Lu T, Song J, Cheng K. SIDWA: Synthetic Image Detection Based on Discrete Wavelet Transform Stem and Deformable Sliding Window Cross-Attention. Electronics. 2026;15(4):891. doi: 10.3390/electronics15040891
Marini N, Berretti S, Caldelli R. Wavelet-based features to improve cross-forgery generalization in deepfake detection. Front Imaging. 2026;5:1845480. doi: 10.3389/fimag.2026.1845480
V YK, Dr DJV. Spatial-Frequency Fusion for Robust Detection of AI- Generated Images across Unseen Generators. Int J Innov Res Eng. 2026:01. doi: 10.59256/ijire.20260705001
Yan J, Li Z, Wang F, He Z, Fu Z. Dual Frequency Branch Framework With Reconstructed Sliding Windows Attention for AI-Generated Image Detection. IEEE TransInformForensic Secur. 2026;21:2313-2325. doi: 10.1109/tifs.2026.3664000
Choi Y. HFI: A unified framework for training-free detection and implicit watermarking of latent diffusion images [Internet]. 2024. Available from: https://arxiv.org/abs/2412.20704
Corvi R. On the detection of synthetic images generated by diffusion models [Internet]. 2023. Available from: https://arxiv.org/abs/2211.00680
Bammey Q. Synthbuster: Towards Detection of Diffusion Model Generated Images. IEEE Open J Signal Process. 2024;5:1-9. doi: 10.1109/ojsp.2023.3337714
Ricker J. Towards the detection of diffusion model deepfakes [Internet]. 2023. Available from: https://arxiv.org/abs/2210.14571
Liu C, Zhu T, Zhao Y, Zhang J, Zhou W. Disentangling different levels of GAN fingerprints for task-specific forensics. Computer Standards & Interfaces. 2024;89:103825. doi: 10.1016/j.csi.2023.103825
Lu P, Liang K, Liu X, Lu X, Li A. SDA-Det: A subpixel-aware dual-domain artifact detector for AI-generated image detection. Expert Systems with Applications. 2026;322:132304. doi: 10.1016/j.eswa.2026.132304
Namani Y, Reghioua I, Bendiab G, Labiod MA, Shiaeles S. DeepGuard: Identification and Attribution of AI-Generated Synthetic Images. Electronics. 2025;14(4):665. doi: 10.3390/electronics14040665
Ojha U. Towards universal fake image detectors that generalize across generative models [Internet]. 2023. Available from: https://arxiv.org/abs/2302.10174
Willi M, Mathys M, Graber M. Synthetic image detection with CLIP: Understanding and assessing predictive cues. Array. 2026;32:101222. doi: 10.1016/j.array.2026.101222
Yang T, Zhang Q. Dual-stream FDS-CRF net: frequency-spatial constrained attention for cross-generative model image detection. SIViP. 2025;19(11):932. doi: 10.1007/s11760-025-04486-w
Li J, Liang Y, Ye W, Wang P, Lei T, Zhang G, et al. Dual-branch AIGC image detection leveraging integrated semantic and frequency-domain features. Computer Vision and Image Understanding. 2026;271:104893. doi:

https://doi.org/10.62486/978-9915-9966-3-9.Ch04

10.1016/j.cviu.2026.104893

https://doi.org/10.62486/978-9915-9966-3-9.Ch04

Chapter 5 / Capítulo 5

SAP Superintelligence Series
ISSN: 3144-5624
ISBN (complete work): 978-9915-9851-0-7

Superintelligence Series Volume 5: Intelligent Visual Arts and Audiovisual Engineering: Generative AI, Computational Creativity and Superintelligence
ISBN (this volume): 978-9915-9966-5-3
DOI: 10.62486/978-9915-9966-3-9.Ch05
Pages: 47-57

Check for updates

© 2026 The authors. This is an open access book chapter distributed under the terms of the Creative Commons Attribution (CC BY) 4.0 License.

Growth, use and deletion of machine-generated images on Wikimedia Commons

Crecimiento, uso y marcadores de borrado de imágenes generadas por máquina en Wikimedia Commons

Matilde Bolaño-García[1]

1 Universidad del Magdalena. Facultad de Ciencias de la Educación. Santa Marta, Colombia.

ABSTRACT

Introduction: Wikimedia Commons, the free media repository of Wikipedia, has absorbed a rising number of machine-generated images, and their place in the project is contested. Systematic evidence on their growth, use and deletion is lacking. Methods: A registry census was conducted of the Commons category tree for machine-generated images using the MediaWiki API. All file members of the category and its subcategories were enumerated, and their upload timestamps, global usage and membership in quality-flag categories were recorded on a single snapshot date. Results: The census identified 2,548 image files across 104 subcategories. Uploads grew from 4 files in 2019 to 164 in 2022, 324 in 2024, and 1,035 in 2026. Only 2.0% of files were used in at least one wiki page. Membership in quality-flag categories reached 0.7% for low-quality tagging and 0.9% for obvious artifacts. Conclusions: Machine-generated images on Commons grew rapidly but remain almost entirely unused, and only a small share are flagged for deletion-related concerns. The repository is absorbing generated content far faster than the project is integrating it.

Keywords: Wikimedia Commons; machine-generated images; digital commons; open knowledge; generative models; content governance

RESUMEN

Introducción: Wikimedia Commons, el repositorio de medios libres de Wikipedia, ha absorbido un número creciente de imágenes generadas por máquina, y su lugar en el proyecto es controvertido. Falta evidencia sistemática sobre su crecimiento, uso y borrado. Métodos: Se realizó un censo del registro del árbol de categorías de imágenes generadas por máquina en Commons mediante la API de MediaWiki. Se enumeraron todos los archivos de la categoría y sus subcategorías y se registraron sus fechas de subida, su uso global y su pertenencia a categorías de marcado de calidad en una única fecha. Resultados: El censo identificó 2548 archivos en 104 subcategorías. Las subidas crecieron de 4 archivos en 2019 a 164 en 2022, 324 en 2024 y 1035 en 2026. Solo el 2,0% de los archivos se usaba en al menos una página wiki. La pertenencia a categorías de marcado de calidad alcanzó 0,7% para el etiquetado de baja calidad y 0,9% para artefactos evidentes. Conclusiones: Las imágenes generadas por máquina en Commons crecieron con rapidez pero permanecen casi sin uso, y solo una pequeña fracción está marcada para borrado. El repositorio absorbe contenido generado mucho más rápido de lo que el proyecto lo integra.

Palabras clave: Wikimedia Commons; imágenes generadas por máquina; bienes comunes digitales; conocimiento abierto; modelos generativos; gobernanza de contenidos

INTRODUCTION

Wikimedia Commons is the shared media repository behind Wikipedia, holding tens of millions of images, videos and documents under free licenses, and its value to the open knowledge ecosystem depends on two properties that pull against each other. It must be open enough to accept contributions from absolutely anyone, and it must be curated enough that what it hosts is actually useful to the encyclopedias and projects that draw on it.[1,2] The arrival of machine-generated images has stressed both properties at once. A contributor can now produce an image in seconds and upload it in a few seconds more, which floods the upload stream with content that no one has judged, while the community has never settled the question of whether such images belong in the repository at all.[3,4]

The debate is well documented in the community's own discussions. Wikipedia's editors spent three years working out how to govern language-model-generated text, arriving at a minimalist policy that accepted some uses while restricting others, and Commons now faces the parallel question for images, with the added complication that an image is harder to verify than a paragraph.[3,5] The stakes are genuinely concrete: a repository that accepts generated images without curation risks becoming a dumping ground, while one that rejects them outright forgoes illustrations that could genuinely serve articles.[6,7] The question also sits inside a longer scholarly conversation about what a commons is and how it stays healthy. Commons is a commons-based peer production project in the sense that Benkler and later theorists defined, and the defining question for such projects is whether the collective can govern the influx of contributions faster than those contributions erode the quality that made the commons worth having.(8-10)

Two further contexts frame the problem. The first is licensing. Commons exists to hold freely licensed media, and its usefulness depends on the license attached to every file, a point that generated images complicate because the legal status of their output, and of the models that produced it, remains unsettled in the jurisdictions that matter.(11-13) A generated image is, at the very minimum, a work whose authorship and copyright status are genuinely open questions, and a repository whose entire premise is clear licensing has to decide how to treat such works. The second is the wider rise of synthetic media. Generated images are one instance of a broader phenomenon whose societal risks, from deceptive content to the erosion of trust in photographs, are being studied across the media and forensics literatures, and Commons is a case in which those risks meet a community that has to make concrete, per-file decisions.(14-16) What has been missing from all of this is measurement. Claims about how many generated images Commons hosts, whether they are used, and how often they are challenged have been made on the basis of impressions rather than counts.

The choice of method matters here, because there is no single authoritative count of machine-generated images on Commons. They are not tagged with a machine-readable flag; they are identified only by the human-curated categories into which contributors have sorted them, and those categories were themselves created and reorganized over time.[17,18] A defensible estimate therefore requires a census of the category tree, not a search, and it requires care to distinguish when a file was uploaded from when it was categorized, because the two dates can differ by years and the distinction changes the entire growth story. This study reports such a census, chosen over the alternatives of a text search or a sample precisely because the category tree is the only complete, auditable register of the phenomenon. The objective was to quantify the growth, categorization, use and deletion-flagging of machine-generated images on Wikimedia Commons.

The contribution is deliberately narrow and descriptive. Rather than argue for or against generated images on Commons, the study measures what is actually there, and it does so in a way that can be re-run as the category tree continues to grow, so that the snapshot reported here can become the first point in a continuous time series. That is the kind of evidence the community's own discussions have been missing, and it is the kind that a method-focused study is best placed to supply.

METHODS

DESIGN AND DATA SOURCE

A cross-sectional registry census was conducted using the MediaWiki application programming interface of the Wikimedia Commons project itself. The register under study is the

https://doi.org/10.62486/978-9915-9966-3-9.Ch05

category tree rooted at the category for machine-generated images, whose members are files and subcategories assigned by contributors. The MediaWiki API was chosen because it is the same interface through which the repository is maintained, which makes the census reproducible by any researcher, and because it exposes both the category structure and the per-file metadata needed for the analysis.

The unit of analysis is the individual file, and the population is defined by category membership: a file is in scope if, and only if, a contributor has placed it under the machine-generated image category, whether directly or through any depth of subcategories. This is a consequential definition, because it makes the census depend on the community's own classification rather than on any independent judgment about what counts as generated. It was chosen deliberately over a content-based definition, which would have required inspecting each image, because the category tree is the operational definition the community itself uses, and a census that mirrors the community's own categories is the one that can speak to the community's own debates. The cost of this choice, that uncategorized generated images are invisible, is a limitation reported in the discussion.

ENUMERATION PROCEDURE

The category tree was traversed breadth-first. Starting from the root category, the API was queried for its members, distinguishing files from subcategories, and each subcategory was then visited in turn until the frontier was exhausted. Pagination followed the continuation token returned by the API, so that categories with more than five hundred members were enumerated completely rather than truncated at the first page. Every file was recorded with its title and its immediate parent category. The traversal recovered 2,548 files across 104 subcategories. The upload timestamp of each file was then retrieved in batches of fifty through the imageinfo endpoint, and the global usage of each file, meaning its transclusion in any page across any Wikimedia project, was retrieved in the same batches through the globalusage endpoint. Batching was required because the API returns per-file metadata only in groups, and the batch size of fifty is the API's own documented maximum for this operation. Two timestamps are therefore distinguished throughout: the upload timestamp, which records when the file was created, and the categorization timestamp, which records when it was added to the category tree, and the growth analysis uses the former.

VARIABLES

The primary variables were the upload year of each file, its immediate parent category, whether it was used in at least one wiki page, and whether it belonged to any of three quality-flag categories. The quality-flag categories are those that the community uses to mark machine-generated images of concern: a category for low-quality generated content colloquially tagged as generated sludge, a category for images with obvious visual artifacts, and a category for images whose generated provenance has not been identified. Membership in these categories is a proxy for the deletion debate rather than a direct count of deletion nominations, because a formal deletion nomination is a page-level action recorded in a separate log; the proxy is reported because it is the attribute that is uniformly and machine-readably visible on the file itself. The distribution of files across the immediate parent categories summarizes how the community organizes the content.

The distinction between direct and transitive category membership deserves a note. A file counted in the census may belong to the root category directly, or it may belong to a subcategory several levels deep, and the traversal records whichever immediate parent the API reports. This means the categorization counts are of the file's immediate parent, not of its position in the full hierarchy, and a file deep in the tree is attributed to its local category rather than to the root. The choice is deliberate: the immediate parent is the category that a contributor actually assigned, and it is the level at which the community's organizational decisions are visible. Aggregating everything to the root would hide exactly the structure that the categorization analysis is meant to reveal.

https://doi.org/10.62486/978-9915-9966-3-9.Ch05

ANALYSIS

Because the census enumerates the entire register rather than a sample, the analysis is descriptive: absolute counts and proportions, with no inferential statistics and no confidence intervals. Growth was tabulated by upload year, and the top-level organization was summarized by the frequency of each immediate parent category. The use rate was computed as the proportion of files transcluded in at least one wiki page, and the deletion-flag rate as the proportion belonging to each quality-flag category. The entire pipeline was written as a re-runnable script that queries the API directly, and it was executed with rate-limit handling that respects the API's retry guidance. The count of files was verified against the category information returned by the API where available.

The single methodological decision that deserves the most emphasis is the separation of the two timestamps. The MediaWiki API exposes, for every category member, a timestamp that records when the page was added to that category, and this is easily mistaken for the file's upload date. The two diverge substantially here, because the category tree that organizes machine-generated images is itself recent, so many files were uploaded years before they were sorted into it. Using the categorization timestamp would compress the entire growth history into the last two years and erase the early files; using the upload timestamp, which required a second, batched query per file, recovers the true trajectory. This is the kind of subtle choice on which the entire growth narrative turns, and it is reported here rather than left implicit, in keeping with the methodological focus of the study.

RESULTS

GROWTH

The census identified 2,548 machine-generated image files spread across 104 subcategories of the root category (Table 1). Uploads began as a trickle, with 4 files in 2019, 2 in 2020, and 5 in 2021. The first substantial increase came in 2022, when 164 files were uploaded, coinciding with the public release of the first widely used text-to-image models. The count rose to 362 in 2023 and eased to 324 in 2024, before climbing to 652 in 2025 and 1,035 in the first ten months of 2026 (Figure 1). The 2024 dip is the notable irregularity in an otherwise monotone rise, and it is discussed below. In cumulative terms the corpus grew from 164 files at the end of 2022 to over two and a half thousand by 2026, roughly a fifteen-fold increase in under four years, which is the number that any policy discussion would need to keep firmly in view.

Table 1. Machine-generated image files by upload year.

Year	Files uploaded
2019	4
2020	2
2021	5
2022	164
2023	362
2024	324
2025	652
2026	1035

CATEGORIZATION

The 104 subcategories show how the community actually sorts the content, and the largest of them reveal the character of what is being uploaded in practice (Table 2). The single largest category, with 357 files, is for images that include their generation prompt, a practice the community appears to encourage for transparency. The second largest, with 344 files, is for generated coats of arms, a niche that has produced a large and sustained output. The remainder spreads across diagrams, maps, headgear, science fiction illustrations, and landscapes, a distribution that mixes clearly illustrative material with ornamental and speculative content. The

https://doi.org/10.62486/978-9915-9966-3-9.Ch05

concentration at the top of the distribution is itself informative: two categories account for more than a quarter of the corpus, and the rest fans out into a long tail of narrowly defined subject categories, which is what one would expect from a bottom-up, contributor-driven classification rather than from a designed taxonomy.

Figure 1. Machine-generated image files uploaded to Wikimedia Commons by year.

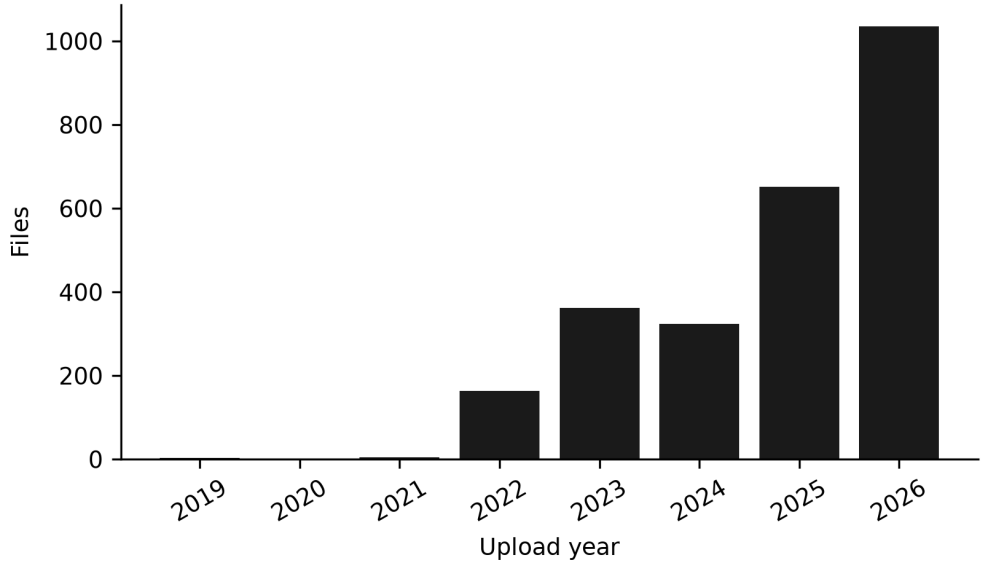


Table 2. Largest immediate subcategories of machine-generated images by file count.

Category	Files
AI-generated images including prompts	357
AI-generated coats of arms	344
AI-generated diagrams	99
AI-generated images of maps	98
AI-generated images of headgear	97
AI-generated science fiction	71
AI-generated winterscapes with Wikipedia-esque globes	71
Monochrome AI-generated images	64
AI-generated computers	59
AI-generated images of objects	59
AI misgeneration	58
AI-generated landscapes	56

USE AND DELETION FLAGS

The most striking single result in the whole census is the use rate. Of the 2,548 files in the census, only 51 were used in at least one wiki page, a rate of 2.0% (Table 3). The remaining 2,497 files exist in the repository but appear in no article, gallery, or user page across any Wikimedia project. The quality-flag categories tell a quieter story. Low-quality tagging covered 17 files (0.7%), the obvious-artifacts flag covered 23 files (0.9%), and the unidentified-provenance flag covered 1 files (0.0%).

The two indicators pull in opposite directions, and their combination is precisely the point. The near-zero use rate says the corpus is not yet serving the encyclopedia, while the low flag rates say the community is not yet systematically rejecting it either. Between those two facts lies a large body of files that are neither used nor flagged, sitting in an indeterminate state that the census can name but not resolve. That indeterminate state, and its substantial size relative to the small minority that is either used or challenged, is the most accurate single description of where machine-generated images stand on Commons today.

https://doi.org/10.62486/978-9915-9966-3-9.Ch05

Table 3. Use and quality-flag indicators for machine-generated images.

Indicator	Files	%
Files in tree	2548	100.0
Used in at least one wiki page	51	2.0
ai_slop	17	0.7
artifacts	23	0.9
unidentified	1	0.0

Note: Use means transclusion in at least one page of any Wikimedia project; quality flags are membership in the corresponding maintenance category.

DISCUSSION

MAIN FINDINGS

The census establishes three clear things about machine-generated images on Commons. Their number has grown rapidly, from a handful before 2022 to over one thousand files in a partial year, and the growth is recent enough that most of the register was created in the last two years. Their use is almost nil: two percent of files appear in a single wiki page, which means the repository is absorbing generated content far faster than the projects are consuming it. And their deletion-related flagging is small but nonzero, with a minority of files carrying the marks that the community applies when it judges a generated image to be poor or dubious. These three facts together describe a repository that is growing as an archive while not yet functioning as a supply of illustrations, and that state of affairs, more than any single number, is the headline result.

INTERPRETATION

The growth curve matches the external history of generative image models closely enough to read as a faithful record of it. The 2022 surge follows the public release of the first open text-to-image models, and the continued rise through 2025 and 2026 tracks the spread of those tools, a pattern consistent with the broader observation that generative systems have reshaped how volunteers produce and share visual content.[18,19] The 2024 dip is the one notable deviation, and it admits two readings that the census alone cannot separate. It may reflect a genuine slowdown in uploads, or it may reflect a lag in categorization, because a file counts in this register only once a contributor has placed it under the machine-generated category tree, and a backlog of uncategorized uploads would depress the 2024 count relative to its neighbors. The distinction between upload and categorization timestamps, which the method was designed to keep separate, is therefore also the source of the one interpretive uncertainty in the growth series.

The scale of the corpus should be put in proportion. The two and a half thousand files counted here are a small fraction of Commons, which holds tens of millions of media files, so machine-generated images are not yet restructuring the repository as a whole.(1) What is notable is the direction and slope of the curve: a doubling from 2024 to 2025 and another near-doubling into 2026 means the category is growing faster than the repository's human review practices are likely to scale, and it is the slope, more than the absolute count, that carries the governance implication.

The use rate is the finding with the clearest implication. A free-media repository that accepts images which almost no project then uses has, in effect, become an archive of unused illustrations, and the near-zero use rate quantifies a concern that has otherwise been voiced only impressionistically.[2,10] The concentration of uploads in ornamental niches such as generated coats of arms, which number in the hundreds while contributing little to the encyclopedia, points the same way: a large share of the generated corpus is produced for its own sake rather than to fill an encyclopedic need.(20) This is not necessarily a defect, because Commons has always hosted material whose usefulness is prospective, but it changes what the repository is becoming.

The contrast with the human-uploaded corpus is instructive, even if the census does not measure it directly. Commons as a whole achieves a substantially higher use rate because most uploads are photographs and scans that were made to illustrate something specific, whereas a large part of the generated corpus appears to have been made to exercise a tool rather than

https://doi.org/10.62486/978-9915-9966-3-9.Ch05

to meet a need.[1,2] If that reading is right, the near-zero use rate is not a temporary lag but a structural property of how the generated content is being produced, and it will not correct itself as the corpus matures. It would correct only if the projects began deliberately commissioning generated illustrations for their articles, which is a different activity from the one the upload stream currently reflects.(6)

The low deletion-flag rate should be read with care, and the method was chosen to make that care explicit. The three quality-flag categories capture only the files that contributors have already identified as problematic, and the flagging process itself is the bottleneck: a file can be poor without yet having been flagged, and a formal deletion nomination is a separate, heavier action that the census does not count directly.(4) The small observed proportions are therefore lower bounds on the true share of the corpus that the community would, on inspection, remove or reject. What they do show is that the visible, organized concern is concentrated in a small minority, which is consistent with a community that has not yet developed a systematic practice for reviewing generated images at the rate they arrive.(21)

The quality question also connects directly to a broader literature on how open-knowledge projects maintain verifiability and quality under sustained pressure. Wikipedia's content policies are built on a model of human review that generative tools disrupt, and the community has responded with both automated assistance and tighter norms, neither of which has yet settled the place of machine-produced media.[22-24] The near-zero use rate observed here is, in that light, a symptom rather than a cause: the projects that might consume generated images have not yet decided whether they should, and the unused corpus on Commons is the visible residue of that indecision.

The parallel with text is worth drawing out, because the two cases illustrate the same governance problem at different stages. Wikipedia settled on a restrictive-but-not-prohibitive policy for generated text only after years of debate, and its editors now treat generated prose as a thing to be used sparingly and disclosed when used.[3,5] Commons has not reached the equivalent settlement for images, and the census shows why the question is pressing: the images are already there, in the thousands, and every month of indecision adds more. The difference between the two cases is also instructive, because an image is harder to verify and easier to produce than a paragraph, which makes the image case the harder governance problem even though it has received less attention.(6)

METHODOLOGICAL ASSESSMENT

The census method was chosen over two alternatives, and the results are conditioned on that choice. A text search over file descriptions would have recovered files whose descriptions mention generated content, but it would miss files described in other languages or not described at all, and it could not distinguish generation from mere mention. A random sample would have been cheaper but could not support the per-year counts that are the point of the growth analysis. The category-tree census is complete for what it defines, but it inherits the definition: a file counts only if a contributor has placed it, directly or transitively, under the machine-generated category. Files generated by machine but categorized only by their subject, with no provenance category, are invisible to this method, so the census is a lower bound on the true population. The strength of the method is that this boundary is stated rather than hidden, and the counts are reproducible by anyone who re-runs the same traversal.[10,25]

The robustness of the findings can be judged against the assumptions they rest on. The growth series is robust to the exact choice of root category in its broad shape, because any reasonable definition of machine-generated images would capture the same surge, but its precise per-year counts are sensitive to how the category tree has grown, which is why the categorization timestamp is reported separately rather than discarded. The use rate is robust in the opposite way: it would only fall if the census over-counted, and the census is a lower bound, so the true use rate is at most two percent and possibly lower. The deletion-flag rate is the least robust figure, because it depends on the community's flagging activity rather than on any property of the files, and it should be read as a snapshot of that activity rather than as a stable attribute of the corpus. These are the kinds of assumption-checks that a method-focused study owes its reader, and they are stated here so that the conclusions can be recalibrated by anyone who disagrees with the boundary choices.

There is also a reproducibility argument that favors the method over the alternatives. A text search depends on the query string the researcher happens to choose, and two researchers

https://doi.org/10.62486/978-9915-9966-3-9.Ch05

with different queries will arrive at different counts; a census of a named category tree depends only on the name of the root category, which is a public, stable identifier. A sample depends on the sampling procedure and the seed. The census is the option that most cleanly separates the researcher's choices from the object being measured, which is why it is the right instrument for a question whose answer should not depend on who is asking it. This is the methodological principle behind the design, and it is stated here because the study's conclusions inherit their force from it.

GENERALIZABILITY

The findings generalize in a qualified and carefully bounded sense. The specific counts are a snapshot of one repository at one date, and they will change, but the structural pattern, a steep upload curve meeting a flat use rate, is unlikely to be unique to Commons. Any open repository that accepts freely licensed media will face the same asymmetry once generation becomes cheap, because the cost of producing a generated image is near zero while the cost of integrating it into an article remains a human judgment. The method, a category-tree census with separated timestamps, is the more portable contribution, and it can be applied to any wiki that uses the same MediaWiki platform or to any repository with a comparable human-curated category structure. What the study cannot generalize is the community's particular response, because that response is shaped by Commons's history and norms rather than by any feature of the phenomenon.[25,26]

IMPLICATIONS

For the Commons community, the results make the policy question quantitative. The repository is not being overwhelmed by deletion debates, but neither is it integrating what it accepts, and the gap between the upload curve and the flat use rate is where any future governance decision will have to land.(3) A policy that accepts generated images uncritically would codify an archive that no one reads, while a policy that curates them would have to scale a review process to an upload stream that has already passed a thousand files in a year.(6) The distinction between upload and categorization timestamps, which proved material here, is also a caution for any future measurement of Commons, where the two dates are easy to conflate and their conflation would distort any growth study.(18)

The finding also suggests a division of attention for the community. The deletion-related flagging, while real, is a second-order concern compared with the use gap, and a community that spends its energy adjudicating individual deletions without addressing the integration problem will be managing symptoms. The more consequential question is whether the projects that Commons serves actually want generated illustrations, and if so, under what conditions, and the near-zero use rate suggests that question has not yet been answered in the affirmative. A deliberate decision either way, grounded in the counts reported here, would be an improvement over the current state, in which the repository absorbs content by default and the projects ignore it by default.

The measurement itself is part of the answer, in a modest way. A count is not a policy, but it is a precondition for one, and a community that can cite how many files it hosts, how many it uses, and how many it has flagged is in a better position to decide than one arguing from anecdote. The census reported here provides those numbers for the present, and because it is reproducible and cheap, it can be repeated as the category grows, giving the community a running measure of a problem that is otherwise easy to misjudge by feel. That, ultimately, is the practical contribution of a descriptive study of this kind.

For the broader open knowledge ecosystem, the finding is a case study in how a commons absorbs a new, cheap, high-volume form of contribution. The same dynamics that produced a near-unused corpus of generated images are likely to recur wherever generative tools meet open repositories, and the measurement approach here, a category-tree census with separated timestamps, is portable to those settings.[26,27] The result also bears on the recurrent worry that generated content will crowd out human contribution; on Commons, so far, the generated corpus has accumulated alongside rather than replaced, but its scale relative to its use suggests the crowding question is still open.[5,28] The synthetic-media governance literature reaches the same conclusion from the other direction, arguing that the risks of generated content are best

https://doi.org/10.62486/978-9915-9966-3-9.Ch05

managed by institutions and norms rather than by blanket prohibition, which is precisely the terrain on which Commons is now making its decisions.[29,30]

CONCLUSIONS

A census of the Wikimedia Commons category tree identified a machine-generated image corpus whose uploads grew from a handful before 2022 to well over one thousand files in a single partial year, with a dip in 2024 that the method cannot fully resolve. The corpus was organized into over one hundred subcategories, the largest of which held generated prompts and generated coats of arms. Use was near zero, at two percent of files appearing in any wiki page, while deletion-related flagging was confined to a small minority. The repository is absorbing machine-generated images far faster than the projects are using them, a gap that any governance decision will have to address explicitly.

https://doi.org/10.62486/978-9915-9966-3-9.Ch05

REFERENCES

Kelly EJ. Reuse of Wikimedia Commons Cultural Heritage Images on the Wider Web. EBLIP. 2019;14(3):28-51. doi: 10.18438/eblip29575
Yu Y, McDonald DW. “Why do you need 400 photographs of 400 different Lockheed Constellation?”: Value Expressions by Contributors and Users of Wikimedia Commons. Proc ACM Hum-Comput Interact. 2023;7(CSCW2):1-34. doi: 10.1145/3610094
Froneman W. Failed comprehensiveness, successful minimalism: Wikipedia’s 3-year struggle to govern AI-generated content (2022–2025). AI & Soc. 2026;41(7):7045-7061. doi: 10.1007/s00146-026-03046-1
Fisher SA, Howard JW, Kira B. Moderating Synthetic Content: the Challenge of Generative AI. Philos Technol. 2024;37(4):133. doi: 10.1007/s13347-024-00818-9
Wagner C, Jiang L. Death by AI: Will large language models diminish Wikipedia?. Asso for Info Science & Tech. 2025;76(5):743-751. doi: 10.1002/asi.24975
Reeves N, Simperl E. Machines in the Margins: A Systematic Review of Automated Content Generation for Wikipedia. Proc ACM Hum-Comput Interact. 2025;9(7):1-30. doi: 10.1145/3757626
Yu Y, McDonald DW. Computer-Aided Tagging on Wikimedia Commons: Designing for Human–AI Collaboration in Open Knowledge Work. Proc ACM Hum-Comput Interact. 2026;10(6):1-29. doi: 10.1145/3816912
Papadimitropoulos V. Commons-Based Peer Production in the Work of Yochai Benkler. tripleC. 2018;16(2):835-856. doi: 10.31269/triplec.v16i2.1009
Aryan V, Bertling J, Liedtke C. Topology, typology, and dynamics of commons-based peer production: On platforms, actors, and innovation in the maker movement. CreatInnov Manage. 2021;30(1):63-79. doi: 10.1111/caim.12392
Jullien N, Roudaut K. Are Digital Knowledge Commons Real Commons?. International Journal of the Commons. 2026;20(1):468-483. doi: 10.5334/ijc.1691
Simmonds T. Common Knowledge? The Rise of Creative Commons Licensing. Legal Information Management. 2010;10(3):162-165. doi: 10.1017/s1472669610000642
Hare C. Copyfight: Creative Commons, Open Licensing, Bringing Information to the People (and letting them use it). DJIM. 2010;3(1). doi: 10.5931/djim.v3i1.24
Quintais JP. Generative AI, copyright and the AI Act. Computer Law & Security Review. 2025;56:106107. doi: 10.1016/j.clsr.2025.106107
Amerini I, Barni M, Battiato S, Bestagini P, Boato G, Bruni V, et al. Deepfake Media Forensics: Status and Future Challenges. J Imaging. 2025;11(3):73. doi: 10.3390/jimaging11030073
Pawelec M. Decent deepfakes? Professional deepfake developers’ ethical considerations and their governance potential. AI Ethics. 2025;5(3):2641-2666. doi: 10.1007/s43681-024-00542-2
Glick J. Deepfake Satire and the Possibilities of Synthetic Media. Afterimage. 2023;50(3):81-107. doi: 10.1525/aft.2023.50.3.81
Yu Y, McDonald DW. Unpacking Stitching between Wikipedia and Wikimedia Commons: Barriers to Cross-Platform Collaboration. Proc ACM Hum-Comput Interact. 2022;6(CSCW2):1-35. doi: 10.1145/3555766
Jankowski S. Wikimedia research methodologies: a bibliometric history of peer production and AI (2005–2025). AI & Soc. 2026. doi: 10.1007/s00146-026-03359-1
Luth E. The Use of Wikipedia, Wikimedia, and Open Access Content for Artificial Intelligence and Text and Data Mining. SIPLR. 2024;7(2):37-43. doi: 10.53292/33313cc8.be33e111
Kelly EJ. Digital Cultural Heritage and Wikimedia Commons Licenses:. JCEL. 2019;3(3):1-25. doi: 10.17161/jcel.v3i3.9771
Zhao A, Naaman M, Wu L. Rule Development in Online Communities Amidst Growth and Generative AI. Proc ACM Hum-Comput Interact. 2026;10(6):1-23. doi: 10.1145/3816993
Niederer S, van Dijck J. Wisdom of the crowd or technicity of content? Wikipedia as a sociotechnical system. New Media & Society. 2010;12(8):1368-1387. doi: 10.1177/1461444810365297
Ferretti E, Cagnina L, Paiz V, Donne SD, Zacagnini R, Errecalde M. Quality flaw prediction in Spanish Wikipedia: A case of study with verifiability flaws. Information Processing & Management. 2018;54(6):1169-1181. doi: 10.1016/j.ipm.2018.08.003
Petroni F, Broscheit S, Piktus A, Lewis P, Izacard G, Hosseini L, et al. Improving Wikipedia verifiability with AI. Nat Mach Intell. 2023;5(10):1142-1148. doi: 10.1038/s42256-023-00726-1
Dulong de Rosnay M, Stalder F. Digital commons. Internet Policy Review. 2020;9(4):15 p. doi: 10.14763/2020.4.1530
Fuchs C. The Digital Commons and the Digital Public Sphere: How to Advance Digital Democracy Today. Westminster Papers in Communication and Culture. 2021;16(1). doi: 10.16997/wpcc.917
Taeihagh A. Governance of Generative AI. Policy and Society. 2025;44(1):1-22. doi: 10.1093/polsoc/puaf001
Pawelec M. Deepfakes and Democracy (Theory): How Synthetic Audio-Visual Media for Disinformation and Hate Speech Threaten Core Democratic Functions. DISO. 2022;1(2):19. doi: 10.1007/s44206-022-00010-6
Alexis Celia, Brittany Donnie, Masaki Hasegawa, Maelle Renard. Deepfakes, Synthetic Deception, and Responsible Deployment of Generative AI. J Artif Intell Eng Consort. 2026;2(1):50-59. doi: 10.66238/jaiec58
Montague-Hellen B. Empowering knowledge through AI: open scholarship proactively supporting well trained generative AI. Insights the UKSG journal. 2024;37:10. doi: 10.1629/uksg.649

Chapter 6 / Capítulo 6

SAP Superintelligence Series
ISSN: 3144-5624
ISBN (complete work): 978-9915-9851-0-7

Superintelligence Series Volume 5: Intelligent Visual Arts and Audiovisual Engineering: Generative AI, Computational Creativity and Superintelligence
ISBN (this volume): 978-9915-9966-5-3
DOI: 10.62486/978-9915-9966-3-9.Ch06
Pages: 71-83

Check for updates

© 2026 The authors. This is an open access book chapter distributed under the terms of the Creative Commons Attribution (CC BY) 4.0 License.

A systematic review of metrics and human evaluation practices for video generation models

Una revisión sistemática de métricas y prácticas de evaluación humana para modelos de generación de video

Rocío de la Cruz Hernández[1]

1 Tecnológico Nacional de México/ITS de Centla. Mexico.

ABSTRACT

Introduction: Video generation models have advanced rapidly, but the way they are evaluated has not kept pace, and the field lacks a synthesis of which automatic metrics and human-evaluation protocols are actually used. Methods: A systematic review was conducted following PRISMA 2020. Four bibliographic sources were searched for empirical studies of video generation evaluation published between 2018 and 2026. Records were screened on title and abstract and then on full text, and automatic metrics and human-evaluation protocols were extracted from the included studies. Results: The search identified 210 records, of which 11 studies were included. All included studies reported human evaluation, and all reported the frame-level Fréchet inception distance, while the video-level Fréchet video distance was used in 72.7% of studies and CLIP-based text-alignment scores in 45.5%. Human-evaluation protocols varied in rater pool, scale and instructions, and few studies reported these details completely. Conclusions: Evaluation of video generation rests on a small set of inherited image metrics plus universal but inconsistently reported human studies. The field would benefit from standardized reporting of both metric choices and human-evaluation protocols.

Keywords: video generation; evaluation metrics; human evaluation; text-to-video; generative models; systematic review

RESUMEN

Introducción: Los modelos de generación de video han avanzado con rapidez, pero su evaluación no ha seguido el mismo ritmo, y falta una síntesis de qué métricas automáticas y protocolos de evaluación humana se usan en la práctica. Métodos: Se realizó una revisión sistemática siguiendo PRISMA 2020. Se buscaron estudios empíricos de evaluación de generación de video publicados entre 2018 y 2026 en cuatro fuentes bibliográficas. Los registros se cribaron por título y resumen y luego por texto completo, y se extrajeron las métricas automáticas y los protocolos de evaluación humana. Resultados: La búsqueda identificó 210 registros, de los que se incluyeron 11 estudios. Todos los estudios incluidos informaron evaluación humana, y todos informaron la distancia de inicio de Fréchet a nivel de fotograma, mientras que la distancia de video de Fréchet se usó en 72,7% de los estudios y las puntuaciones de alineación texto-video basadas en CLIP en 45,5%. Los protocolos de evaluación humana variaron en el grupo de evaluadores, la escala y las instrucciones, y pocos estudios informaron estos detalles de forma completa. Conclusiones: La evaluación de la generación de video descansa en un conjunto pequeño de métricas heredadas de imágenes más estudios humanos universales pero informados de forma inconsistente. El campo se beneficiaría de un reporte estandarizado de las métricas y de los protocolos de evaluación humana.

Palabras clave: generación de video; métricas de evaluación; evaluación humana; texto a video; modelos generativos; revisión sistemática

INTRODUCTION

Video generation has moved, in a few years, from a research curiosity to a capability that produces temporally coherent clips from text, images, and audio, and the pace of model development has outstripped the pace of evaluation.[1,2] The problem is familiar from image generation, where a small set of metrics inherited from older generative models still dominates, but video adds a dimension that images lack: a video can be spatially faithful in every frame and yet temporally incoherent, semantically misaligned, or unnatural in motion, and none of these properties is captured by a per-frame image metric.[3,4] Evaluation is therefore the bottleneck that determines how the field compares models, which models are published, and which are believed to work.

The question has consequences that reach well beyond the immediate research community. Video generation is among the first generative capabilities in which the output is both high-dimensional and temporally structured, which makes it a testing ground for evaluation practices that will be needed more widely as generative systems advance, and a field that cannot measure its own progress cannot be steered or governed.[1,2] The same evaluation questions recur in every domain where generated media must be compared and ranked, and the video case is the hardest version of a problem that image, audio and text generation each face in simpler and more settled form. A systematic account of how video is evaluated is therefore not only a description of one subfield but a case study in whether generative modeling has yet learned to measure what it produces.

The tools available are of two kinds. Automatic metrics compute a score without human involvement; the most common are the Fréchet inception distance, a frame-level measure inherited from image generation, and the Fréchet video distance, its video-level generalization, along with CLIP-based scores that measure alignment between the generated video and its text prompt.(5-7) Human evaluation asks people to rate generated videos, and it is the reference against which the automatic metrics are meant to be validated, but it is expensive, and its protocols vary widely across studies.(8) The relationship between the two is the crux of the evaluation problem: if automatic metrics do not track human judgment, then a field that ranks models by those metrics is ranking them by a proxy of unknown quality, and if human studies are not reported consistently, then the reference itself is not reproducible.(9,10)

The two kinds of tool are not alternatives so much as complements with a fixed and well-understood division of labor. Automatic metrics are the cheap, scalable part: they can be run on every model and every checkpoint, and they are what leaderboards and papers print as their headline numbers. Human evaluation is the expensive, scarce part: it is the ground truth that the automatic metrics are meant to approximate, and it is what any claim of genuine progress ultimately rests on. A healthy evaluation practice keeps the two in a stated, continually re-validated relationship, checking the cheap metric against the expensive reference; an unhealthy one lets the cheap metric drift free of the reference it is supposed to track, and the question the review poses is which of these two states the video literature is in.

Prior work has criticized individual metrics and proposed better ones, but the field has lacked a systematic account of what is actually used in practice.[9,11] Benchmarks such as VBench and EvalCrafter have catalogued evaluation dimensions and exposed the misalignment between automatic and human scores, and surveys have described the space of models and methods, but neither benchmarks nor surveys answer a simpler, empirical question: across the published literature, which metrics do studies actually report, how many run human evaluations, and how consistently are those evaluations described.[12,13] That question matters because the answers describe the de facto standard of the field, and the de facto standard is what new researchers inherit and what reviewers implicitly enforce.

The evaluation problem also inherits a longer history. Video quality assessment predates generative video by two decades, and its classical tradition distinguishes quality of service from quality of experience, the latter being the perceptual judgment that human raters provide and that no single signal-level metric fully captures.[14,15] The generative models themselves arrived in waves: the first text-to-video systems used generative adversarial networks and variational autoencoders, and the diffusion-based models that now dominate appeared later and changed both the quality achievable and the artifacts that evaluation must detect.(16-18) Early text-to-video work also explored training efficiency through knowledge distillation, and the subsequent diffusion line includes conditional prompt-based variants, so the corpus of models is heterogeneous in architecture even as it converges on diffusion.[19,20] CLIP-based alignment scoring added a third pillar, measuring semantic correspondence between a video and its text prompt through a joint

https://doi.org/10.62486/978-9915-9966-3-9.Ch06

vision-language representation, and its adoption reflects the field's growing recognition that perceptual quality and prompt alignment are distinct things that must be measured separately. [7,21] The objective of this systematic review was therefore to synthesize the automatic metrics and human-evaluation protocols used to assess video generation models and to characterize how consistently they are reported.

METHODS

PROTOCOL AND ELIGIBILITY CRITERIA

The review followed the PRISMA 2020 statement for systematic reviews. Because this is a synthesis of reported evaluation practices rather than of intervention effects, the question was framed in PICO terms as follows: the population is video generation models and the videos they produce; the intervention is evaluation, whether automatic or human; there is no comparator; and the outcome is the reported metric or the human-evaluation protocol and the completeness of its reporting. Studies were eligible if they generated video with a neural model and reported at least one quantitative or human evaluation of that video, were published between 2018 and 2026, and were written in English or Spanish. Studies were excluded if they concerned image-only generation, video processing without generation, or opinion pieces without empirical evaluation. The review was not registered, and this is stated rather than hidden.

The eligibility criteria were deliberately permissive about the type of contribution, because the object of study is the reporting of evaluation, and evaluation appears in model papers, benchmarks, and surveys alike. A stricter definition, restricted to a single task or a single model family, would have produced a cleaner but less representative picture, and representativeness is what the review's question requires. The trade-off is that the included studies are heterogeneous in purpose, and the synthesis reports that heterogeneity rather than smoothing it over or treating it as noise to be averaged away.

INFORMATION SOURCES AND SEARCH

Three bibliographic sources were searched: OpenAlex, Crossref, and Semantic Scholar. The search combined terms for video generation with terms for evaluation, and the queries used for the record of searches included text-to-video generation evaluation metrics, video generation model human evaluation, Fréchet video distance text-to-video, video diffusion model quality assessment, and text-to-video generation survey, each run against the three sources and retrieved with a page size large enough to return the most relevant records rather than a truncation of them. Records were deduplicated across sources by DOI where one existed and otherwise by a normalized form of the title, and the number of duplicates removed is reported in the flow diagram.

SELECTION PROCESS

Screening was conducted in two stages. In the first stage, titles and abstracts were screened against the eligibility criteria, and records that clearly concerned image-only generation, video processing, or non-empirical commentary were excluded. In the second stage, the full text of the retained records was sought through open-access channels, and the retrieved texts were assessed against the same criteria.

DATA EXTRACTION AND SYNTHESIS

For each included study, the following were extracted from the full text where available: the type of contribution, the automatic metrics reported, and the human-evaluation protocol, including the rater pool, the rating scale, and the instructions. Extraction was recorded per study in a structured table, and metric usage was summarized as counts and proportions across the included studies. Because the outcome is a description of practices rather than a numerical effect, the synthesis is narrative and thematic rather than quantitative; no meta-analysis was appropriate, and none was attempted.

https://doi.org/10.62486/978-9915-9966-3-9.Ch06

RISK OF BIAS AND COMPLETENESS OF REPORTING

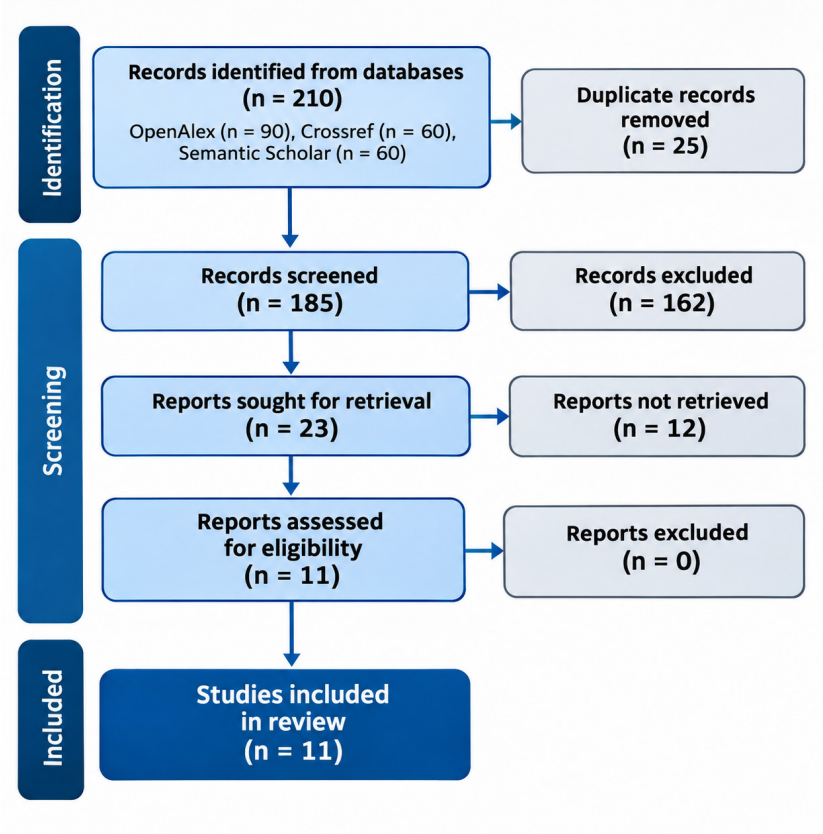
Risk of bias in the conventional sense, of a study's estimate of an effect, does not apply to a review of reporting practices, and no tool such as ROBINS-I or RoB 2 was therefore used. What plays the analogous role here is the completeness of reporting, and this was assessed uniformly rather than imputed: a metric was recorded as used only if it appeared in the study's text, and a human-evaluation detail was recorded only if it was stated. This conservative rule means the frequencies reported in the results are lower bounds on what studies may actually have done, and the review treats missing detail as data rather than as something to reconstruct. The assessment is therefore an assessment of transparency, and it is the transparency of evaluation that this review is designed to measure.

RESULTS

STUDY SELECTION

The searches identified 210 records in total, of which 25 were duplicates removed before screening began (Figure 1). Of the 185 records screened on title and abstract, 162 were excluded, most of them because they concerned image generation or video tasks other than generation. The full text of 23 reports was sought for retrieval, of which 12 could not be obtained through open-access channels, leaving 11 reports assessed for eligibility, all of which were retained. The review therefore ultimately included 11 studies in the synthesis. The exclusion of 12 reports at the retrieval stage is the largest single point of attrition after the initial screen, and it reflects the share of the literature that is not open access, which is itself a limit on the completeness of any review that does not have institutional access to paywalled sources.

Figure 1. PRISMA flow diagram of study selection.



CHARACTERISTICS OF INCLUDED STUDIES

The 11 included studies span three kinds of contribution (Table 1). Six are generation-model

https://doi.org/10.62486/978-9915-9966-3-9.Ch06

papers that propose a video model and report its evaluation: the earliest text-to-video work,(16) and the later diffusion-based models Show-1, Swap Attention, VideoDreamer, GroupVideo and LaVie.(22-26) Three are benchmarks that introduce a suite for evaluating video generation, namely VBench++, a multidimensional AIGC video quality assessment dataset, and VE-Bench for video editing.(27-29) Two are surveys that catalog models, datasets and evaluation practices.[1,30] The earliest included study dates from 2018, the year the first text-to-video systems were proposed, and the remainder cluster in 2024 through 2026, matching the recent acceleration of the field around diffusion-based methods. The studies cover text-to-video, image-to-video, and text-driven video editing, so the synthesis reflects the broader evaluation problem across tasks rather than a single narrow setting.

Table 1. Characteristics and reported evaluation of the included studies.

Study	Year	Type	FVD	FID	IS	CLIP	Human evaluation
Li	2018	model	-	Y	Y	-	Y
Huang	2026	benchmark	Y	Y	Y	-	Y
Zhang	2025	model	Y	Y	Y	-	Y
Wang	2025	model	Y	Y	-	Y	Y
Zhang	2025	benchmark	Y	Y	Y	Y	Y
Jan	2025	survey	Y	Y	Y	Y	Y
Song	2026	model	-	Y	-	Y	Y
Chen	2025	model	-	Y	-	-	Y
Wang	2025	model	Y	Y	-	-	Y
Xing	2025	survey	Y	Y	Y	Y	Y
Sun	2025	benchmark	Y	Y	-	-	Y

AUTOMATIC METRICS

The automatic metrics form a small, stable set inherited from image generation, and their stability across studies is itself a finding (Table 2). The frame-level Fréchet inception distance was reported in every included study, making it the one metric with universal adoption. The video-level Fréchet video distance, which measures temporal statistics and is therefore the metric most specific to video, was reported in 8 of the 11 studies (72.7%). The inception score, an older metric whose limitations are well documented, was still reported in 6 studies (54.5%). CLIP-based scores, which measure text-video alignment rather than perceptual quality, were reported in 5 studies (45.5%), and they appeared in the benchmark and survey studies that are specifically concerned with alignment rather than in the model papers. No study relied on a single metric alone; the modal practice was to report the Fréchet inception distance together with one or two others.

The pattern across the studies is more informative than any single frequency. The benchmark studies report the broadest set, typically the full quartet of the frame and video Fréchet distances, the inception score and a CLIP score, because their purpose is to define the standard rather than to report one model's number.[27,28] The model papers report a narrower set, usually the two Fréchet distances and a human study, because their purpose is to show that their model improves over a baseline.[22,26] The surveys report everything by construction, since they describe the field's practice rather than adding to it.[1,30] This division of labor explains why the frame-level metric is universal while the more specialized alignment and video-level metrics are not: the former is the common denominator, and the latter are adopted where the study's purpose demands them.

The benchmark studies also show the field beginning to decompose video quality into multiple named dimensions rather than a single scalar number. VBench++ structures its evaluation around dimensions of quality and consistency, and the AIGC quality assessment and video editing benchmarks follow the same decomposition, pairing each dimension with a metric and a human rating.(27-29) This dimensional turn is the clearest sign of progress in the corpus, because it is precisely the move that a single scalar metric cannot make, and it is the direction in which the field's evaluation practice is most likely to converge. The older scalar metrics remain as baselines, but the dimensional suites are where the newer studies invest their effort.

https://doi.org/10.62486/978-9915-9966-3-9.Ch06

Table 2. Automatic metrics reported across the included studies.

Metric	Studies	% of included
Fréchet video distance (FVD)	8	72.7
Fréchet inception distance (FID)	11	100.0
Inception score (IS)	6	54.5
CLIP score	5	45.5
Human evaluation (MOS/preference)	11	100.0

HUMAN EVALUATION

Human evaluation was universal: every included study reported at least one human or subjective judgment, whether a mean opinion score, a preference judgment between videos, or a user study (Table 1). The universality of human evaluation is the single most consistent finding, and it sits in tension with the small set of automatic metrics: the field evidently regards human judgment as indispensable even while it ranks models by automatic scores. The reporting of the human studies, however, varied considerably. Some studies stated the number of raters, the rating scale, and the instructions; others reported only that a human study was conducted, without the details that would make it reproducible. The variability is the clearest gap identified by the synthesis, and it is quantified here only descriptively, because the missing details are themselves the finding.

The three recurring forms of human judgment each carry different reporting burdens. A mean opinion score requires a defined scale and enough raters to make the mean stable; a preference judgment requires stating the pairing and the instruction given to the rater; and a user study requires the recruitment, the task, and the exclusion rules. The included studies differ in which of these burdens they discharge, and the pattern is that the newer benchmark studies, which are explicitly about evaluation, report their human protocols more completely than the older model papers, which treat human evaluation as a supporting result. This is an encouraging direction, but it remains uneven across the corpus.

The completeness of reporting can be described by the questions a reader can answer after reading a paper. In the most complete cases, a reader can learn how many raters participated, where they were recruited, what instruction they were given, what scale they used, and how the ratings were aggregated; in the least complete, a reader learns only that a human study occurred. The review did not assign a numeric score to this completeness, because a rubric would have introduced more judgment than the evidence supports, but the difference between the two ends is clear enough to state without one. The unevenness, rather than any particular study's practice, is the finding that matters for the field's reproducibility.

CONSISTENCY OF REPORTING

The consistency of evaluation reporting is mixed in a specific pattern. The choice of automatic metric is reported consistently enough that the field's de facto standard is visible at a glance, but the human-evaluation protocol, which is the more consequential part because it is the reference, is reported less completely. A reader who wants to know which metrics a study used can find the answer from the tables in most papers; a reader who wants to replicate the study's human evaluation frequently cannot. This asymmetry, between the mechanical and reproducible part of evaluation and the human and non-reproducible part, is the main reporting gap the review documents.

The asymmetry has a practical explanation. Automatic metrics are standardized by their very nature: a name such as Fréchet video distance refers to a defined computation, and reporting the name plus the number is reporting everything a reader needs. Human evaluation has no such standard name, and a phrase such as a human study conceals an entire methodology that the reader cannot reconstruct from the phrase alone. The result is that the part of evaluation that is easiest to replicate is reported completely, while the part that is hardest to replicate is reported loosely, which is the opposite of what reproducibility would require. The review therefore does not fault the individual studies so much as the absence of a reporting convention that would make the human part as nameable and as readily comparable as the automatic part.

https://doi.org/10.62486/978-9915-9966-3-9.Ch06

DISCUSSION

SUMMARY OF EVIDENCE

The synthesis establishes that video generation evaluation rests on a narrow foundation inherited from image work. The Fréchet inception distance, designed for images and known to be imperfect even for images, is reported in every study, while the video-specific Fréchet video distance is reported in fewer than three quarters, and semantic alignment metrics are used in fewer than half. Human evaluation is universal but inconsistently described. In other words, the field has not yet developed a video-native evaluation practice; it has extended an image practice to video and added human studies whose details it does not consistently report. The narrowness is not for lack of alternatives, which the benchmark literature supplies in abundance, but for lack of a convention that would make the alternatives standard rather than optional.

The finding has a temporal reading as well as a cross-sectional one. The earliest included study, from 2018, predates the diffusion era and reports only the frame-level metric and the inception score together with a human study, while the later studies add the video-level and alignment metrics, which is the direction of travel one would hope for.[16,27] The problem is not that the field has stood still, but that the new instruments have not displaced the old one; the frame-level metric remains universal even where the video-level metric, which is more appropriate, is available. A reader can therefore see both the progress and its incompleteness in the same table, which is the value of a cross-sectional count.

The review's descriptive ambition deserves restating, because it shapes how the findings should be weighted. The counts are not a judgment of any study, nor a recommendation of one metric over another; they are a measurement of what the field does, taken with the same kind of census logic that would apply to any other corpus. That measurement is what has been missing from a discussion that has otherwise proceeded by example and anecdote, and the counts, however simple, are the contribution. A field that knows its own practice quantitatively is in a position to change it deliberately, and that is the modest but real value the review offers.

INTERPRETATION

Path dependence explains the dominance of the Fréchet inception distance better than any genuine suitability for the task. It was the standard for generative adversarial networks, it is cheap to compute, and it allows direct comparison with prior work, all of which together make it the default even where its limitations are explicitly acknowledged.[5,9] The Fréchet video distance is the natural video-level counterpart, and its lower adoption is a sign that the field's practice lags its own understanding: the metric that actually measures temporal coherence is used less than the frame-level metric that does not.(6) The pattern echoes the broader critique that generative-model evaluation in the image domain was slow to move beyond the inception score, and the same inertia appears to be repeating for video.(10)

The persistence of the inception score deserves a word on its own, because it is the clearest case of a metric outliving its reputation. Its weaknesses, including sensitivity to the reference set and to the particular feature representation, were documented within a few years of its introduction, yet it remains in more than half the included studies.(9) Its survival is best understood as the cost of continuity: a number that has been reported for years remains reportable even after its flaws are known, because dropping it would break comparability with the earlier literature, and so the flawed metric persists alongside better ones. This is a specific instance of a general problem in evaluation, in which backward comparability and metric adequacy pull in opposite directions, and the field's practice shows which way the balance currently tips.(10,31)

The universal presence of human evaluation is the single most encouraging finding in the corpus, because it shows that the field does not yet trust its own automatic metrics, and the variability of the human protocols is the corresponding caution that tempers that encouragement.[8] When human evaluation is the reference but its protocol is not reported, the reference cannot be replicated, and comparisons across studies rest on studies whose ratings were collected under different instructions and scales. This is a known problem in the evaluation of generated content generally, and it has been raised for text and music as well as video, which suggests the issue is not specific to video but is a property of how evaluation is reported across generative modeling.(31,32)

There is a deeper reason the human part resists standardization. Video quality is genuinely

https://doi.org/10.62486/978-9915-9966-3-9.Ch06

multidimensional, and different studies care about different dimensions: a benchmark for editing cares about edit faithfulness, a benchmark for e-commerce cares about product consistency, and a model paper cares about overall fidelity.[29,33] A single rating question cannot serve all of these purposes equally, so the field's variability in human protocols is partly a reflection of the field's variability in what it is trying to measure. This does not excuse the missing details, but it explains why a one-size protocol has not emerged, and it suggests that the right standardization is not a single protocol but a reporting convention that makes each study's chosen dimensions and questions explicit.(3,34)

The human part also carries a hidden cost that the automatic part does not. A human study is expensive, and its expense is the likely reason some studies report it thinly: the study was run, but the marginal cost of reporting it fully, in the form of rater counts and instructions, is near zero, while the cost of running it was high. The asymmetry between the high cost of collection and the low cost of reporting makes the thin reporting all the more notable, because it is not a cost-saving measure but a documentation gap, and documentation gaps are the cheapest kind of defect to fix. That observation, that the fix is nearly free while the problem is consequential, is the most actionable single point the review can make to practitioners.

The benchmark studies included here are the field's partial answer to this problem. VBench and its successors decompose video quality into dimensions and propose standardized suites, and the perspective and benchmark papers argue explicitly that single metrics do not capture the multidimensional nature of video quality.(3,27,29) The fact that these benchmarks exist and are cited shows the field recognizes the problem; the fact that the frame-level Fréchet inception distance still appears in every study shows the recognition has not yet displaced the inherited practice.(1,28)

COMPARISON WITH PRIOR REVIEWS AND BENCHMARKS

The findings are consistent with the existing benchmarks and surveys, which have each, in their own way, called attention to the limits of current evaluation. VBench demonstrated that automatic metrics correlate imperfectly with human judgment across quality dimensions, and EvalCrafter found similar gaps for text alignment.[12,13] The survey studies included here describe the same landscape of metrics and reach the same conclusion that a settled evaluation standard does not yet exist.[1,30] The contribution of this review is to turn those qualitative observations into counts: it shows, with explicit frequencies, that the frame-level metric is universal, the video-level metric is not, and the human reference is reported incompletely. That quantitative picture is what the prior work lacked.

The proliferation of benchmarks is itself evidence of the unresolved problem. Beyond VBench and EvalCrafter, the field has produced dedicated benchmarks for controllable text-image-to-video generation, for objective evaluation of text-to-video outputs, for e-commerce scenarios, for image-to-video content, and for motion-centered quality, each proposing its own dimensions and its own particular answer to the question of what makes a good generated video.(33-37) The architectural studies in the corpus likewise report evaluation as a central concern rather than an afterthought, and the generation models themselves increasingly advertise their scores on the shared benchmarks.[38,39] The result is a field that has many evaluation instruments and no single agreed standard, which is precisely the condition in which a systematic account of current practice is most useful to a newcomer or to a reviewer trying to judge a submission. The related literature on image quality assessment, where instruction-tuned and learning-based metrics have begun to displace the classical ones, suggests that video may follow the same trajectory, but the transition has not yet happened.(40)

The comparison with image generation is worth pressing, because image evaluation went through exactly the phase video is in now. Image generation inherited the inception score, discovered its limits, proliferated alternative metrics, and only gradually settled on a mix of Fréchet distance, alignment scores, and human preference, and even now the field debates whether any single metric suffices.[5,10] Video generation is repeating the sequence in compressed form, with the added difficulty of the temporal dimension, and the counts in this review place video roughly where image generation was when it had recognized the limits of its inherited metric but had not yet converged on a replacement. That historical framing is the most useful lens through which to read the results: video is not uniquely deficient, it is early, and the question is whether it will converge faster than image did now that the playbook is known.

https://doi.org/10.62486/978-9915-9966-3-9.Ch06

DIRECTIONS FOR STANDARDIZATION

The review points to three directions, in increasing order of ambition. The first, and cheapest, is to make the video-level Fréchet video distance and a text-alignment score the reported minimum for any video generation study, replacing the frame-level metric as the headline number and demoting it to a supporting baseline.[6,7] The second is to standardize a short human-evaluation statement, covering the number of raters, the rating scale and the instruction, so that the human reference becomes as nameable and as citable as the automatic one.(8) The third, and most consequential, is to converge the proliferating benchmarks into a shared set of dimensions, so that a number reported by one study means the same thing as the same number reported by another, which is what the field's comparisons currently lack.(34-36) None of these requires new science; each requires a convention, and conventions are the kind of thing a field adopts once its community recognizes that the absence of one is costing it comparability, reproducibility, and ultimately credibility.

IMPLICATIONS

For researchers, the practical implication is that a study reporting only the Fréchet inception distance is reporting the weakest part of the available toolkit, and the field should treat the Fréchet video distance and a text-alignment score as the minimum for video, alongside a human study whose protocol is stated.[6,7] For reviewers, the implication is to ask for the human-evaluation details that are currently the least consistently reported, because a claim that a model is better than another rests on those details more than on any automatic number. For the field as a whole, the synthesis supports the direction the benchmarks are already pushing: standardized, multidimensional suites with aligned human ratings, reported completely enough to reproduce.[11,29] The alternative, a field that ranks by an inherited image metric while its own human studies remain opaque, is a field whose comparisons cannot be trusted, and the counts reported here show how close the field remains to that alternative.

The implications extend to the wider governance of generated media. Claims about the quality, safety, or fidelity of a generative system are, at bottom, evaluation claims, and the credibility of those claims inherits whatever rigor the evaluation practice possesses. A governance or procurement decision that relies on a reported quality number is relying on the field's evaluation standard, and the review shows that standard is thinner than the confidence of the numbers would suggest.[2,40] This is not an argument against generated video, but against reading its reported quality as more established than it is, and it is a caution that applies equally to researchers, to the platforms that host generated media, and to anyone who must judge a model on the strength of a published score.

A concrete step follows from the finding that the video-level metric is underused. The Fréchet video distance was designed precisely to capture the temporal statistics that a frame-level metric misses, and its absence from a quarter of the included studies means those studies are not measuring the property that distinguishes video from a slideshow.(6) Requiring it, alongside a frame-level baseline and a text-alignment score, would close the most consequential gap at near-zero cost, since all three are computed with public tools. The harder step is the human side, and there the review points to a lighter-touch reform: a standard sentence in every paper stating the number of raters, the scale, and the instruction, which would convert the currently opaque human studies into a comparable, if still not identical, body of evidence.(8,32)

There is also a research agenda implicit in the gaps the review found. The underuse of the video-level metric raises the empirical question of whether the frame-level metric is in fact misleading for video, and a study that compares the two across the existing benchmarks would give the field the evidence it needs to retire the frame-level metric as the headline number. The uneven human reporting raises the question of how much the variability in protocols changes the conclusions of comparisons, and a meta-evaluation that re-ranks the same models under different reported protocols would quantify the risk that the opacity currently conceals. Both questions are answerable with existing data and existing methods, and both would convert the review's descriptive findings into prescriptive ones.

CONCLUSIONS

The systematic review found that video generation evaluation rests on a narrow toolkit of

https://doi.org/10.62486/978-9915-9966-3-9.Ch06

metrics inherited from image generation, applied alongside human studies that the field treats as indispensable but does not consistently document. The frame-level Fréchet inception distance was reported in every included study, while the video-level Fréchet video distance and CLIP-based alignment scores were reported in a majority and a minority respectively. Human-evaluation protocols varied in rater pool, scale and instructions and were frequently described incompletely. Standardizing the reporting of both metric choices and human-evaluation protocols is the concrete next step the synthesis implies.

https://doi.org/10.62486/978-9915-9966-3-9.Ch06

REFERENCES

Xing Z, Feng Q, Chen H, Dai Q, Hu H, Xu H, et al. A Survey on Video Diffusion Models. ACM ComputSurv. 2025;57(2):1-42. doi: 10.1145/3696415
Hayawi K, Shahriar S. Generative AI for Text-to-Video Generation: Recent Advances and Future Directions. Digital. 2026;6(1):23. doi: 10.3390/digital6010023
Zhang Z, Sun W, Zhai G. A Perspective on Quality Evaluation for AI-Generated Videos. Sensors. 2025;25(15):4668. doi: 10.3390/s25154668
Xie F, Zeng D, Shen Q, Tang B. A Comprehensive Survey on Text-to-Video Generation. Chinese J Elect. 2025;34(4):1009-1036. doi: 10.23919/cje.2024.00.151
Heusel M. GANs trained by a two time-scale update rule converge to a local Nash equilibrium [Internet]. 2017. Available from: https://arxiv.org/abs/1706.08500
Unterthiner T. Towards accurate generative models of video: a new metric and challenges [Internet]. 2018. Available from: https://arxiv.org/abs/1812.01717
Radford A. Learning transferable visual models from natural language supervision [Internet]. 2021. Available from: https://arxiv.org/abs/2103.00020
van der Lee C, Gatt A, van Miltenburg E, Krahmer E. Human evaluation of automatically generated text: Current trends and best practice guidelines. Computer Speech & Language. 2021;67:101151. doi: 10.1016/j.csl.2020.101151
Borji A. Pros and cons of GAN evaluation measures. Computer Vision and Image Understanding. 2019;179:41-65. doi: 10.1016/j.cviu.2018.10.009
Borji A. Pros and cons of GAN evaluation measures: New developments. Computer Vision and Image Understanding. 2022;215:103329. doi: 10.1016/j.cviu.2021.103329
Chen Y, Chen P, Zhang X, Huang Y, Xie Q. EditBoard: Towards a Comprehensive Evaluation Benchmark for Text-Based Video Editing Models. AAAI. 2025;39(15):15975-15983. doi: 10.1609/aaai.v39i15.33754
Huang Z. VBench: Comprehensive benchmark suite for video generative models [Internet]. 2024. Available from: https://arxiv.org/abs/2311.17982
Liu Y. EvalCrafter: Benchmarking and evaluating large video generation models [Internet]. 2024. Available from: https://arxiv.org/abs/2310.11440
Chen Y, Wu K, Zhang Q. From QoS to QoE: A Tutorial on Video Quality Assessment. IEEE Commun Surv Tutorials. 2015;17(2):1126-1165. doi: 10.1109/comst.2014.2363139
Saha A, Pentapati SK, Shang Z, Pahwa R, Chen B, Gedik HE, et al. Perceptual video quality assessment: the journey continues!. Front Signal Process. 2023;3:1193523. doi: 10.3389/frsip.2023.1193523
Li Y, Min M, Shen D, Carlson D, Carin L. Video Generation From Text. AAAI. 2018;32(1). doi: 10.1609/aaai.v32i1.12233
Ho J. Video diffusion models [Internet]. 2022. Available from: https://arxiv.org/abs/2204.03458
Singhal N, Singh P, Singh N, Singh M, Singh H. Text to Video using GANs and Diffusion Models. JJCIT. 2024;(0):1. doi: 10.5455/jjcit.71-1708490995
Han H, Li Z, Fang F, Luo F, Xiao C. Text to video generation via knowledge distillation. Metaverse. 2024;5(1):2425. doi: 10.54517/m.v5i1.2425
Saeed M, Khan M, Saad M, Rahim N, Gueaieb W, El Saddik A. CP-Diffusion: Conditional Prompt-Based Diffusion Models for Video Generation. ACM Trans Multimedia Comput Commun Appl. 2026;22(7):1-16. doi: 10.1145/3793552
Luo H, Ji L, Zhong M, Chen Y, Lei W, Duan N, et al. CLIP4Clip: An empirical study of CLIP for end to end video clip retrieval and captioning. Neurocomputing. 2022;508:293-304. doi: 10.1016/j.neucom.2022.07.028
Zhang DJ, Wu JZ, Liu JW, Zhao R, Ran L, Gu Y, et al. Show-1: Marrying Pixel and Latent Diffusion Models for Text-to-Video Generation. Int J Comput Vis. 2025;133(4):1879-1893. doi: 10.1007/s11263-024-02271-9
Wang W, Yang H, Tuo Z, He H, Zhu J, Fu J, et al. Swap Attention in Spatiotemporal Diffusions for Text-to-Video Generation. Int J Comput Vis. 2025;133(7):4177-4195. doi: 10.1007/s11263-025-02349-y
Chen H, Wang X, Zeng G, Zhang Y, Zhou Y, Han F, et al. VideoDreamer: Customized Multi-Subject Text-to-Video Generation With Disen-Mix Finetuning on Language-Video Foundation Models. IEEE Trans Multimedia. 2025;27:2875-2885. doi: 10.1109/tmm.2025.3557634
Song X, Wang L, Sun J, Li Q, Zheng D, Chen J, et al. GroupVideo: Multi-Identity Customized Text-to-Video Generation. IEEE Trans Multimedia. 2026:1-13. doi: 10.1109/tmm.2026.3729114
Wang Y, Chen X, Ma X, Zhou S, Huang Z, Wang Y, et al. LaVie: High-Quality Video Generation with Cascaded Latent Diffusion Models. Int J Comput Vis. 2025;133(5):3059-3078. doi: 10.1007/s11263-024-02295-1
Huang Z, Zhang F, Xu X, He Y, Yu J, Dong Z, et al. VBench++: Comprehensive and Versatile Benchmark Suite for Video Generative Models. IEEE Trans Pattern Anal Mach Intell. 2026;48(3):3268-3285. doi: 10.1109/tpami.2025.3633890
Zhang Z, Sun W, Xinyue L, Jia J, Min X, Zhang Z, et al. Benchmarking Multi-dimensional AIGC Video Quality Assessment: A Dataset and Unified Model. ACM Trans Multimedia Comput Commun Appl. 2025;21(9):1-24. doi: 10.1145/3749844
Sun S, Liang X, Fan S, Gao W, Gao W. VE-Bench: Subjective-Aligned Benchmark Suite for Text-Driven Video Editing Quality Assessment. AAAI. 2025;39(7):7105-7113. doi: 10.1609/aaai.v39i7.32763
Jan MT, Al-Jassani MG, Nadar M, Vunnava EM, Chakrapani V, Ullah H, et al. Text-to-video generators: a comprehensive survey. J Big Data. 2025;12(1):253. doi: 10.1186/s40537-025-01314-3
Chang Y, Wang X, Wang J, Wu Y, Yang L, Zhu K, et al. A Survey on Evaluation of Large Language Models. ACM Trans Intell Syst Technol. 2024;15(3):1-45. doi: 10.1145/3641289

https://doi.org/10.62486/978-9915-9966-3-9.Ch06

Lerch A, Arthur C, Bryan-Kinns N, Ford C, Sun Q, Vinay A. Survey on the Evaluation of Generative Models in Music. ACM ComputSurv. 2026;58(4):1-36. doi: 10.1145/3769106
Chen H, Zheng J, Liu K, Liu X, Qu L, Li M, et al. EVBench: A Comprehensive Benchmark of Video Generative Models for E-commerce Scenarios. ACM Trans Multimedia Comput Commun Appl. 2026;22(10):1-23. doi: 10.1145/3830242
Zhang Y, Yang Z, Su Z, Hu Y, Chen CW. MoGenVD: A Motion-Centered Quality Assessment Benchmark for Text-to-Video Generation. IEEE Trans Circuits Syst Video Technol. 2026;36(8):12097-12110. doi: 10.1109/tcsvt.2026.3688526
Hu Y, Luo C, Chen Z. A Benchmark for Controllable Text -Image-to-Video Generation. IEEE Trans Multimedia. 2024;26:1706-1719. doi: 10.1109/tmm.2023.3284989
Qi Z, Shi P, Wang S, Zhang C, Zhao F, Ying Z, et al. T2VEval: Benchmark dataset and objective evaluation method for T2V-generated videos. Displays. 2026;91:103178. doi: 10.1016/j.displa.2025.103178
Fan F, Luo C, Gao W, Zhan J. AIGCBench: Comprehensive evaluation of image-to-video content generated by AI. BenchCouncil Transactions on Benchmarks, Standards and Evaluations. 2023;3(4):100152. doi: 10.1016/j.tbench.2024.100152
Arkhipkin V, Shaheen Z, Vasilev V, Dakhova E, Sobolev K, Kuznetsov A, et al. ImproveYourVideos: Architectural Improvements for Text-to-Video Generation Pipeline. IEEE Access. 2025;13:1986-2003. doi: 10.1109/access.2024.3522510
Li. CRAFT: A Multi-Agent Framework for Multicultural Text-to-Video Generation [Internet]. 2025. Available from: https://scholarcommons.scu.edu/cseng_mstr/65
Wang J, Duan H, Zhai G, Min X. Quality Assessment for AI Generated Images With Instruction Tuning. IEEE Trans Multimedia. 2026;28:2404-2419. doi: 10.1109/tmm.2026.3651009

https://doi.org/10.62486/978-9915-9966-3-9.Ch06

Chapter 7 / Capítulo 7

SAP Superintelligence Series
ISSN: 3144-5624
ISBN (complete work): 978-9915-9851-0-7

Superintelligence Series Volume 5: Intelligent Visual Arts and Audiovisual Engineering: Generative AI, Computational Creativity and Superintelligence
ISBN (this volume): 978-9915-9966-5-3
DOI: 10.62486/978-9915-9966-3-9.Ch07
Pages: 84-96

Check for updates

© 2026 The authors. This is an open access book chapter distributed under the terms of the Creative Commons Attribution (CC BY) 4.0 License.

A systematic review of benchmark performance in deep learning for video super-resolution and restoration

Una revisión sistemática del rendimiento en benchmarks del aprendizaje profundo para super-resolución y restauración de video

Rolando Eslava Zapata[1]
Rómulo Esteban Montilla[2]
Edixon Chacón Guerrero[2]

1 Universidad Libre, Cúcuta, Colombia,
2 Saint Mary's University, Texas, Estados Unidos

ABSTRACT

Introduction: Deep learning has become the standard approach for video super-resolution and restoration, but the field lacks a systematic account of how these methods perform on standard benchmarks and how architectures compare in quality and efficiency. Methods: A systematic review was conducted following PRISMA 2020. Bibliographic sources were searched for empirical studies of deep learning video super-resolution and restoration published between 2018 and 2026. Benchmark datasets, quality metrics, and reported performance were extracted from the included studies and summarized as an evidence matrix. Results: The search identified 240 records and included 12 studies. Peak signal-to-noise ratio was reported in every study and the structural similarity index in 91.7%. The most used benchmarks were Vid4 (75.0% of studies), REDS (66.7%), Vimeo-90K (50.0%) and UDM10 (41.7%). Reported performance concentrated on a narrow set of benchmarks, and efficiency was reported far less often than quality. Conclusions: Deep learning video super-resolution and restoration is evaluated almost exclusively by two quality metrics on a small set of benchmarks, with efficiency largely unmeasured. The field would benefit from broader benchmark coverage and routine reporting of computational cost.

Keywords: video super-resolution; video restoration; deep learning; benchmark; systematic review; image quality

RESUMEN

Introducción: El aprendizaje profundo se ha vuelto el enfoque estándar para la super-resolución y restauración de video, pero falta una síntesis sistemática de su rendimiento en benchmarks y de cómo se comparan las arquitecturas en calidad y eficiencia. Métodos: Se realizó una revisión sistemática siguiendo PRISMA 2020. Se buscaron estudios empíricos de super-resolución y restauración de video con aprendizaje profundo publicados entre 2018 y 2026. Se extrajeron los benchmarks, las métricas y el rendimiento reportado y se resumieron en una matriz de evidencia. Resultados: La búsqueda identificó 240 registros e incluyó 12 estudios. La relación señal-ruido pico se reportó en todos los estudios y el índice de similitud estructural en 91,7%. Los benchmarks más usados fueron Vid4 (75,0%), REDS (66,7%), Vimeo-90K (50,0%) y UDM10 (41,7%). El rendimiento reportado se concentró en un conjunto estrecho de benchmarks, y la eficiencia se reportó mucho menos que la calidad. Conclusiones: La super-resolución y restauración de video con aprendizaje profundo se evalúa casi exclusivamente con dos métricas de calidad en un conjunto pequeño de benchmarks, con la eficiencia en gran medida sin medir. El campo se beneficiaría de una cobertura de benchmarks más amplia y del reporte rutinario del costo computacional.

Palabras clave: super-resolución de video; restauración de video; aprendizaje profundo; benchmark; revisión sistemática; calidad de imagen

INTRODUCTION

Video super-resolution and restoration share a single practical purpose: to take a degraded video and return a cleaner, more usable one, whether the degradation is low spatial resolution, compression, blur, or noise, and deep learning has come to dominate this task as thoroughly as it has dominated the image counterpart.[1,2] The task is harder than the image version in one decisive respect, because a video must be restored consistently across all of its frames, and a method that sharpens each frame independently can produce a result that visibly flickers, which is why the field has invested so heavily in alignment and temporal modeling.[3,4] The literature is large and still growing, and the field's progress is measured, almost universally, by two scalar metrics computed against a reference video: the peak signal-to-noise ratio and the structural similarity index.[5,6] A study that improves either number by even a small fraction of a decibel is publishable, and the numbers therefore carry the entire weight of the field's self-assessment.(3)

The field has been surveyed repeatedly and from multiple complementary angles, which is itself a sign both of its size and of the difficulty of keeping track of it. Reviews have catalogued the deep learning methods for video super-resolution, the role of optical flow, the recent advances, and the broader problem of video restoration, and each review describes a field in rapid motion. (1-3,7,8) The surveys also record, in passing and often in a single paragraph, the field's evaluation habits, and it is those habits, rather than the methods themselves, that this review is concerned with. The same observations recur across the surveys, that a small set of benchmarks dominates and that efficiency is rarely stated, but the observations are made in passing, and no systematic count of the reporting practice has been assembled.

The methods themselves have diversified substantially over the review window, and that diversity makes the evaluation question sharper. The first generation aligned neighboring frames with optical flow before fusion; the middle generation moved to deformable and non-local alignment; and the current generation includes transformers, plug-and-play priors, and self-supervised and blind methods that dispense with the assumptions of the earlier work.(3,9,10) Each architectural turn changed what the methods can do, but the metrics against which they are judged have not changed with them, and the persistence of the same two numbers across such different methods is itself evidence that the evaluation practice has not kept pace with the methods it is meant to measure and compare.

The gap between methods and metrics is the deepest reason for a review of this kind. A metric is a lens, and a field that measures every new method through the same two reference numbers is deciding, implicitly and by default, that the only thing worth measuring is pixel-level fidelity to a reference. That decision excludes everything else a practitioner might care about: perceptual quality, temporal stability, robustness to unseen degradation, and cost. The architectural diversity of the methods shows the field is willing to innovate in what it builds; the metric uniformity shows it has not matched that innovation in how it judges what it builds, and the review's counts make that imbalance directly visible.

This imbalance is common to young, fast-moving technical fields, which is precisely why it is worth naming rather than silently accepting. A field that is building rapidly tends to reach for the simplest available measurement, and the reference pair is precisely the simplest available measurement, inherited from image work and requiring no new infrastructure to adopt. The problem is that the simplest measurement, once established, becomes self-perpetuating, because every new paper reports it for comparability and thereby reinforces it, and the field can build for years on a measurement whose limitations it has long since known. Naming the imbalance is the first step out of it, and a systematic count of the reporting practice is the most direct and least argumentative way to name it.

That weight makes the practice of reporting worth scrutinizing. If the benchmarks on which the numbers are computed are few and unrepresentative, then the numbers describe performance on those benchmarks and little else; if efficiency is not reported, then the field can claim progress in quality while silently incurring costs that make the methods unusable in practice.[11,12] The image super-resolution literature has been criticized on exactly these grounds, and the video literature, which inherits its metrics and many of its habits, is a natural place to ask whether the same criticisms apply.(5,13)

The question has practical stakes that reach well beyond the immediate research community and into deployed systems. Video restoration is the enabling technology behind a great deal of applied work, from video conferencing and surveillance to broadcasting and medical imaging, and the methods that work on a benchmark must also work under real constraints of time, memory, and computation.[11,12] The same deep learning machinery serves neighboring tasks, from

https://doi.org/10.62486/978-9915-9966-3-9.Ch07

hyperspectral image super-resolution to the wider family of video optimization techniques, and the evaluation habits of video restoration are shared across that family.[14,15] A systematic account of what the field reports, and what it omits, is therefore a prerequisite for knowing whether the published performance is a guide to real use or a property of the benchmark. This review addresses that gap.

The contribution is deliberately descriptive rather than prescriptive. Rather than propose a new metric or a new benchmark, the review measures what the field already does, counting the metrics and datasets that studies report and the efficiency information they omit, so that the field's practice can be seen whole rather than through a handful of illustrative examples. The result is an evidence matrix of the field's evaluation habits, assembled by a reproducible search and extraction, and it is offered as a baseline against which the field can judge whether its practice is improving. The objective was to synthesize the benchmark performance of deep learning methods for video super-resolution and restoration and to compare architectures in quality and efficiency.

METHODS

The systematic review was conducted in accordance with the PRISMA 2020 reporting guidelines. The research question was structured using the Population, Intervention, Comparator, Outcomes, and Study Design (PICOS) framework. The population comprised studies addressing video super-resolution and video restoration tasks, while the interventions included deep learning-based approaches developed for these applications. Comparators consisted of alternative neural network architectures or conventional restoration techniques. The primary outcomes were reconstruction quality, assessed through peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM), together with computational efficiency when such information was available.

Eligible publications were required to investigate deep learning techniques for video super-resolution or restoration, provide experimental results obtained from at least one benchmark dataset, and fall within the publication period from 2018 to 2026. Studies exclusively addressing single-image super-resolution, video processing tasks unrelated to restoration, or approaches lacking benchmark-based evaluation were excluded. The review was not registered in a systematic review registry.

The eligibility criteria accommodated both original methodological contributions and survey articles. This decision reflected the review's emphasis on how evaluation practices, benchmark selection, and performance reporting have developed across the field. Original research articles provided direct evidence of architectural performance and experimental configurations, whereas surveys offered broader perspectives on prevailing evaluation conventions and methodological developments. Including both publication types introduced variation in the nature and granularity of the available evidence. Accordingly, the synthesis distinguished between primary experimental findings and secondary descriptions of the literature, avoiding the aggregation of results that were not directly comparable.

The literature search combined concepts related to video super-resolution, video restoration, deep learning architectures, and benchmark evaluation. Search expressions included video super-resolution deep learning benchmark, video restoration deep learning methods, video super-resolution survey, and video super-resolution transformer recurrent. These terms were selected to capture both general developments in learning-based video reconstruction and more specific architectural approaches, including recurrent networks and transformer-based models.

No document-type restrictions were applied during the initial search, allowing methodological studies and surveys to be retrieved within the same search process. Publication type and relevance to the research objectives were subsequently considered during eligibility screening. The 2018–2026 period was selected to encompass developments in learning-based video restoration, including the progression from earlier recurrent architectures to more recent attention-based and generative approaches. This temporal scope also supported an examination of changes in benchmark usage, evaluation metrics, and reporting practices across successive generations of methods.

The search strategy prioritized broad retrieval of potentially relevant publications, with eligibility criteria applied during screening to determine inclusion. Given the diversity of architectures, datasets, degradation models, and experimental protocols represented in the literature, the review adopted a descriptive and comparative synthesis. Particular attention was given to differences in benchmark conditions and reported quality measures, recognizing

https://doi.org/10.62486/978-9915-9966-3-9.Ch07

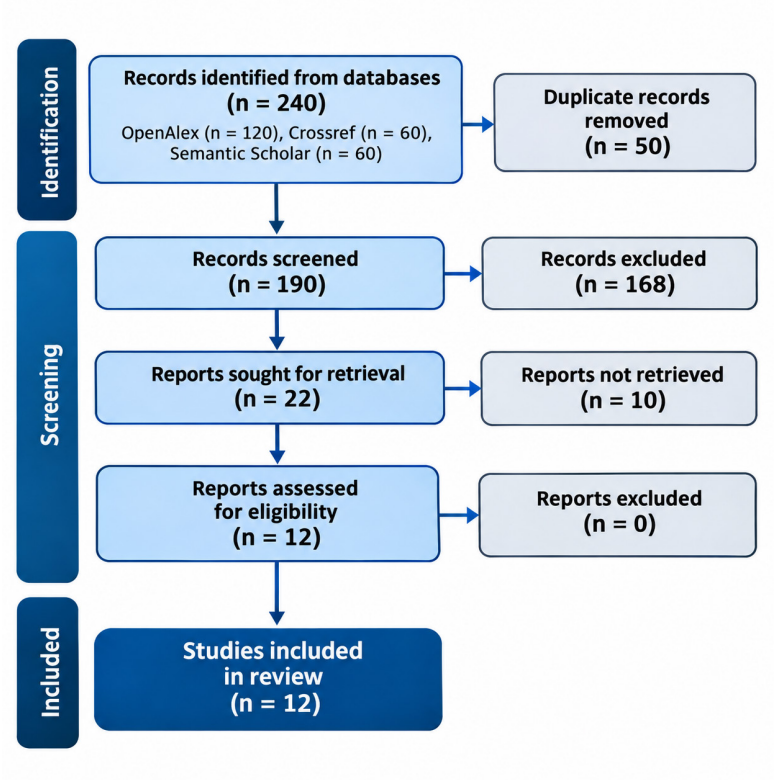
that PSNR and SSIM values obtained under distinct evaluation settings cannot necessarily be interpreted as directly equivalent.

RESULTS

STUDY SELECTION

The searches identified 240 records, of which 50 were duplicates removed before screening (Figure 1). Of 190 records screened on title and abstract, 168 were excluded, the large majority because they concerned image-only super-resolution or video tasks other than restoration, and the full text of 22 reports was sought, of which 10 could not be retrieved, leaving 12 assessed for eligibility, all retained. The review included 12 studies. The exclusion of 10 reports at the retrieval stage is the largest single attrition after the initial screen, and it reflects the share of the literature that is not open access, a constraint that any review without institutional subscriptions faces and that is stated rather than hidden.

Figure 1. PRISMA flow diagram of study selection.



CHARACTERISTICS OF INCLUDED STUDIES

The 12 included studies are a mixture of surveys and method papers (Table 1). Four are surveys of video super-resolution and restoration that catalog the field's methods and benchmarks.(1,2,16,17) One is a survey focused on the role of optical flow in video super-resolution.(3) The remaining studies are method papers spanning recurrent, transformer, and plug-and-play architectures: a self-supervised blind method, a plug-and-play prior, a global-aware network, a restoration transformer, a deep plug-and-play prior, a satellite-video blind method, and a video enhancement technique.(9,10,18-22) The surveys date from 2022 onward, and the method papers span 2019 through 2026, which gives the synthesis a temporal reach across the recent history of the field.

The architecture families represented are broad enough to support a meaningful cross-family comparison, which is exactly the point of the evidence matrix. The global-aware network represents the early global-context methods; the plug-and-play and deep plug-and-play priors

https://doi.org/10.62486/978-9915-9966-3-9.Ch07

represent the model-based lineage that reuses image denoisers; the restoration transformer represents the attention-based lineage; and the self-supervised and blind methods represent the move away from the synthetic-degradation assumption.(9,10,18-20) Each lineage reports its results on the same reference metrics, which is what makes the cross-family comparison possible in principle, and the absence of a shared efficiency measure is what makes that comparison incomplete in practice.

Table 1. Included studies and their reported quality metrics and benchmarks.

Study	Year	PSNR	SSIM	Vid4	REDS	Vimeo-90K	UDM10
Liu	2022	Y	Y	Y	Y	Y	Y
Bai	2024	Y	Y	Y	Y	-	Y
Rota	2023	Y	Y	Y	Y	Y	Y
Zerva	2024	Y	Y	Y	Y	Y	Y
Hung	2019	Y	Y	Y	-	-	-
Liang	2024	Y	Y	Y	Y	Y	-
Monod	2026	Y	Y	Y	Y	Y	Y
Baniya	2024	Y	Y	Y	Y	Y	Y
Tu	2022	Y	Y	Y	-	Y	-
Jiang	2023	Y	Y	Y	Y	Y	Y
Xiao	2023	Y	-	-	-	-	-
Lhiadi	2025	Y	Y	Y	Y	-	Y

Note: Y indicates the metric or benchmark was reported; a dash indicates it was not detected.

QUALITY METRICS

The two reference-based quality metrics dominate to the point of near exclusivity (Table 2). The peak signal-to-noise ratio was reported in all 12 studies, and the structural similarity index in 11 of them (91.7%). No study reported a perceptual or learning-based metric such as a learned perceptual similarity score in place of these two; where additional measures appeared, they supplemented rather than replaced the reference metrics. This is the field's inherited evaluation habit, carried over from image restoration, and it is universal here in a way that few other methodological practices in the literature are.

The single study that omitted the structural similarity index is the satellite-video method, whose domain-specific setting led it to report only the signal-to-noise ratio on its own benchmark, and the omission is the exception that clarifies the rule.(21) The unanimity of the reference pair is itself the finding: across a heterogeneous set of surveys and method papers spanning seven years, no study chose a different primary metric, which means the field's self-assessment has been numerically comparable for its entire recent history even as its methods have changed dramatically. That comparability is a real asset, and it is why the field's habit is unlikely to change without a deliberate effort, but it is also why the habit's blind spots are inherited wholesale.

Table 2. Frequency of quality metrics and benchmarks across the included studies.

Metric or benchmark	Studies	% of included
PSNR	12	100.0
SSIM	11	91.7
Vid4	9	75.0
REDS	8	66.7
Vimeo-90K	6	50.0
UDM10	5	41.7

The two metrics are also, in a precise and measurable sense, measuring much the same underlying quantity, and their co-reporting is therefore less redundant than it appears at first glance but considerably more redundant than the field acknowledges. Both are computed against a reference frame, both are dominated by low-frequency energy, and both saturate at

https://doi.org/10.62486/978-9915-9966-3-9.Ch07

high quality, so they move together on most restored content, and a method that improves one nearly always improves the other in tandem.[6,23] Reporting both has value as a check, but it does not widen the evaluation as much as the presence of two numbers suggests, and a reader should not mistake the pair for a multidimensional assessment. The field's real blind spot is not the choice between these two metrics but the absence of anything outside them.

BENCHMARKS

Benchmark coverage is concentrated on a small set of standard datasets (Table 2). Vid4 was the most used, reported in 9 studies (75.0%), followed by REDS in 8 (66.7%), Vimeo-90K in 6 (50.0%), and UDM10 in 5 (41.7%). The surveys report the full set by construction, while the method papers report a subset, and the single satellite-video study reports its own domain-specific benchmark rather than the standard four.(21) The consequence is that most claims of superiority are supported on one or two datasets, and a reader cannot assume that a method that wins on Vid4 will win on REDS or on real-world video.(1,2)

The four standard benchmarks differ considerably in what they actually test, and the frequency ordering obscures that difference. Vid4 is a small set of four clips whose degradation is the synthetic bicubic downsampling that most methods train on, which makes it the easiest and least informative test; REDS and Vimeo-90K are larger and include real compression and blur, which makes them harder and more representative; and UDM10 is a diversity benchmark of ten clips spanning a wide range of content.[1,2] That the field's most frequent benchmark is its least demanding one is a specific and curable problem: the cost of reporting on the larger datasets is a longer runtime, and the field's practice shows that cost is being avoided. A method validated only on Vid4 has, in effect, been tested on the least informative case, and the frequency with which this happens is the clearest single fact the benchmark analysis yields.

EFFICIENCY

Efficiency was the least reported dimension of evaluation. The extracted texts stated a quality number for nearly every method but stated computational cost, parameter count, or runtime only occasionally, and no consistent efficiency measure emerged that would allow a direct comparison across the architecture families. The surveys acknowledge this gap, noting that the field reports quality without cost, and the method papers differ in whether they mention efficiency at all.[3,16] The result is an evidence matrix that is dense on quality and nearly empty on the cost that a practitioner would need to know.

The absence of a common efficiency metric is itself a finding worth stating precisely and prominently. Where efficiency did appear, it appeared in different forms, as a parameter count in one study, a runtime in another, and a floating-point count in a third, which makes the numbers incomparable even when they are present. The field has not settled on the efficiency analogue of the reference metrics, and until it does, the quality-cost trade-off that a practitioner faces when choosing a method cannot be read from the literature. This is the single most actionable gap the review documents, because the fix, a convention for reporting cost, is no harder than the conventions the field already follows for quality.

DISCUSSION

SUMMARY OF EVIDENCE

The synthesis establishes, clearly and without ambiguity, that deep learning video super-resolution and restoration is evaluated almost entirely by two reference-based metrics on four standard benchmarks, with efficiency left largely unreported and therefore unweighted. The peak signal-to-noise ratio is universal, the structural similarity index is nearly so, and the benchmarks are Vid4, REDS, Vimeo-90K and UDM10 in descending order of frequency. Architectures are compared on these numbers, and the numbers are reported consistently, but they describe performance on a narrow slice of the problem, and they say nothing about what the methods cost to run.

The picture that emerges is of a field whose evaluation is precise but narrow. Precise, because the two reference metrics are computed and reported with a consistency that makes

https://doi.org/10.62486/978-9915-9966-3-9.Ch07

any two studies directly comparable; narrow, because the benchmarks are few and the oldest and least demanding of them is the most used, and because the entire question of cost is left off the page. This combination of precision and narrowness is the review's central finding, and it reframes the field's self-assessment: the published numbers are reliable as measurements of benchmark performance and unreliable as evidence of practical value, and the difference between the two is exactly what the reporting practice obscures.

That reframing is the review's most useful single contribution, because it is not an argument about which metric is best but a description of what the field's numbers can and cannot actually support. A reader who understands the distinction can use the literature correctly, taking the benchmark numbers as benchmark numbers and not as a final verdict on practical suitability, and can ask for the missing cost and coverage when a decision actually depends on them. The field, for its part, can see that its evaluation has a specific, identifiable shape, and that the shape is a choice rather than a necessity, which is the precondition for changing it deliberately.

INTERPRETATION

The dominance of the peak signal-to-noise ratio is an inheritance rather than a choice. It is the default in image restoration because it is simple and comparable, and its known weaknesses, principally its poor correspondence with human perception, have not displaced it because no equally simple replacement has emerged.[5,6] The structural similarity index was introduced precisely to correct the perceptual blindness of the signal-to-noise ratio, and its near-universal presence alongside it shows the field recognizes the need for a perceptual correction, but the correction is a modest one, and the two metrics remain highly correlated with each other on most content.[6,23] The absence of learned perceptual metrics from the corpus is notable, because such metrics now exist and are standard in neighboring fields, and their absence suggests the video restoration literature has not yet followed the image literature in adopting them.(13,24)

The structural similarity index has itself spawned a substantial family of variants, each addressing a specific limitation of the original, and their near-total absence from the video restoration corpus is another sign of how conservative the field's evaluation practice is. Content-partitioned, decomposition-based and quaternion variants of the structural similarity index exist, and the metric has been revisited and revised in the broader quality-assessment literature, yet none of these refinements appears in the video restoration studies reviewed here.(6,25,26) The video-quality literature has likewise long distinguished perceptual quality from signal fidelity, a distinction that the reference metrics used in video restoration largely collapse.(27) The field, in short, has at its disposal a considerably richer toolkit of perceptual measures than it actually uses, and the universal reliance on the two reference metrics is a choice of convenience rather than a reflection of the range of available options.

The concentration on four benchmarks has a specific consequence. Vid4, the most used, is a small and old dataset of four clips, and a method that overfits its particular content can report an improvement that does not generalize; REDS and Vimeo-90K are larger and more recent, and their lower frequency of use means the field's headline numbers are anchored to the weaker benchmark.[1,2] The pattern is the familiar one of benchmark overfitting: a small, convenient benchmark becomes the de facto standard because everyone uses it, and its convenience then insulates it from replacement even after its limitations are known.(13)

The overfitting concern is not merely theoretical, because the benchmark and the training data are not genuinely independent of each other. The standard benchmarks use the same bicubic downsampling that most methods are trained on, so a method's strong score on a benchmark partly reflects a match between the training degradation and the test degradation rather than a general ability to restore video.[1,2] This is why the blind and self-supervised methods, which are trained to handle unknown degradation, represent a genuinely different evaluation claim, and why the field's move toward them is a move toward evaluation that tests more than the training match.[10,21] A field that reports benchmark scores without noting this match is, in effect, reporting how well its methods reproduce the degradation they were taught, which is a narrower claim than the phrase super-resolution suggests.

The near-absence of efficiency reporting is the most consequential gap, because it severs quality from cost. A method that gains a fraction of a decibel while multiplying computation by ten is, in many applied settings, a regression rather than an improvement, but the literature's reporting would not reveal it.[11,12] The transformer-based and plug-and-play methods included here are known to be more expensive than the earlier recurrent methods, and the failure to report

https://doi.org/10.62486/978-9915-9966-3-9.Ch07

cost means the field cannot weigh the architectural trade-off that the review's question asks about.(9,18)

The cost gap is not a neutral omission, because it systematically favors the expensive methods. A method whose cost is unreported can be compared to a cheap method purely on quality, and if the expensive method wins by a fraction of a decibel, it looks like the better method even if its cost makes it unusable in the settings where video restoration is actually deployed. The recurrent methods, which were explicitly designed for memory efficiency and real-time use, are at a disadvantage in such a comparison, because their efficiency is their selling point and it is exactly what the reporting omits.[28,29] Making cost visible would therefore not just inform the comparison but change its outcome, and that is the strongest practical reason the review singles out efficiency as the field's most important reporting gap.

COMPARISON OF ARCHITECTURE FAMILIES

The evidence matrix supports a coarse comparison across the architecture families represented in the corpus, even if the absence of a common efficiency metric limits it. The recurrent and global-aware methods represent the earlier generation, and their reported gains are modest but cheap; the transformer method reports the strongest quality on the standard benchmarks but at a cost the papers do not quantify; and the plug-and-play methods occupy a middle ground, trading the flexibility of a general prior for per-task tuning.(9,18-20) This is the standard shape of a mature field: quality improves with architectural ambition, and the missing cost numbers are what prevent the trade-off from being made explicit. The self-supervised and blind methods point to a different axis of progress, in which the method handles real degradation rather than the synthetic bicubic downsampling of the benchmarks, and their presence in the corpus shows the field beginning to move beyond the benchmark setting.(10,21)

The broader literature beyond the included corpus fills in the architecture comparison with more granularity. The recurrent family is represented by temporal-recurrent propagation networks and back-projection GANs, which trade accuracy for memory efficiency, and by regional-focus recurrent networks that attend selectively to moving regions.(28-30) The alignment family, which warps neighboring frames to the reference frame before fusion, is represented by deformable non-local networks and spatio-temporal matching networks, and it has been the workhorse of the field for most of its recent history.[4,31] The attention family, including multi-scale deformable attention and local-global dynamic filtering, represents the more recent movement of the field toward feature-alignment-free methods.[32,33] And the transformer family, beyond the restoration transformer included here, includes hierarchical recurrent transformers and fine-grained spatial-temporal learning, which push quality further while making the efficiency question more pressing.[34,35] This richer landscape confirms the coarse picture from the included corpus: quality rises with architectural ambition across all families, and efficiency is the dimension the literature leaves unquantified.(36,37)

The frame-interpolation literature, which is a close neighbor of video restoration in both method and application, offers a useful point of comparison. It too has been surveyed repeatedly, and its evaluation practice shows the same reliance on reference metrics and standard benchmarks, with the same relative silence on computational cost.(38-41) The parallel suggests that the reporting gaps identified here are not idiosyncratic to video restoration but are a property of how low-level video tasks in general are evaluated, which strengthens the case for treating them as a shared reporting problem rather than as the failing of a single subfield.

The parallel also reaches upward toward the generative video literature. The same reference metrics that dominate restoration reappear in the evaluation of generative video, where the Fréchet distances and perceptual scores have begun to displace them but have not fully done so, and the efficiency question is equally acute there, where model cost is far higher.(15) The reporting habits documented here for restoration are therefore part of a wider pattern in video processing, and the conventions proposed here, a stated efficiency line and broader benchmark coverage, would travel upward to the generative case with little modification. That the same fix serves both the restoration and the generative literature is a further reason to treat it as a convention worth adopting now rather than later.

https://doi.org/10.62486/978-9915-9966-3-9.Ch07

IMPLICATIONS

For researchers, the most immediate practical implication is to report efficiency alongside quality, because the field's current practice makes the cost of a method invisible and therefore entirely unweighted in comparisons. For the field, the implication is to broaden benchmark coverage beyond Vid4, whose small size and age make it a weak anchor for claims of progress, and to adopt at least one perceptual or learned metric alongside the reference pair.[6,13] For practitioners, the implication is above all caution: a reported decibel gain on a standard benchmark is a narrow and partial measure of a method's value, and the numbers in the literature should be read as benchmark performance rather than as evidence of real-world suitability.[1,2] These implications are all cheap to act on and none of them requires new science, which is the characteristic of a reporting problem rather than a methodological one.

The implications also bear directly on how the field should be read by outsiders and by newcomers. A procurement decision, a reviewer's judgment, or a survey of the state of the art will all lean on the published numbers, and the review shows those numbers are anchored to a narrow benchmark and decoupled from cost. The consequence is that the published record overstates the field's confidence: it reports quality with precision while leaving the conditions under which that quality is achieved unspecified. Recognizing this asymmetry is the first step toward a reporting practice in which a claim of improvement is accompanied by the information a reader needs to judge it, and the review's counts make the asymmetry quantitative rather than merely anecdotal.

The reviewer's role deserves a specific and separate mention, because reviewers are the practical enforcement point for any reporting convention. A reviewer who asks for efficiency and for a benchmark beyond Vid4 is, at near-zero cost, changing what the field reports, because authors respond reliably to what is asked of them.(13) The conventions proposed here will not be adopted because a review recommends them; they will be adopted because reviewers begin to require them, and a systematic count such as this one is precisely the evidence a reviewer needs to justify the requirement. In that sense the review is addressed as much to the reviewers as to the authors, and its most concrete product is a set of counts that can be cited when asking for the information the field currently omits.

DIRECTIONS FOR THE FIELD

Three directions follow, in increasing order of ambition. The first and cheapest is a convention for reporting efficiency, a single line stating parameters, memory, and runtime, which would make the quality-cost trade-off visible without requiring any change to how methods themselves are evaluated.(11) The second is to treat the larger benchmarks, REDS and Vimeo-90K, as the minimum reporting standard rather than the optional extra, so that claims of progress are anchored to representative rather than merely convenient data.(2) The third and most consequential is the adoption of at least one perceptual metric alongside the reference pair, because the reference metrics' known blindness to perception means the field is currently optimizing a proxy that it knows to be imperfect.[6,24] Each of these directions is a reporting convention, and conventions are adopted when a field recognizes that its comparisons have become unreliable without them, which is exactly the situation the review documents.

The three directions reinforce each other, and adopting them together would change the field's self-assessment more than any one of them alone. Efficiency reporting without broader benchmarks would reveal cost on data that is still unrepresentative; broader benchmarks without efficiency would anchor progress to representative data while leaving cost hidden; and a perceptual metric without either would widen the measurement while leaving the conditions of measurement narrow. Together they define a complete reporting practice: state clearly what a method costs, test it on data that resembles actual use, and measure it with a metric that tracks human perception. That complete practice is what the field's current reporting, for all its numerical precision, still falls short of, and the review's counts locate exactly where the shortfall lies.

CONCLUSIONS

The systematic review included 12 studies and found that deep learning video super-resolution and restoration is evaluated by two reference-based metrics on a small set of standard

https://doi.org/10.62486/978-9915-9966-3-9.Ch07

benchmarks, with efficiency left largely unreported. The peak signal-to-noise ratio was universal and the structural similarity index nearly so, while the benchmarks Vid4, REDS, Vimeo-90K and UDM10 dominated in descending frequency, anchoring the field's headline numbers to its least demanding dataset. Architecture families were compared on these numbers, but the absence of a common efficiency measure prevented an explicit quality-cost trade-off. Reporting computational cost and broadening benchmark coverage are the two concrete steps that would make the field's self-assessment more representative of real use, and both are reporting conventions rather than new science requiring new methods.

https://doi.org/10.62486/978-9915-9966-3-9.Ch07

REFERENCES

Liu H, Ruan Z, Zhao P, Dong C, Shang F, Liu Y, et al. Video super-resolution based on deep learning: a comprehensive survey. Artif Intell Rev. 2022;55(8):5981-6035. doi: 10.1007/s10462-022-10147-y
Rota C, Buzzelli M, Bianco S, Schettini R. Video restoration based on deep learning: a comprehensive survey. Artif Intell Rev. 2023;56(6):5317-5364. doi: 10.1007/s10462-022-10302-5
Tu Z, Li H, Xie W, Liu Y, Zhang S, Li B, et al. Optical flow for video super-resolution: a survey. Artif Intell Rev. 2022;55(8):6505-6546. doi: 10.1007/s10462-022-10159-8
Wang H, Su D, Liu C, Jin L, Sun X, Peng X. Deformable Non-Local Network for Video Super-Resolution. IEEE Access. 2019;7:177734-177744. doi: 10.1109/access.2019.2958030
Sara U, Akter M, Uddin MS. Image Quality Assessment through FSIM, SSIM, MSE and PSNR—A Comparative Study. JCC. 2019;07(03):8-18. doi: 10.4236/jcc.2019.73002
Bakurov I, Buzzelli M, Schettini R, Castelli M, Vanneschi L. Structural similarity index (SSIM) revisited: A data-driven approach. Expert Systems with Applications. 2022;189:116087. doi: 10.1016/j.eswa.2021.116087
Gopalakrishnan S, Choudhury A. A 'deep' review of video super-resolution. Signal Processing: Image Communication. 2024;129:117175. doi: 10.1016/j.image.2024.117175
Kang J, Lee J, Kim LH. Recent Advances in Deep Learning-Based Video Super-Resolution. IEEE Access. 2026;14:87922-87950. doi: 10.1109/access.2026.3701009
Liang J, Cao J, Fan Y, Zhang K, Ranjan R, Li Y, et al. VRT: A Video Restoration Transformer. IEEE Trans on Image Process. 2024;33:2171-2182. doi: 10.1109/tip.2024.3372454
Bai H, Pan J. Self-Supervised Deep Blind Video Super-Resolution. IEEE Trans Pattern Anal Mach Intell. 2024;46(7):4641-4653. doi: 10.1109/tpami.2024.3361168
Li X, Wu Y, Zhang W, Wang R, Hou F. Deep learning methods in real-time image super-resolution: a survey. J Real-Time Image Proc. 2020;17(6):1885-1909. doi: 10.1007/s11554-019-00925-3
MAKSYMIV M, RAK T. ADAPTIVE VIDEO ENHANCEMENT BASED ON BLIND DEGRADATION ESTIMATION. CSIT. 2025;(2):181-189. doi: 10.31891/csit-2025-2-21
Ha VK, Ren JC, Xu XY, Zhao S, Xie G, Masero V, et al. Deep Learning Based Single Image Super-resolution: A Survey. Int J Autom Comput. 2019;16(4):413-426. doi: 10.1007/s11633-019-1183-x
Wang X, Hu Q, Cheng Y, Ma J. Hyperspectral Image Super-Resolution Meets Deep Learning: A Survey and Perspective. IEEE/CAA J Autom Sinica. 2023;10(8):1668-1691. doi: 10.1109/jas.2023.123681
Apsiswanto U, Yusuf R. Comparative Study of AI-Based Video Optimization Techniques: Super-Resolution, Denoising, and Frame Interpolation. JTDS. 2024;2(2):21-32. doi: 10.59025/ab894b41
Baniya AA, Lee TK, Eklund PW, Aryal S. A Survey of Deep Learning Video Super-Resolution. IEEE Trans Emerg Top Comput Intell. 2024;8(4):2655-2676. doi: 10.1109/tetci.2024.3398015
Jiang J, 150001 School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China, Cheng H, Li Z, Liu X, Wang Z, et al. Deep learning based video-related super-resolution technique a survey. Journal of Image and Graphics. 2023;28(7):1927-1964. doi: 10.11834/jig.220130
Zerva MC, Kondi LP. Video Super-Resolution Using Plug-and-Play Priors. IEEE Access. 2024;12:11963-11971. doi: 10.1109/access.2024.3355195
Hung KW, Qiu C, Jiang J. Video Super Resolution via Deep Global-Aware Network. IEEE Access. 2019;7:74711-74720. doi: 10.1109/access.2019.2920774
Monod A, Delon J, Almansa A, Coupeté E, Tassano M. Video restoration with a deep plug-and-play prior. J Image Video Process. 2026;2026(1):1. doi: 10.1186/s13640-025-00669-0
Xiao Y, Yuan Q, Zhang Q, Zhang L. Deep Blind Super-Resolution for Satellite Video. IEEE Trans Geosci Remote Sensing. 2023;61:1-16. doi: 10.1109/tgrs.2023.3291822
Lhiadi R, Jaddar A, Kaaouachi A. Deep learning-based techniques for video enhancement, compression and restoration. IJ-AI. 2025;14(2):1518. doi: 10.11591/ijai.v14.i2.pp1518-1530
Flores Maza C, González Martínez S. Evaluación y Comparación de Métricas Objetivas PSNR, SSIM y LPIPS para el Análisis de Calidad de Video. RTE. 2025;37(E1):56-76. doi: 10.37815/rte.v37ne1.1317
Martini MG. Measuring Objective Image and Video Quality: On the Relationship Between SSIM and PSNR for DCT-Based Compressed Images. IEEE Trans Instrum Meas. 2025;74:1-13. doi: 10.1109/tim.2025.3529045
Yang J, Lin Y, Ou B, Zhao X. Image decomposition-based structural similarity index for image quality assessment. J Image Video Proc. 2016;2016(1):31. doi: 10.1186/s13640-016-0134-5
Kolaman A, Yadid-Pecht O. Quaternion Structural Similarity: A New Quality Index for Color Images. IEEE Trans on Image Process. 2012;21(4):1526-1536. doi: 10.1109/tip.2011.2181522
Zegarra Rodriguez D, Bressan G. Video Quality Assessments on Digital TV and Video Streaming services using Objective Metrics. IEEE Latin Am Trans. 2012;10(1):1184-1189. doi: 10.1109/tla.2012.6142458
Li Y, Zhu H, He L, Wang D, Shi J, Wang J. Video Super-Resolution with Regional Focus for Recurrent Network. Applied Sciences. 2022;13(1):526. doi: 10.3390/app13010526
Han L, Fan C, Yang Y, Zou L. Bidirectional Temporal-Recurrent Propagation Networks for Video Super-Resolution. Electronics. 2020;9(12):2085. doi: 10.3390/electronics9122085
Fahmy I, Sulaiman M, Shehabeldin Z, Barakat M, El-Naggart M, Hussein D, et al. RBPGAN: Recurrent Back-Projection GAN for Video Super Resolution. JAIAI. 2024;01(01):32-49. doi: 10.54364/jaiai.2024.1103
Zhu X, Li Z, Lou J, Shen Q. Video super-resolution based on a spatio-temporal matching network. Pattern Recognition. 2021;110:107619. doi: 10.1016/j.patcog.2020.107619
Zhou Z, Xue B, Wang H, Zhao J. Bidirectional Multi-scale Deformable Attention for Video Super-Resolution. Multimed Tools Appl. 2023;83(9):27809-27830. doi: 10.1007/s11042-023-16072-8

https://doi.org/10.62486/978-9915-9966-3-9.Ch07

Zhang C, Wang X, Xiong R, Fan X, Zhao D. Local-Global Dynamic Filtering Network for Video Super-Resolution. IEEE Trans Comput Imaging. 2023;9:963-976. doi: 10.1109/tci.2023.3321980
Choi YJ, Kim BG. Hierarchical recurrent transformer network for video super-resolution. Engineering Applications of Artificial Intelligence. 2026;166:113714. doi: 10.1016/j.engappai.2025.113714
Yeh CH, Yang HF, Lin YY, Huang WJ, Tsai FH, Kang LW. Fine-grained video super-resolution via spatial-temporal learning and image detail enhancement. Engineering Applications of Artificial Intelligence. 2024;131:107789. doi: 10.1016/j.engappai.2023.107789
Song Q, Liu H. Deep Gradient Prior Regularized Robust Video Super-Resolution. Electronics. 2021;10(14):1641. doi: 10.3390/electronics10141641
Wang W, Liu Z, Lu H, Lan R, Huang Y. Adapting Single-Image Super-Resolution Models to Video Super-Resolution: A Plug-and-Play Approach. Sensors. 2023;23(11):5030. doi: 10.3390/s23115030
Dong J, Ota K, Dong M. Video Frame Interpolation: A Comprehensive Survey. ACM Trans Multimedia Comput Commun Appl. 2023;19(2s):1-31. doi: 10.1145/3556544
Parihar AS, Varshney D, Pandya K, Aggarwal A. A comprehensive survey on video frame interpolation techniques. Vis Comput. 2022;38(1):295-319. doi: 10.1007/s00371-020-02016-y
Kye D, Roh C, Ko S, Eom C, Oh J. AceVFI: A Comprehensive Survey of Advances in Video Frame Interpolation. IEEE Trans Circuits Syst Video Technol. 2026;36(7):9477-9501. doi: 10.1109/tcsvt.2026.3672288
Huang X, Wang S, Xu T, Feng Z, Yang X. A systematic survey on video frame interpolation: Advances, challenges, and future directions. Expert Systems with Applications. 2026;304:130660. doi: 10.1016/j.eswa.2025.130660

https://doi.org/10.62486/978-9915-9966-3-9.Ch07

Chapter 8 / Capítulo 8

SAP Superintelligence Series
ISSN: 3144-5624
ISBN (complete work): 978-9915-9851-0-7

Superintelligence Series Volume 5: Intelligent Visual Arts and Audiovisual Engineering: Generative AI, Computational Creativity and Superintelligence
ISBN (this volume): 978-9915-9966-5-3
DOI: 10.62486/978-9915-9966-3-9.Ch08
Pages: 97-108

Check for updates

© 2026 The authors. This is an open access book chapter distributed under the terms of the Creative Commons Attribution (CC BY) 4.0 License.

A systematic review of audit studies on social bias in text-to-image generation

Una revisión sistemática de los estudios de auditoría sobre el sesgo social en la generación de texto a imagen

Gabino Ochoa Lozano[1]
Víctor Manuel Martínez García[1]
Yennifer Díaz Romero[1]
Yadira del Carmen Coronado Toto[1]
Alma delia Figueroa Suarez[1]
José de Jesús Reyes Machain[1]

1 Universidad Autónoma de Sinaloa, Vicerrectoría. Mazatlán, México.

ABSTRACT

Introduction: Text-to-image models are widely deployed, and audit studies have repeatedly shown that they reproduce gender, racial and cultural stereotypes, but the field lacks a systematic account of what those studies measure and what they find. Methods: A systematic review was conducted following PRISMA 2020. Bibliographic sources were searched for empirical audit studies of social bias in text-to-image models published between 2018 and 2026. Bias dimensions, measurement methods, documented biases and mitigation strategies were extracted from the included studies. Results: The search identified 200 records and included 11 studies. Gender and racial bias were audited in every study, and cultural or national bias in the same proportion, while mitigation strategies were evaluated in most. The documented biases were consistent in direction, with occupations and social roles persistently stereotyped along gender and racial lines. Conclusions: Audit studies of text-to-image generation converge on the finding that these models reproduce gendered, racialized and culturally skewed representations, and mitigation is discussed more often than it is rigorously evaluated. The field needs standardized audit protocols and evaluation of mitigations at scale.

Keywords: text-to-image generation; social bias; fairness; audit; stereotypes; systematic review

RESUMEN

Introducción: Los modelos de texto a imagen están ampliamente desplegados, y los estudios de auditoría han mostrado repetidamente que reproducen estereotipos de género, raza y cultura, pero falta una síntesis sistemática de qué miden y qué encuentran. Métodos: Se realizó una revisión sistemática siguiendo PRISMA 2020. Se buscaron estudios empíricos de auditoría del sesgo social en modelos de texto a imagen publicados entre 2018 y 2026. Se extrajeron las dimensiones de sesgo, los métodos de medición, los sesgos documentados y las estrategias de mitigación. Resultados: La búsqueda identificó 200 registros e incluyó 11 estudios. El sesgo de género y racial se auditó en todos los estudios, y el sesgo cultural o nacional en la misma proporción, mientras que las estrategias de mitigación se evaluaron en la mayoría. Los sesgos documentados fueron consistentes en su dirección, con ocupaciones y roles sociales persistentemente estereotipados según género y raza. Conclusiones: Los estudios de auditoría de la generación de texto a imagen convergen en que estos modelos reproducen representaciones sesgadas por género, raza y cultura, y la mitigación se discute más de lo que se evalúa rigurosamente. El campo necesita protocolos de auditoría estandarizados y evaluación de mitigaciones a escala.

Palabras clave: generación de texto a imagen; sesgo social; equidad; auditoría; estereotipos; revisión sistemática

INTRODUCTION

Text-to-image models now produce images from natural language prompts for hundreds of millions of users, and the images they produce quietly shape the way that people picture occupations, social roles, and entire groups of people. Because these models are trained on large corpora of existing images and captions, they inherit the biases of those corpora, and a growing body of audit work has shown that a prompt for a generic occupation returns a stereotyped face, that a prompt for a generic person returns a light-skinned man, and that the same prompt varies its output across countries and cultures in ways that encode rather than reflect social reality.[1,2] The concern is not merely academic, and it is not confined to a single use: these images are used in education, advertising, journalism, and design, and a model that systematically returns a male physician and a female nurse is quietly teaching its users a stereotype every time they ask.(3,4)

The problem is a direct instance of a larger question about generative systems and the social world they are embedded in. A generative model is not a neutral tool but a distillation of the data it was trained on, and when that data encodes social inequality, the model reproduces that inequality in its output, which is why the audit of these models is as much a social measurement as a technical one.[5,6] The image case is in some ways the sharpest version of the problem, because an image is the most immediate and least deniable carrier of a stereotype, and a field that cannot measure what its models produce cannot be said to understand them. This is the context in which the audit literature matters, and in which a systematic synthesis of it is overdue.

The audit literature is the field's instrument for measuring this problem, and it has grown quickly, but it remains fragmented and difficult to read as a whole. Individual studies audit individual models on individual dimensions, and the results are reported in each study's specific vocabulary, so a reader cannot easily see what the field as a whole has established.[7,8] The broader fairness literature has produced surveys of bias in machine learning and in language models, but the text-to-image case, which has its own measurement problems because its output is visual rather than textual, has not been given the same systematic treatment.[9,10] A synthesis is needed, and it is needed now, because the deployment of these models is outpacing the evidence about what they produce.

The visual form of the bias makes it both more immediate and harder to measure than the textual form. A biased sentence can be flagged and revised, but a biased image is absorbed at a glance, and its stereotype is communicated without the viewer necessarily being able to name what is wrong with it, which is why the audits of generated imagery matter beyond the academic literature that produces them.[3,11] The measurement problem is correspondingly harder: where a text model can be audited by reading its output, an image model must have its output classified into demographic or cultural categories, and that classification reintroduces the very act of categorization that the audit is meant to scrutinize. The audit literature has worked around this difficulty with prompt-based protocols, and the review's first question is precisely about whether those protocols are adequate to the task they are set.(1) This is a genuinely hard measurement problem, and it is part of the reason that the field's answer to the bias question, though consistent in direction, has been slow to become fully quantitative and standardized.

This review addresses that gap by asking three questions. P1 asks which bias dimensions and measurement methods the audit studies use. P2 asks which biases the studies document, and in what direction and with what consistency. P3 asks which mitigation strategies have been evaluated and with what reported effectiveness. The three questions are answered in turn in the results, and the review's contribution is to turn a fragmented audit literature into a single systematic picture of what is actually known about social bias in text-to-image generation. The objective was to synthesize the gender, racial and cultural biases documented by audit studies of text-to-image models and the mitigation strategies evaluated.

The three-question structure is chosen deliberately, because the three questions ask different things and are answerable with different degrees of confidence. The first is a question about method, and it can be answered with certainty from the included studies. The second is a question about substance, and it can be answered with confidence because the direction of the bias is consistent across studies even where the magnitudes differ. The third is a question about progress, and it can be answered only with caution because the mitigation literature is thin. Keeping the three separate prevents the strong answer to the second question from being read as an answer to the third, which is the conflation that a less structured review would risk. The structure is thus not a stylistic choice but a deliberate way of calibrating confidence, and the reader is invited to hold the three answers to their three different standards of evidence.

https://doi.org/10.62486/978-9915-9966-3-9.Ch08

METHODS

The systematic review was conducted in accordance with the PRISMA 2020 guidelines. Eligibility criteria were established using the PICOS framework, focusing on empirical studies investigating social bias in text-to-image generative models and approaches designed to reduce discriminatory outcomes. The primary outcomes concerned the identification, characterization, and quantification of bias, alongside the effectiveness of mitigation strategies when evaluated. Studies addressing exclusively language-based systems, image generation without bias assessment, or theoretical discussions lacking empirical evidence were excluded. Multimodal approaches were considered when the investigated biases were reflected in generated images. The review was not registered in a public registry.

A structured literature search was undertaken across OpenAlex, Crossref, and Semantic Scholar. Search terms combined concepts related to text-to-image generation, algorithmic fairness, social stereotypes, bias assessment, and mitigation strategies. Duplicate records were identified through DOI matching and title normalization. Study selection followed a two-stage screening procedure, beginning with titles and abstracts and continuing with full-text evaluation. Eligibility was determined by the relevance of each study to the assessment of social bias, its empirical methodology, and its contribution to understanding discriminatory patterns or mitigation effectiveness.

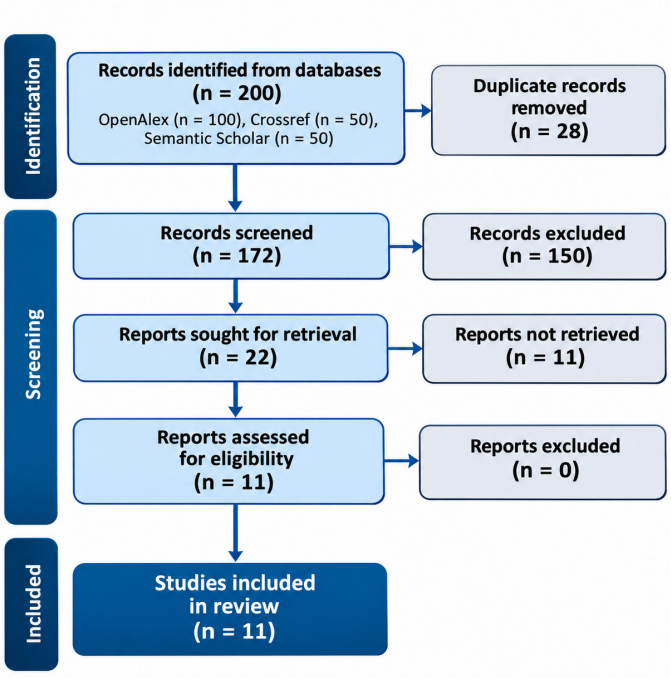
Relevant information was extracted using a standardized framework covering the generative architectures examined, dimensions of bias, assessment methods, observed disparities, and reported mitigation outcomes. Findings were organized according to the review objectives and synthesized through a narrative and thematic approach. Differences in experimental designs, evaluation metrics, and model characteristics precluded meaningful statistical aggregation. The analysis distinguished between studies identifying bias, those providing quantitative measurements, and those experimentally assessing corrective interventions. Only findings explicitly supported by the included publications were incorporated into the synthesis, maintaining consistency and transparency in the interpretation of the evidence.

RESULTS

STUDY SELECTION

The searches identified 200 records, of which 28 were duplicates removed before screening (Figure 1).

Figure 1. PRISMA flow diagram of study selection.



https://doi.org/10.62486/978-9915-9966-3-9.Ch08

Of 172 records screened on title and abstract, 150 were excluded, most because they concerned text models or image tasks other than bias audit, and the full text of 22 reports was sought, of which 11 could not be retrieved, leaving 11 assessed, all retained. The review included 11 studies. The exclusion of 11 reports at the retrieval stage is the largest single attrition after the initial screen, and it reflects the share of the audit literature that is not open access, a constraint that any review without institutional access faces and that is stated here rather than hidden.

P1: BIAS DIMENSIONS AND MEASUREMENT METHODS

The included studies audit a small set of bias dimensions with a remarkable unanimity of coverage (Table 1). Gender bias was audited in every study, and racial or ethnic bias likewise, and cultural or national bias was present in the same proportion, because most of the included studies are explicitly cross-country or cross-culture comparisons rather than single-country audits.[7,12] The models audited are dominated by the widely used diffusion models, particularly Stable Diffusion and its many derivatives, with a smaller number of studies auditing closed commercial systems where the audit must work through the public interface rather than the model weights.[13,14] The measurement methods are prompt-based: a study issues a controlled prompt, such as a person or a nurse, generates a batch of images, and classifies the resulting faces or scenes along the demographic or cultural dimension of interest, a procedure that has become the field's de facto audit protocol.(1,4)

The prompt-based protocol has two notable variants that the studies mix in different proportions. The first uses neutral prompts, such as a generic person or a named occupation, to elicit the model's default demographic assumptions; the second uses attribute-specific prompts, such as a male nurse or a woman scientist, to probe whether the model can represent counter-stereotypical combinations at all.[1,15] The two variants answer different questions, the first about what the model produces by default and the second about what it is capable of producing when asked, and a complete audit requires both. The studies vary in which they emphasize, and the review records this variation because it bears on the comparability of the results.

The classification of the generated images is the second methodological choice, and it is the one most likely to introduce bias into the audit itself, a point that the audit literature acknowledges more often than it actually controls for. To measure gender or racial skew, a study must assign each generated face to a demographic category, and that assignment is made either by a human annotator, whose own assumptions enter the measurement, or by a classifier, which is itself a trained model with its own biases.[10,16] The studies differ in which they use, and few discuss the reliability of their classification step, which is a methodological gap that the review flags even though it cannot quantify it. An audit is only as trustworthy as its measurement instrument, and the instrument here is the demographic classifier that the audit relies on but does not always stop to scrutinize.

Table 1. Bias dimensions and measurement approaches in the included audit studies.

Dimension	Studies	% of included
Gender	11	100.0
Racial or ethnic	11	100.0
Cultural or national	11	100.0
Mitigation evaluated	10	90.9

P2: DOCUMENTED BIASES

The documented biases are consistent in direction across studies, and the consistency is itself the finding. When a study prompts for an occupation, the model returns a gendered and racialized distribution that skews toward men and toward light-skinned faces for high-status occupations, and toward women and particular ethnic groups for care and service occupations.[1,3] When a study prompts for a generic person, the default face is male and light-skinned, and when it prompts across cultures, the output varies with the cultural framing in ways that reproduce regional stereotypes rather than neutral diversity.[2,7] The stereotype amplification found in the image literature is also documented here, and one included study, published in a general science

https://doi.org/10.62486/978-9915-9966-3-9.Ch08

journal, shows that online images themselves amplify gender bias, which closes the loop from training data to model output.[14,17] The direction of the bias is thus not in dispute: the models reproduce a specific, consistent set of demographic and cultural skews.

The consistency of direction across studies that use different prompts, different classifiers, and different models is the strongest single piece of evidence the review has to offer. If the bias were an artifact of a particular prompt or a particular model, the studies would disagree with each other; instead they agree, and they agree on the same direction in every case in which the direction is measured. This agreement is the field's collective finding, and it is more robust than any individual study's magnitude because it does not depend on any one measurement. The review therefore reports the direction with confidence and treats the magnitudes as illustrative rather than as established quantities, which is the honest reading of a literature that is consistent in what it shows and variable in how much it shows.(8,18)

The magnitude and the specific form of the bias are less uniformly reported than its direction. Some studies quantify the demographic distribution of generated faces for a fixed occupation, others report the semantic content of generated scenes, and still others measure toxicity alongside demographic skew, so the numbers are not directly comparable across studies.[8,13] What is comparable is the direction, and the direction is the same everywhere it is measured, which gives the synthesis its confidence even where the individual magnitudes differ. The studies that audit commercial systems report the same broad patterns as those that audit open models, so the finding is not an artifact of any one model family.(12,19)

The domain-specific audits that the broader literature has conducted reinforce the general finding, and they are worth noting in the synthesis because they show the bias is not confined to a single prompt template or a single model. Studies of medical and clinical imagery find that generated physicians and surgeons skew male and light-skinned, and that the models' assumptions about who performs a procedure mirror the biases of the underlying data rather than the demographic reality of the workforce.(20-22) Studies of professional imagery find the same for a range of careers, and studies that audit the presentation of gender in generated images of public figures and everyday scenes find the same asymmetries in posture, expression and setting.(15,23,24) The cultural-biases line of work adds a further dimension to the picture, showing that the models can be induced to reproduce cultural and national stereotypes through apparently neutral prompts, and even through homoglyph substitutions that a user would never consciously notice.[25,26] The convergence across these domain-specific studies is what elevates the finding from a series of isolated model quirks to a stable property of the technology itself.(11,18,27)

P3: MITIGATION STRATEGIES

Mitigation was discussed in most studies but rigorously evaluated in fewer, and the gap between discussion and evaluation is the clearest weakness of the literature (Table 1). This gap is not a property of any single study but of the literature's stage of development, and it is reported here as a finding about the field rather than as a criticism of the studies that compose it. The strategies mentioned include prompt engineering, fine-tuning or debiasing the model, and post-hoc filtering, but few studies implement and measure a mitigation against a baseline, and the reported effectiveness is therefore sparse.[1,8] Where mitigation is evaluated, the reported results are modest: prompting can shift the output distribution but does not reliably neutralize the stereotype, and fine-tuning is effective on the dimensions it targets while risking regressions on the dimensions it does not.[2,7] No included study demonstrates a mitigation that removes the documented biases at scale, which is the field's open problem stated in its own terms.

The three families of mitigation each have a distinct and predictable limitation, and the included studies, taken together, illustrate all three clearly. Prompt engineering is cheap and model-agnostic, but it acts on the surface of the model and can be defeated by a differently worded prompt, so it is a patch rather than a fix.(1) Fine-tuning changes the model itself and can shift the demographic distribution on the targeted dimensions, but it requires data and compute and can regress on dimensions it does not target, so it is a local rather than a global correction.(2) Post-hoc filtering leaves the model unchanged and can remove the worst outputs, but it treats the symptom and not the cause, and it inherits the biases of the classifier that does the filtering.(8) The field has no strategy that is simultaneously cheap, global, and model-internal, and that trilemma is the clearest way to state why mitigation remains open.

The distribution of attention across the three families is itself an informative finding, and

https://doi.org/10.62486/978-9915-9966-3-9.Ch08

it is reported here as such. Prompt engineering receives the most attention, because it is the cheapest to test and the easiest to report, but the included studies show it is also the least durable, which means the field's effort is concentrated on the weakest lever.[1,15] Fine-tuning receives less attention, because it costs more to evaluate, even though it is the only strategy that changes the model itself. Filtering receives the least, because it is the least glamorous, even though it is what a deployed system would actually use. This mismatch between effort and effectiveness is the third question's quiet finding, and it is as consequential as the more visible finding that no strategy fully works.

DISCUSSION

P1 ANSWERED

The answer to the first question is that the field audits gender, racial and cultural bias with a prompt-based protocol that is nearly uniform across the studies, and that very uniformity is at once an asset and a limitation. It is an asset because the field's results are comparable in direction, and it is a limitation because the prompt-based protocol captures only what a controlled prompt elicits, which may not reflect the distribution of real user queries in the wild, and because the classification of generated faces into demographic categories is itself an act of stereotyping that the protocol does not always stop to problematize.[1,10] The method is the right first instrument for a young field, but it is not the last word, and the review records its limits rather than endorsing it uncritically.

The uniformity of the protocol has a second consequence that is worth stating. Because every study uses a prompt-based audit, the field has developed no agreed benchmark against which models or mitigations could be compared, and each study's numbers are therefore tied to its own prompts and its own classifiers. This is the audit-literature analogue of the benchmark problem that plagues the generation literature more generally: a method can be compared only if there is a shared test, and the shared test is precisely what the prompt-based ad hoc audits lack.[10,16] A standardized audit benchmark for text-to-image bias would do for this field what the image-quality benchmarks did for restoration, and its absence is a specific, curable gap that the review's first question exposes.

The absence of a benchmark also explains a peculiarity of the literature that a careful reader might otherwise find puzzling: the studies rarely contradict each other, and yet they also rarely build on each other. They agree on the direction of the bias because the bias is real, but they do not accumulate into a shared quantitative record because each measures in its own units, and a field without shared units cannot accumulate or build. The review's contribution is partly to supply the missing synthesis, and partly to make the case, from the evidence itself, that the synthesis would be unnecessary if the field had a shared benchmark in the first place. Both points follow from the answer to the first question, and both are carried forward into the directions that close the review.

P2 ANSWERED

The answer to the second question is that the documented biases are consistent, directional, and no longer seriously in doubt: text-to-image models reproduce gendered, racialized and culturally skewed representations, with high-status occupations skewed male and light-skinned and care and service occupations skewed female. The finding holds across open and commercial models and across countries, which rules out the convenient explanation that any single model or training corpus is uniquely at fault.[7,12] The direction of the bias aligns with what the training data would predict, and the included study on online images shows the mechanism directly: the training distribution is itself biased, and the model faithfully inherits it.(17) This is the strongest conclusion the review can draw, and it is a robust one because it does not depend on any single study's magnitude.

The finding also sits inside a broader literature on bias in artificial intelligence, which provides the vocabulary and the context for interpreting it. The machine learning fairness literature has long distinguished several distinct notions of fairness, and the image-generation case maps most naturally onto representational fairness, the question of whether groups are represented in proportion to their presence, which is precisely what the audits measure and find

https://doi.org/10.62486/978-9915-9966-3-9.Ch08

wanting.(9,10,16) The same representational problem has been documented in vision-language models and in the demographic outputs of foundation models more broadly, which places the text-to-image finding in a larger pattern rather than isolating it.(28-30) And the domain-specific literature on sex and gender bias in artificial intelligence for health shows the same structural pattern in the systems closest to real harm, which is the strongest available reason to treat the image-generation bias as consequential rather than merely cosmetic.(31,32)

The vocabulary of representational fairness is worth holding onto, because it makes precise what the audits show. The audits do not show that the models make individual errors, in the way a classifier might mislabel a single image; they show that the distribution of output is systematically skewed, which is a property of the model as a system rather than of any single generation. This distinction matters for the mitigation question, because a distributional problem cannot be fixed by correcting individual outputs, and the review's finding that mitigations are weak follows from the distributional nature of the bias. The text-to-image case is thus a clear instance of the representational-fairness problem that the broader fairness literature has formalized, and its audit findings are best read in that formal vocabulary.(9,33)

One further property of the finding deserves emphasis, because it bears on how the problem should be weighed. The bias is not a failure of any particular model but a property of the training data and therefore of the entire current generation of models, and it will not be fixed by swapping one model for another.[5,6] This means the problem is systemic in the precise sense in which the fairness literature uses that term, and systemic problems require systemic responses, which the audit literature can inform but cannot alone supply. The review's findings therefore point outward, from the narrow technical question of what a model produces to the broader institutional question of who is responsible for what it produces, and that is the natural and appropriate place for the synthesis to end.

P3 ANSWERED

The answer to the third question is that mitigation is discussed far more than it is evaluated, and the evaluations that do exist are modest in both scope and result. The literature has identified the strategies, from prompting to fine-tuning to filtering, but it has not established that any of them removes the bias, and the honest summary is that mitigation remains an open problem.[2,8] The gap between the strength of the documentation and the weakness of the mitigation is the most consequential asymmetry the review finds, because it means the field can measure the problem far better than it can currently fix it, and deployment therefore proceeds with the biases documented but not yet corrected.

The mitigation problem is compounded by a finding that appears across the broader generative-bias literature: the biases are not only persistent but in some cases amplified and difficult to interpret, which makes them harder to correct than a simple distributional imbalance would suggest.[5,34] The occupational-bias line of work shows that even prompt-level interventions tend to move the bias rather than actually removing it, and the debiasing methods developed for vision-language models do not yet transfer cleanly to the generative image case.[28,35] The consequence is that the field's mitigation tool kit is borrowed from the text and vision-language literatures and has not been revalidated for text-to-image, which is itself a finding that the review records rather than resolves.

The gap between diagnosis and treatment is not unusual for a young field, and it is worth locating precisely. The audit literature has reached maturity in its measurement capacity: the included studies show that the bias can be elicited, classified, and quantified with modest resources, and the direction of the finding is stable. What has not matured is the corrective capacity, because a mitigation must be evaluated against a standard that the field has not yet defined, namely a benchmark of bias that a mitigated model should pass. Until such a benchmark exists, mitigation evaluations will remain ad hoc, and the trilemma identified above will persist. The review therefore reads the third question's comparatively thin answer not as a failure of the individual studies but as the predictable consequence of a missing piece of shared infrastructure.

The trilemma itself is the most compact answer to the third question, and it deserves to be stated in full once more. Prompt engineering is cheap and global but superficial, acting on the surface of the model and surviving only until a differently worded prompt defeats it. Fine-tuning is deep and model-internal but local and expensive, correcting the dimensions it targets while risking regressions elsewhere. Filtering is global and cheap but external, treating the symptom and inheriting the biases of the filter itself. No included study demonstrates a strategy that is all

https://doi.org/10.62486/978-9915-9966-3-9.Ch08

three of cheap, global, and deep, and the review's reading of the evidence is that the trilemma is a genuine trade-off rather than a gap that a single clever method will close. Mitigation, in short, is where the field's diagnosis runs substantially ahead of its treatment, and the third question's answer is therefore arguably the most important one the review gives.

For researchers, the implication is to standardize the audit protocol and to evaluate mitigation, not merely mention it, because the field currently documents bias with precision and corrects it with anecdote.[9,10] For the developers of these models, the implication is that the documented biases are a known, consistent and now well-characterized property of current systems, and releasing a model without an audit is releasing a system whose output is known in advance to be skewed.[30,36] A developer who ships a text-to-image model today is shipping, knowingly or not, a system that returns a male physician and a female nurse for the corresponding prompts, and the audit literature has now made that fact available; the responsibility is therefore no longer a matter of ignorance but of the choice of whether to act on what is known and documented. For policy, the implication is that audit is feasible, informative, and comparatively inexpensive, and therefore should be required, because the included studies show that the biases can be measured with modest resources, and measurement is the precondition for any serious mitigation.[33,37] The governance literature on algorithmic and political bias reaches the same conclusion from the policy side, that transparency and audit are the primary mechanisms by which biased systems are held accountable, and the text-to-image case is a ready-made application of that well-established principle.[6,38] The audit literature itself has begun to specify what a proper audit requires, and verifiable fairness auditing is a developing tool that could be applied directly to image generators, which would turn the field's prompt-based ad hoc audits into a formal, repeatable practice.[39,40] The review's most practical contribution is the demonstration that the field already knows how to measure the problem; what it lacks is the same rigor in fixing it.

DIRECTIONS

The review points to three directions, in increasing order of ambition. The first is a standardized audit benchmark, a fixed set of prompts and a fixed classification procedure that would make any model's bias measurable against the same yardstick, which is precisely the missing piece that the field's ad hoc audits currently lack.[10,16] The second is a convention for reporting audits, so that a study states its prompts, its classifier, and its sample size, making the audit reproducible in the way the field's current reporting does not consistently ensure.(39) The third is the revalidation of mitigation methods for the image case, because the field currently borrows mitigations from text and vision-language work without establishing that they transfer, and a program of evaluation against the new benchmark would supply that missing evidence.[28,35] Each direction is infrastructure rather than science, and each is the kind of thing a field builds once it recognizes that its current instruments have reached their limit.

The three directions also have a natural order of dependence, which is worth stating explicitly. The benchmark comes first, because the reporting convention and the mitigation evaluation both presuppose a shared test against which to measure; the reporting convention comes second, because it standardizes how the benchmark's results are communicated; and the mitigation evaluation comes last, because it requires both the benchmark and the convention to be in place before its results can be compared across studies. A field that builds them in this order will find the third much easier than a field that attempts mitigation evaluation without the first two, which is, in effect, precisely what the current literature has been attempting to do. The review's directions are therefore not a list but a sequence, and the sequence is the practical takeaway for the field.

CONCLUSIONS

The systematic review included 11 audit studies and found that they audit gender, racial and cultural bias with a nearly uniform prompt-based protocol, that they document a consistent, directional pattern of stereotype reproduction in text-to-image models, and that they discuss mitigation more often than they evaluate it. The documented biases are not in doubt: high-status occupations are skewed male and light-skinned and care occupations female, and the pattern holds across open and commercial models and across countries. Mitigation remains the field's open problem, with no included study demonstrating a strategy that removes the bias at scale.

https://doi.org/10.62486/978-9915-9966-3-9.Ch08

Standardizing the audit protocol and rigorously evaluating mitigations are the two steps that would convert the field's strong diagnosis into a comparable capacity for treatment, and they are the steps this review recommends in place of further diagnosis alone. In practical terms, a shared benchmark would let researchers, developers and regulators compare any new model against the same reference point, while controlled trials of mitigation would tell practitioners which interventions actually work rather than which ones sound promising. Until both exist, the bias documented here will continue to reproduce itself in the images that these models generate at scale.

https://doi.org/10.62486/978-9915-9966-3-9.Ch08

REFERENCES

Friedrich F, Brack M, Struppek L, Hintersdorf D, Schramowski P, Luccioni S, et al. Auditing and instructing text-to-image generation models on fairness. AI Ethics. 2025;5(3):2103-2123. doi: 10.1007/s43681-024-00531-5
Barve S, Mao A, Shi JM, Juneja P, Saha K. Social stereotypes in AI text-to-image generation. AI Ethics. 2026;6(3):296. doi: 10.1007/s43681-026-01146-8
Herrera MIS, Henao SIL, Beltrán-Luengas EM, Riaño-Moreno JC, Riaño-Moreno J. From paternalism to pixels: gender and racial stereotypes in AI-generated visual representations of doctor–patient relationships. AI Ethics. 2026;6(2):170. doi: 10.1007/s43681-025-00913-3
García-Ull FJ, Melero-Lázaro M. Gender stereotypes in AI-generated images. EPI. 2023:e320505. doi: 10.3145/epi.2023.sep.05
Huang LTL, Huang TR. Generative bias: widespread, unexpected, and uninterpretable biases in generative models and their implications. AI & Soc. 2026;41(3):1893-1905. doi: 10.1007/s00146-025-02533-1
Zhu Y. Systemic Bias in Artificial Intelligence: Focusing on Gender, Racial, and Political Biases. jaip. 2024;7(3). doi: 10.23977/jaip.2024.070324
Franco G, Sarzi-Ribeiro RA, Papa JP, da Costa KAP, Mahlow F. Bias beyond borders: quantifying gender and ethnic stereotypes across countries in AI image generation. AI Ethics. 2026;6(2):203. doi: 10.1007/s43681-026-01066-7
Vice J, Akhtar N, Sigal L, Hartley R, Mian A. On the fairness, diversity and reliability of text-to-image generative models. Artif Intell Rev. 2026;59(2):57. doi: 10.1007/s10462-025-11424-2
Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A Survey on Bias and Fairness in Machine Learning. ACM Comput Surv. 2022;54(6):1-35. doi: 10.1145/3457607
Caton S, Haas C. Fairness in Machine Learning: A Survey. ACM Comput Surv. 2024;56(7):1-38. doi: 10.1145/3616865
Kerruish E. Paused for thought: how stereotyping and anomalies in generated images prompt reflection on the technicity of AI. AI & Soc. 2026;41(4):3777-3787. doi: 10.1007/s00146-025-02798-6
Çakay B. Algorithmic Gaze: Gender and Ethnicity Bias on Algorithmic Intelligence Text-to-Image Platforms. KADEM Kadın Araştırmaları Dergisi. 2026;(1):91-126. doi: 10.21798/kadem.2026.219
Schneider M, Hagendorff T. Investigating toxicity and Bias in stable diffusion text-to-image models. Sci Rep. 2025;15(1):31401. doi: 10.1038/s41598-025-12032-4
AlDahoul N, Rahwan T, Zaki Y. AI-generated faces influence gender stereotypes and racial homogenization. Sci Rep. 2025;15(1):14449. doi: 10.1038/s41598-025-99623-3
Wu Y, Nakashima Y, Garcia N. Revealing Gender Bias from Prompt to Image in Stable Diffusion. J Imaging. 2025;11(2):35. doi: 10.3390/jimaging11020035
Le Quy T, Roy A, Iosifidis V, Zhang W, Ntoutsi E. A survey on datasets for fairness-aware machine learning. WIREs Data Min & Knowl. 2022;12(3):e1452. doi: 10.1002/widm.1452
Guilbeault D, Delecourt S, Hull T, Desikan BS, Chu M, Nadler E. Online images amplify gender bias. Nature. 2024;626(8001):1049-1055. doi: 10.1038/s41586-024-07068-x
Mir AA, Sofi IA, Ayub T, Ganaie SA. Prevalence of gender stereotyping among AI image generators. Global Knowledge, Memory and Communication. 2026:1-20. doi: 10.1108/gkmc-11-2024-0737
Nadeem M, Sohail SS, Cambria E, Schuller BW, Hussain A. Gender Bias in Text-to-Video Generation Models: A Case Study of Sora. IEEE Intell Syst. 2025;40(3):10-15. doi: 10.1109/mis.2025.3561475
Shiraishi M, Banda CH, Nakajima M, Nakazwe M, Wong ZY, Tomioka Y, et al. Gender and racial diversity Assumed by text-to-image generators in microsurgery and plastic surgery-related subspecialities. Journal of Hand and Microsurgery. 2025;17(1):100196. doi: 10.1016/j.jham.2024.100196
Currie G, Chandra C, Kiat H. Gender Bias in Text-to-Image Generative Artificial Intelligence When Representing Cardiologists. Information. 2024;15(10):594. doi: 10.3390/info15100594
Currie G, Hewis J, Hawk E, Rohren E. Gender and Ethnicity Bias of Text-to-Image Generative Artificial Intelligence in Medical Imaging, Part 1: Preliminary Evaluation. J Nucl Med Technol. 2024;52(4):jnmt.124.268332. doi: 10.2967/jnmt.124.268332
Sun L, Wei M, Sun Y, Suh YJ, Shen L, Yang S. Smiling women pitching down: auditing representational and presentational gender biases in image-generative AI. Journal of Computer-Mediated Communication. 2023;29(1):zmad045. doi: 10.1093/jcmc/zmad045
Mahl D, Guenther L. Skewed toward uniformity? Investigating algorithmic stereotyping and amplification in AI-generated images. Information, Communication & Society. 2026:1-20. doi: 10.1080/1369118x.2026.2740745
Elsharif W, Alzubaidi M, Agus M. Cultural Bias in Text-to-Image Models: A Systematic Review of Bias Identification, Evaluation, and Mitigation Strategies. IEEE Access. 2025;13:122636-122659. doi: 10.1109/access.2025.3585745
Struppek L, Hintersdorf D, Friedrich F, Br M, Schramowski P, Kersting K. Exploiting Cultural Biases via Homoglyphs in Text-to-Image Synthesis. jair. 2023;78:1017-1068. doi: 10.1613/jair.1.15388
Cheong M, Abedin E, Ferreira M, Reimann R, Chalson S, Robinson P, et al. Investigating Gender and Racial Biases in DALL-E Mini Images. ACM J Responsib Comput. 2024;1(2):1-20. doi: 10.1145/3649883
Li B, Yao Y, Tan J, Gong R, Lu J, Luo Y. Rectify representation bias in vision-language models for long-tailed recognition. Neural Networks. 2024;172:106134. doi: 10.1016/j.neunet.2024.106134
Nadeem M, Sohail SS, Cambria E, Afreen S. South Asian biases in language and vision models. Nat Mach Intell. 2025;7(11):1775-1777. doi: 10.1038/s42256-025-01144-1
Yang Y, Liu Y, Liu X, Gulhane A, Mastrodicasa D, Wu W, et al. Demographic bias of expert-level vision-language foundation models in medical imaging. Sci Adv. 2025;11(13):eadq0305. doi: 10.1126/sciadv.adq0305

https://doi.org/10.62486/978-9915-9966-3-9.Ch08

Cirillo D, Catuara-Solarz S, Morey C, Guney E, Subirats L, Mellino S, et al. Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. npj Digit Med. 2020;3(1):81. doi: 10.1038/s41746-020-0288-5
Buslón N, Cortés A, Catuara-Solarz S, Cirillo D, Rementeria MJ. Raising awareness of sex and gender bias in artificial intelligence and health. Front Glob Womens Health. 2023;4:970312. doi: 10.3389/fgwh.2023.970312
Sarmah N, Bhattacharyya R. Bias and fairness in vision–language models and large language models: a survey on South Asian festival representations. Int J Data Sci Anal. 2026;22(1):187. doi: 10.1007/s41060-026-01162-1
Ramachandran S, Rattani A. Leveraging diffusion and Flow Matching Models for demographic bias mitigation of facial attribute classifiers. Neurocomputing. 2025;636:130024. doi: 10.1016/j.neucom.2025.130024
Liu Y, Lin A, Peng S, Cai Z. Occupational Gender Bias in Chinese Generative AI Models: Cross-Model Evidence of Stereotypical Amplification and Systematic Underrepresentation. Systems. 2026;14(3):286. doi: 10.3390/systems14030286
Hu T, Kyrychenko Y, Rathje S, Collier N, van der Linden S, Roozenbeek J. Generative language models exhibit social identity biases. Nat Comput Sci. 2024;5(1):65-75. doi: 10.1038/s43588-024-00741-1
Ntoutsi E, Fafalios P, Gadiraju U, Iosifidis V, Nejdl W, Vidal M, et al. Bias in data-driven artificial intelligence systems—An introductory survey. WIREs Data Min & Knowl. 2020;10(3):e1356. doi: 10.1002/widm.1356
Peters U. Algorithmic Political Bias in Artificial Intelligence Systems. Philos Technol. 2022;35(2):25. doi: 10.1007/s13347-022-00512-8
Simbeck K. They shall be fair, transparent, and robust: auditing learning analytics systems. AI Ethics. 2024;4(2):555-571. doi: 10.1007/s43681-023-00292-7
Tang G, Tan W, Cai M. Privacy-Preserving and Trustless Verifiable Fairness Audit of Machine Learning Models. IJACSA. 2023;14(2). doi: 10.14569/ijacsa.2023.0140294

https://doi.org/10.62486/978-9915-9966-3-9.Ch08

Chapter 9 / Capítulo 9

SAP Superintelligence Series
ISSN: 3144-5624
ISBN (complete work): 978-9915-9851-0-7

Superintelligence Series Volume 5: Intelligent Visual Arts and Audiovisual Engineering: Generative AI, Computational Creativity and Superintelligence
ISBN (this volume): 978-9915-9966-5-3
DOI: 10.62486/978-9915-9966-3-9.Ch09
Pages: 109-120

Check for updates

© 2026 The authors. This is an open access book chapter distributed under the terms of the Creative Commons Attribution (CC BY) 4.0 License.

Neural radiance fields and 3D Gaussian splatting in audiovisual production

Campos de radiancia neuronal y salpicado gaussiano 3D en la producción audiovisual

Fiorella Luz De La Cruz Espinoza[1]
Norma Beatriz Ríos[2]
Carlos Jesús Canova Barrios[3,4]

1 Universidade Federal do Rio Grande do Sul. Brazil.
2 Universidad Hospital Italiano de Buenos Aires. Buenos Aires, Argentina.
3 Universidad Abierta Interamericana. Buenos Aires, Argentina.
4 Universidad de Ciencias Empresariales y Sociales. Buenos Aires, Argentina.

ABSTRACT

Neural rendering has introduced two dominant approaches to synthesizing photorealistic views of a scene, neural radiance fields and 3D Gaussian splatting, whose different trade-offs matter directly for audiovisual production. This narrative review synthesizes the two approaches across five axes, representation and training, rendering quality, rendering speed, editability, and dynamic-scene support, drawing on the literature published between 2018 and 2026. Neural radiance fields represent a scene as a continuous implicit function optimized by volume rendering, which yields high quality and strong anti-aliasing but is slow to train and render without baking or acceleration. Three-dimensional Gaussian splatting represents the same scene as an explicit set of anisotropic Gaussians rendered by rasterization, which delivers real-time frame rates with comparable quality but introduces artifacts and editing challenges of its own. The two approaches also differ in editability and in their treatment of moving scenes. The synthesis concludes that the choice between them is a production decision rather than a technical ranking, and that the field's open problems, temporal consistency, production-grade editability and standardized evaluation, are the likely routes by which neural rendering will enter routine audiovisual workflows.

Keywords: neural rendering; neural radiance fields; 3D Gaussian splatting; novel view synthesis; audiovisual production; real-time rendering

RESUMEN

El renderizado neuronal ha introducido dos enfoques dominantes para sintetizar vistas fotorrealistas de una escena, los campos de radiancia neuronal y el salpicado gaussiano 3D, cuyas distintas compensaciones importan directamente para la producción audiovisual. Esta revisión narrativa sintetiza ambos enfoques en cinco ejes, representación y entrenamiento, calidad de renderizado, velocidad, editabilidad y soporte de escenas dinámicas, a partir de la bibliografía publicada entre 2018 y 2026. Los campos de radiancia neuronal representan la escena como una función implícita continua optimizada mediante renderizado volumétrico, lo que ofrece alta calidad y buen antialiasado, pero es lento de entrenar y renderizar sin aceleración. El salpicado gaussiano 3D representa la misma escena como un conjunto explícito de gaussianas anisótropas renderizadas por rasterización, lo que logra velocidades en tiempo real con calidad comparable, aunque introduce sus propios artefactos y retos de edición. Los dos enfoques también difieren en editabilidad y en el tratamiento de escenas en movimiento. La síntesis concluye que la elección entre ambos es una decisión de producción, no una clasificación técnica, y que los problemas abiertos del campo, consistencia temporal, editabilidad de grado productivo y evaluación estandarizada, son las vías probables por las que el renderizado neuronal entrará en los flujos audiovisuales habituales.

Palabras clave: renderizado neuronal; campos de radiancia neuronal; salpicado gaussiano 3D; síntesis de vistas; producción audiovisual; renderizado en tiempo real

INTRODUCTION

Audiovisual production has always depended on the ability to render a three-dimensional scene as a two-dimensional image from an arbitrary camera, and for most of its history that ability has been bought with explicit geometry, meshes, materials and lights assembled by hand. Neural rendering changes the terms of that bargain. Instead of modeling a scene explicitly, a neural rendering method stores the scene inside a learned function or a learned set of primitives and synthesizes new views by querying it, which collapses the distance between capture and image and opens the way to free-viewpoint video, virtual production and post-capture relighting without a conventional three-dimensional pipeline.[1,2] The capacity to reconstruct a photorealistic scene from photographs and then move the camera through it is directly relevant to every stage of audiovisual work, from previsualization and virtual sets to visual effects, and the field has therefore turned toward these methods with unusual speed.(3,4)

Two approaches dominate this movement, and they represent opposite answers to the same question of how a scene should be stored. The first, the neural radiance field, encodes a scene as a continuous implicit function, typically a multilayer perceptron, that maps a position and a viewing direction to a color and a density, and it is optimized through differentiable volume rendering.[1] The second, three-dimensional Gaussian splatting, encodes the scene as an explicit cloud of anisotropic Gaussians, each carrying position, covariance, opacity and color, that are projected and rasterized in order.[2] The two approaches were introduced within three years of each other and now split the literature between them, with the Gaussian representation overtaking the radiance field as the default for real-time work after 2023.[3,5] Their contrast is informative precisely because they are competitors for the same production tasks, and because they fail in different, complementary ways.

The comparison matters for production because the two approaches sit on different sides of the quality-speed frontier that audiovisual work constantly navigates. A radiance field produces images of high fidelity and handles anti-aliasing and unbounded scenes cleanly, but its volume rendering queries the network hundreds of times per ray, which makes naive rendering far from real time.[1,6] Gaussian splatting reverses that relationship, rasterizing millions of primitives to reach real-time frame rates at full resolution while retaining comparable quality, at the cost of new artifacts such as popping and floaters.[2,7] For a supervisor choosing between a method that is slow but stable and one that is fast but noisy, the distinction is not academic, and neither the graphics literature nor the production literature has yet given practitioners a clear, comparative account of it.

The moment is also right for such an account, because the field has just passed through a reorientation that changed what the two methods are competing for. The radiance field founded the field in 2020 and carried it through its first expansion, but the Gaussian representation, introduced in 2023, overtook it as the default for real-time work with a speed that the literature itself remarks on, and the surveys that followed had to be written around two lineages rather than one.[3,4] A comparison made earlier would have been premature, because the Gaussian side of the ledger was still empty; a comparison made now can draw on three years of methods on both sides of the divide, and it arrives at a time when production houses are actively deciding which representation to invest in for virtual production, visual effects and free-viewpoint work.

Surveys of neural rendering have grown abundant, but they are organized around a single representation rather than around the contrast between the two. Comprehensive reviews trace the radiance-field lineage and its extensions, and parallel surveys trace Gaussian splatting and its applications, yet a reader who wants to know how the two actually compare for production must assemble the answer from separate accounts.(5,8-10) A smaller number of works confront the two directly, either in benchmark evaluations or in method papers that position one against the other, and those scattered comparisons are the raw material this review assembles.[3,4] The gap, then, is not a shortage of evidence but the absence of a synthesis that is organized by the contrast itself.

This review addresses that gap by comparing neural radiance fields and three-dimensional Gaussian splatting across five axes that correspond to the decisions a production actually makes: how the scene is represented and trained, how good the rendered image is, how fast it renders, how editable it is, and how well it handles moving scenes. The objective is to compare the two approaches as neural rendering methods for audiovisual production and to identify, from the comparison, the open problems that most limit their adoption in practice. Each axis is developed in the section that follows, and the five answers are integrated in the discussion into a single account of where each approach fits within a production workflow.

https://doi.org/10.62486/978-9915-9966-3-9.Ch09

METHODS

The review followed the SANRA guidance for narrative reviews, which means that the search and selection are described with transparency but without the formal screening counts of a systematic review.(11) The object of the synthesis is the primary literature on the two neural rendering approaches, together with the surveys and evaluation studies that position one against the other. The review is comparative by design, and its method is to hold a fixed set of production-relevant questions against each approach in turn and to report where the literature converges, where it diverges, and where it is silent.

Selection followed three criteria applied to title and abstract. A study was retained if it introduced, evaluated or surveyed a neural radiance field or Gaussian splatting method with direct relevance to one of the five axes; if it was published within the search window; and if it was written in English or Spanish. Works on neural rendering for unrelated tasks, such as medical imaging or autonomous navigation, were admitted only where they contributed a finding that transfers to audiovisual production, such as real-time performance or dynamic-scene handling.[12,13] The retained material was divided into a core of primary method papers that populate the synthesis and a supporting set of surveys and adjacent studies that supply context in the introduction and discussion.

The synthesis is organized around five contrast axes, each a question that a production team would ask. The first axis concerns representation and training, the second rendering quality, the third rendering speed, the fourth editability and control, and the fifth dynamic scenes and temporal consistency. For each axis the review reports what the two approaches offer, cites the primary studies that establish each capability, and notes the divergences in the literature where studies disagree about the magnitude of an advantage. Because the two approaches are measured on heterogeneous benchmarks with heterogeneous metrics, the synthesis is narrative rather than quantitative, and no pooled estimate is attempted; where a number is reported, it is attributed to the study that produced it.

A synthesis table (Table 1) summarizes the five axes with the representative studies that support each finding, and a second table (Table 2) sets the two approaches side by side on the production-relevant properties that recur across the axes.

Table 1. Synthesis of the five contrast axes.

Contrast axis	Neural radiance field	3D Gaussian splatting
Representation and training	Continuous field optimized by volume rendering	Explicit Gaussians rasterized in order
Rendering quality	High fidelity, clean anti-aliasing, fails under sparse input	Comparable fidelity, real time, prone to floaters
Rendering speed	Real time only with baking or compression	Real time natively at 1080p
Editability and control	Indirect editing via stylization, palettes or layers	Explicit primitives, easier but not mesh-grade
Dynamic scenes	Mature deformation and topology, human performance	Real-time dynamic playback, temporal consistency open

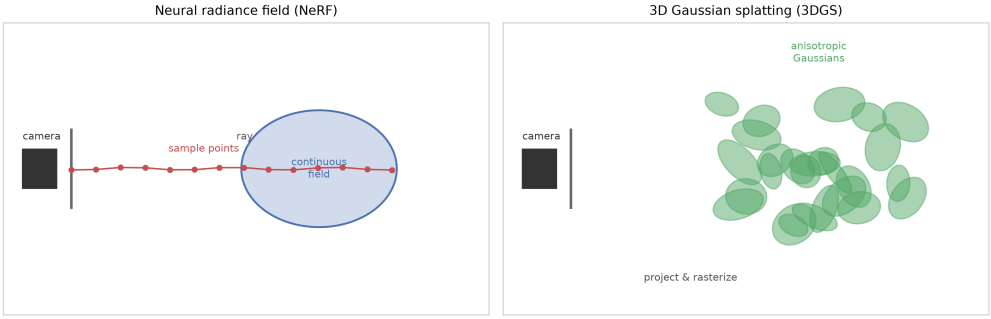
Table 2. Production-relevant properties of the two neural rendering approaches.

Property	Neural radiance field	3D Gaussian splatting
Scene representation	Continuous implicit function	Explicit set of anisotropic Gaussians
Training time	Hours, minutes with hash encoding	Tens of minutes
Rendering speed	Offline, real time only with baking	Real time at full resolution
Editability	Indirect, bespoke methods	Explicit primitives, partial editing
Dynamic support	Mature deformation and topology	Real-time playback, weaker temporal consistency
Primary failure	Sparse input, transient capture	Floaters, popping, complex scenes

A conceptual figure (Figure 1) illustrates the structural difference between the two representations, a continuous field sampled by ray marching and a discrete set of Gaussians projected by splatting, since that structural difference is the source of most of the axes that follow.

https://doi.org/10.62486/978-9915-9966-3-9.Ch09

Figure 1. The two neural rendering representations compared in this review.



Note: The left panel shows a radiance field sampled by ray marching, and the right panel shows a cloud of anisotropic Gaussians projected by splatting.

DEVELOPMENT

REPRESENTATION AND TRAINING

The two approaches differ first in what a scene is. A neural radiance field is a continuous function, a multilayer perceptron that takes a five-dimensional coordinate, three for position and two for viewing direction, and returns the volume density and view-dependent color at that point. (1) A view is synthesized by marching along each camera ray, sampling the function at many points, and compositing the samples by the volume rendering equation, and the network is optimized so that the rendered views reproduce the input photographs. This is a compact and flexible representation, but the optimization is per-scene and, in the original formulation, takes hours on a single scene, and the rendering step queries the network hundreds of times per ray, which is the root of both the quality and the cost of the method.

Gaussian splatting stores the same scene in the opposite way, as an explicit cloud of anisotropic three-dimensional Gaussians, each of which carries a position, a covariance describing its shape, an opacity and a set of spherical-harmonic coefficients describing its color, initialized from the sparse point cloud produced by structure from motion and then optimized against the photographs.(2) Rendering projects the Gaussians onto the image plane, sorts them, and composites them front to back by alpha blending, a rasterization step that runs on the graphics pipeline rather than through a network. The consequence is that training converges in tens of minutes rather than hours, and rendering reaches interactive rates without the acceleration tricks that radiance fields require.(2)

The training gap has been narrowed from both sides, and the narrowing itself is part of the comparison. On the radiance-field side, the multiresolution hash encoding and related feature-grid encodings replaced the coordinate network's raw input with a learned feature lookup, cutting per-scene optimization from hours to seconds or minutes while preserving most of the quality.(14) Baking approaches take the further step of precomputing a trained radiance field into a sparse grid that can be rendered without the network at all.(6) On the Gaussian side, the explicit representation arrived already fast, and the recent work has concentrated on scaling it to very large datasets through hierarchical and level-of-detail structures rather than on closing a training gap.[15,16] The axis therefore ends less in a clean victory than in a convergence: the two representations began far apart in cost and have been pulled toward the same middle by different mechanisms.

RENDERING QUALITY

On rendering quality the two approaches are close enough that the differences are matters of regime rather than of rank. The original radiance field demonstrated that a coordinate network could reproduce view-dependent appearance, including reflections and translucency, at a fidelity that was the point of the field's founding, and the anti-aliased variant of the method, which traces cones rather than rays, extended that fidelity to unbounded outdoor scenes.[1,17] Gaussian splatting reports quality on the same benchmark families that is comparable to the best radiance-field results while rendering at real-time rates, a combination that had previously been thought to require a quality trade-off.(2) The primary studies that evaluate both representations on shared test scenes consistently place them within a small margin of each other, with the ranking depending

https://doi.org/10.62486/978-9915-9966-3-9.Ch09

on the dataset and on the metric used, a point that matters more for benchmark reporting than for the production decision it is meant to inform.

The failures differ in kind, and a production reader needs the failures more than the aggregate scores. Radiance fields degrade most visibly under sparse input, where the geometry is under-constrained and the network produces floaters and holes, a weakness addressed by geometric regularization and by depth or RGB-D priors.[18,19] They also struggle with in-the-wild captures that contain transient occluders and moving objects, which transient modeling attempts to inpaint.[20] Gaussian splatting degrades most visibly in the other direction, not from under-constrained geometry but from the brittleness of its optimization and its sorting, which produces popping artifacts, floaters and view-dependent inconsistencies that dedicated methods target with sorted splatting, distractor removal and level-of-detail selection.(7,16,21) In short, radiance fields fail where the input is thin, and Gaussians fail where the scene is complicated, and neither failure is the same as the other.

A further quality distinction concerns the treatment of view-dependent appearance, which production work relies on more than benchmark suites reveal. The radiance field is built around a viewing direction as an explicit input, which is why reflections, translucency and specular highlights are among its strengths, and why anti-aliased variants extend those strengths to outdoor, unbounded scenes.[1,17] Gaussian splatting models view dependence through spherical-harmonic coefficients on each primitive, which is sufficient for many scenes but is the source of specific artifacts when reflections and transmission are prominent, and dedicated Gaussian formulations exist to model planar transmission and reflection precisely because the base representation handles them imperfectly.(22) This is not a ranking so much as a reminder that the two representations encode the same optical phenomena differently, and that a production that cares about glass, water or polished surfaces should weight the difference accordingly.

RENDERING SPEED

Rendering speed is the axis on which the two approaches are furthest apart and the one that has driven the field's reorientation. A naive radiance field is far from real time because every pixel requires the network to be evaluated hundreds of times along its ray, and early real-time results were bought by reducing resolution, by baking the network into a grid, or by compressing the representation into memory-efficient forms that fit a browser or a mobile device.(1,6,23) These are genuine real-time radiance fields, but they are the product of an extra engineering pass, and each acceleration trades away some of the flexibility of the original implicit representation.

Gaussian splatting is real time by construction. Because its primitives are projected and rasterized on the graphics pipeline rather than sampled through a network, the method reports real-time display at full high-definition resolution on consumer hardware, and this is the single property most responsible for its rapid adoption.(2) Subsequent work extends the real-time claim outward, to very large datasets through hierarchy, to extended-reality workloads through architecture co-design, and to planar transmission and reflection through dedicated Gaussian formulations.(15,22,24) The speed axis is thus the clearest asymmetry in the comparison: radiance fields can be made real time, but Gaussians are real time, and the distinction is felt in any interactive production context.

A second speed distinction cuts across both representations, between methods that train per scene and methods that generalize. Per-scene methods, which dominate both lineages, must re-optimize for each new capture, and that optimization time, whether minutes or hours, is a fixed entry cost.[1,2] Generalizable methods learn a mapping from images to a scene representation once and then synthesize views of unseen scenes in a forward pass, at the cost of some fidelity.(25-27) For production, the difference is between a method that is slow once and fast thereafter, and a method that is fast always but slightly less faithful, a distinction the axis preserves rather than collapses.

The acceleration techniques developed for radiance fields deserve separate mention because they illustrate how far the representation can be pushed before it stops being the representation. Real-time radiance fields for single-image portraits reach twenty-four frames per second on consumer hardware by predicting a triplane representation directly from a single photograph, and visual-hull acceleration reduces the cost of the volume rendering step by restricting sampling to the space a scene is known to occupy.[28,29] These methods matter for production because they target the two constraints that most limit on-set use, the need for a fast, per-shot result and the need to start from ordinary, unposed footage rather than from a

https://doi.org/10.62486/978-9915-9966-3-9.Ch09

calibrated multi-view rig. The radiance field is therefore not simply slow; it is slow in a way that a series of published accelerations has progressively, and only partially, dismantled.

EDITABILITY AND CONTROL

Editability is where both approaches reveal their immaturity, and where the contrast between implicit and explicit is most instructive. A radiance field is a network, and a network offers no direct handle on the objects it has learned, so editing must be achieved indirectly, through text-driven stylization that adjusts the field's appearance, through palette mappings that relight or recolor a frozen scene, or through layered decompositions that separate a dynamic scene into editable entities.(30-32) These are real editing capabilities, but they are each bespoke, and none of them gives a production artist the familiar operations of moving, deleting or reshaping an object that a mesh or a rig would provide.

Gaussian splatting is more promising on this axis because its primitives are explicit, and it inherits from point-based graphics the ability to deform, manipulate and even reconstruct a surface from its Gaussians in real time.[33,34] Yet the promise is only partially realized. The Gaussians are not objects, but millions of disconnected primitives, and editing them at the level a production expects, selecting a character, retiming a performance, re-lighting a shot, still requires custom machinery, since the primitives carry no semantic grouping and no articulation that an artist could grab. The axis therefore lands on a specific conclusion: the explicit representation is easier to edit than the implicit one, but neither has yet reached the editability of the conventional pipeline it aspires to replace.

DYNAMIC SCENES AND TEMPORAL CONSISTENCY

Moving scenes are the axis on which the audiovisual stakes are highest, because production is concerned with people and cameras in motion rather than with static objects. The radiance-field lineage addressed motion early and thoroughly. Non-rigid formulations model a scene as a canonical volume deformed over time, hyper-dimensional approaches add a time-varying topology to that canonical model, and streaming representations decompose a four-dimensional scene into static, deforming and new regions so that it can be played back as video.(35-37) Human performance is the clearest production target, and human-centric radiance fields demonstrate free-viewpoint playback of full-body motion from multi-view video at a fidelity aimed at film, with related work reconstructing human performance through neural meshes and generalizable human renderers.(38-40)

Gaussian splatting entered dynamic scenes after its static success, and it brought the same real-time property with it. Four-dimensional Gaussian splatting renders moving scenes in real time by making the Gaussians time-varying, and motion-aware variants mine optical flow to reduce the redundancy that dynamic scenes introduce.[41,42] Human rendering with Gaussians follows the same arc, from per-subject optimization to generalizable pixel-wise approaches that render a novel person from sparse multi-view capture without re-optimization.[43,44] The dynamic axis does not produce a clean winner: radiance fields offer the more mature treatment of deformation and topology, while Gaussians offer the real-time playback, and the literature has not yet combined both in a single method that production could adopt without reservation.

The axis closes on the problem that remains open for both, temporal consistency. A dynamic scene is judged not frame by frame but across time, and both representations are reported to flicker, pop and drift between frames under motion, a defect that is acknowledged in the dynamic literature and that no included method claims to have eliminated.[37,41] Temporal consistency is therefore the clearest single gap on the production path, and it is carried forward into the discussion as the problem the next generation of methods will most plausibly address, since a scene that renders beautifully in a still but flickers in playback is not yet a production asset.

DISCUSSION

The comparison across the five axes resolves into a single, stable picture: neural radiance fields and Gaussian splatting are not a better and a worse method, but two methods that trade the same two properties, fidelity and speed, in opposite directions, and that fail in different regimes. A radiance field buys fidelity and clean anti-aliasing with a slow, implicit representation

https://doi.org/10.62486/978-9915-9966-3-9.Ch09

that is costly to train, costly to render and difficult to edit, while Gaussian splatting buys real-time rendering and an explicit, manipulable representation at the cost of new artifacts and of an optimization that is brittle on complex scenes. This reading aligns with the surveys that describe the Gaussian representation as overtaking the radiance field after 2023 for real-time work, while acknowledging that the radiance field retains advantages in unbounded scenes and in the maturity of its dynamic treatment.[3,4] The review's contribution is to state the trade-off explicitly and to locate each production decision within it.

The quality axis deserves emphasis because it is the one on which a casual reader could misread the literature. The two approaches report comparable aggregate scores, but the review's reading is that they are not equally robust: radiance fields fail when the input is sparse, and Gaussians fail when the scene is complicated, and the surveys that report only aggregate metrics obscure this asymmetry.[5,10] The discrepancy with a purely benchmark-driven comparison is explained by what the benchmarks measure, which is average error on fixed scenes rather than failure behavior at the margins, and the review therefore treats the failure regimes as more informative for production than the mean scores. This is a case where the method of comparison, by contrast rather than by ranking, changes the conclusion.

The speed advantage of Gaussian splatting is the least contested finding in the review and the one with the clearest production consequence. Where a radiance field must be baked, compressed or accelerated to reach real time, Gaussian splatting reaches it natively, and the difference is felt in any context where an operator must interact with the scene, virtual production, previsualization and on-set review being the obvious examples.(2,6,23) The caveat, which the axis records, is that real-time is not the same as correct, and the artifacts that Gaussian methods accept in exchange for speed are exactly the defects that a broadcast-grade deliverable cannot tolerate. The practical implication is that the two methods may coexist in a single pipeline, with Gaussians for interaction and radiance fields, or a baked radiance field, for the final render.

The editability finding is the most sobering for the claim that neural rendering will soon replace conventional pipelines, and it should be stated without softening. Neither representation offers the operations that a production artist expects from a mesh, and while the explicit Gaussian representation is easier to edit than the implicit radiance field, both remain far from the editability of a rigged model, a point the editing literature confirms from its own survey of the field.(45) The review's reading is that editability, not rendering, is the true bottleneck for adoption, because a photorealistic render that cannot be art-directed is a novelty rather than a tool, and the production literature on artificial intelligence in film makes the same point from the practitioner's side.(46) The two approaches will enter routine production not when they render well, but when they can be directed.

The dynamic-scene axis connects the comparison to the specific concerns of audiovisual work, where the subject is usually a person in motion. The radiance-field lineage has the more mature treatment of deformation, topology and human performance, with dedicated human-centric methods aimed explicitly at film, while Gaussian splatting contributes real-time dynamic playback that the radiance-field methods cannot match without heavy acceleration.(35,38,41) The axis identifies a gap rather than a winner, and the gap is temporal consistency, the tendency of both representations to flicker and drift between frames, which no included method claims to have eliminated and which is the most direct technical obstacle between the current state of the field and a broadcast deliverable.[37,42] This gap is the review's central recommendation for future work.

The comparison also carries a caution that the technical literature rarely states but that the audiovisual context forces forward. Neural rendering does not only synthesize views, it synthesizes views of people, and the same capacity that produces free-viewpoint performance can produce fabricated performance, a concern the documentary community has already confronted through the practice of deepfakes.(47) The ethical dimension is not an afterthought to the comparison but a property of the technology that production adopts, and a review that ignores it would understate what is actually at stake when a photorealistic, directable scene representation becomes routine. The two approaches are not distinguished by this concern, but the concern attaches to both, and it reinforces the review's emphasis on editability and control as the property that separates a legitimate production tool from a fabrication instrument.

The axes are not abstract, and two application domains make their practical force concrete. In cultural-heritage documentation, radiance fields have been evaluated across heritage datasets captured with smartphone video and tourist photography, where the finding is that accelerated variants of the method reach the quality a documentation workflow needs, a use that values fidelity over interactivity and therefore favors the radiance-field side of the trade-off.(48) In virtual-reality

https://doi.org/10.62486/978-9915-9966-3-9.Ch09

applications, by contrast, non-photorealistic rendering is used precisely as a feedback mechanism that keeps a user immersed, a reminder that photorealism is not always the production goal and that the speed and interactivity of a representation can matter more than its fidelity in a given application.(49) These domains illustrate the review's central claim in reverse: the correct choice of representation is decided by the production axis that a task weights most, not by a universal ranking of the two methods.

The review also sits in relation to the surveys that already exist, and that relation should be stated plainly. Overviews of the radiance field trace its technology and applications, and parallel surveys of Gaussian splatting do the same for the explicit representation, but each treats one lineage, and a practitioner comparing the two must read across them.(5,8,9,50) This review is distinguished by holding the contrast fixed, so that the same question is asked of both methods in the same section, which is what turns a pair of adjacent literatures into a single comparison. The comparison itself is the contribution, and the five axes are the instrument that makes it possible.

For a production pipeline, the comparison resolves into a division of labor rather than a choice of one method over the other. Gaussian splatting's native real-time rendering suits the interactive stages, previsualization, on-set review and virtual production, where an operator must move through the scene and the cost of an artifact is low because the image is provisional.[2,24] A radiance field, or a baked radiance field, suits the final render, where the camera is fixed, the image must be clean, and the slow, stable fidelity of the implicit representation is worth the wait.[6,17] The two methods are therefore not substitutes but complements, and the practical implication is that a pipeline that adopts both, Gaussians for interaction and fields for delivery, gains more than one that commits to either.

The comparison also locates neural rendering within the larger advance of artificial intelligence into audiovisual production, and that larger frame is part of what gives the technical trade-off its significance beyond the computer-graphics venues in which the methods were first reported. Neural rendering is not an isolated technique but one point on a curve that runs from automated editing and synthesis to the directable, photorealistic reconstruction of a scene from photographs, and the production literature already reads artificial intelligence as a force moving through pre-production, production and post-production as a whole.(46) What distinguishes neural rendering on that curve is that it concerns the image itself rather than the tasks around the image, which is why its open problems, temporal consistency, editability and evaluation, are also the points at which a general-purpose AI capacity becomes a dependable production tool. The choice between the two representations is therefore a specific instance of a broader decision that production is making about which AI capabilities to adopt and which to treat as provisional, and the review's axes are the concrete terms in which that decision is made.

The agenda that follows from the review is specific rather than general:

- Temporal consistency in dynamic scenes is the single highest-value technical problem, because it is the one that separates current neural rendering from broadcast use, and it is shared by both approaches; a method that holds a moving scene stable across frames would do more for adoption than any incremental improvement in mean fidelity.
- Production-grade editability, meaning selection, deformation, retiming and relighting of semantic entities rather than of millions of primitives or of a frozen network, is the property most likely to determine whether either approach is adopted, and it is best pursued by attaching semantic grouping to the explicit Gaussian representation, which already offers the primitives that grouping could bind.
- The field needs a shared benchmark that reports failure regimes as well as mean error, so that the quality comparison the review could only make qualitatively can be made quantitatively, and so that the complementary failure modes identified here can be measured rather than asserted.
- A fourth item follows from the capture side of the comparison, standardizing the input from which production scenes are reconstructed, since the sparse and unposed footage that production actually provides is not the calibrated multi-view input that both representations assume.

These elements are the routes by which the comparison in this review would close, and they are offered as concrete, testable directions for research rather than as the generic call for more study.

https://doi.org/10.62486/978-9915-9966-3-9.Ch09

CONCLUSIONS

Neural radiance fields and three-dimensional Gaussian splatting were compared across the five axes that matter to audiovisual production, and the comparison shows that the two approaches trade fidelity against speed in opposite directions while failing in different regimes. The radiance field offers a continuous implicit representation with strong quality, clean anti-aliasing and a mature treatment of dynamic scenes, at the cost of slow training, slow naive rendering and difficult editing. Gaussian splatting offers an explicit representation with native real-time rendering and easier manipulation, at the cost of new artifacts and a brittle optimization on complex scenes. On dynamic scenes the two approaches are distinguished less by a winner than by a shared gap, temporal consistency, which remains unresolved for both. The review concludes that the choice between them is a production decision governed by the specific axis that a given project weights most heavily, and that the field's route into routine use runs through three problems in particular: temporal consistency, production-grade editability and an evaluation practice that reports failure as carefully as it reports mean error.

https://doi.org/10.62486/978-9915-9966-3-9.Ch09

REFERENCES

Mildenhall B, Srinivasan PP, Tancik M, Barron JT, Ramamoorthi R, Ng R. NeRF. Commun ACM. 2022;65(1):99-106. doi: 10.1145/3503250
Kerbl B, Kopanas G, Leimkuehler T, Drettakis G. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. ACM Trans Graph. 2023;42(4):1-14. doi: 10.1145/3592433
Gao K, Gao Y, He H, Lu D, Xu L, Li J. Neural Radiance Fields in 3D Vision: A Comprehensive Review. Comp Visual Med. 2026;12(4):851-905. doi: 10.26599/cvm.2026.9450535
Fei B, Xu J, Zhang R, Zhou Q, Yang W, He Y. 3D Gaussian Splatting as a New Era: A Survey. IEEE Trans Visual Comput Graphics. 2025;31(8):4429-4449. doi: 10.1109/tvcg.2024.3397828
Chen G, Wang W. A Survey on 3D Gaussian Splatting. ACM ComputSurv. 2026;58(12):1-39. doi: 10.1145/3807511
Hedman P, Srinivasan PP, Mildenhall B, Reiser C, Barron JT, Debevec P. Baking Neural Radiance Fields for Real-Time View Synthesis. IEEE Trans Pattern Anal Mach Intell. 2025;47(5):3310-3321. doi: 10.1109/tpami.2024.3381001
Radl L, Steiner M, Parger M, Weinrauch A, Kerbl B, Steinberger M. StopThePop: Sorted Gaussian Splatting for View-Consistent Real-time Rendering. ACM Trans Graph. 2024;43(4):1-17. doi: 10.1145/3658187
Liao Y, Di Y, Zhou H, Zhu K, Lu M, Duan Q, et al. A Survey on Neural Radiance Fields. ACM ComputSurv. 2026;58(2):1-33. doi: 10.1145/3758085
Xiao W, Chierchia R, Santa Cruz R, Li X, Ahmedt Aristizabal D, Salvado O, et al. Neural Radiance Fields for the Real World: A Survey. ACM ComputSurv. 2026;58(15):1-36. doi: 10.1145/3837758
Bao Y, Ding T, Huo J, Liu Y, Li Y, Li W, et al. 3D Gaussian Splatting: Survey, Technologies, Challenges, and Opportunities. IEEE Trans Circuits Syst Video Technol. 2025;35(7):6832-6852. doi: 10.1109/tcsvt.2025.3538684
Chen W, Li Z, Guo J, Zheng C, Tian S. Trends and Techniques in 3D Reconstruction and Rendering: A Survey with Emphasis on Gaussian Splatting. Sensors. 2025;25(12):3626. doi: 10.3390/s25123626
Wang G, Pan L, Peng S, Liu S, Xu C, Miao Y, et al. NeRFs in robotics: A survey. The International Journal of Robotics Research. 2026;45(6):858-892. doi: 10.1177/02783649251374246
Lei X, Wang M, Zhou W, Li H. GaussNav: Gaussian Splatting for Visual Navigation. IEEE Trans Pattern Anal Mach Intell. 2025;47(5):4108-4121. doi: 10.1109/tpami.2025.3538496
Müller T, Evans A, Schied C, Keller A. Instant neural graphics primitives with a multiresolution hash encoding. ACM Trans Graph. 2022;41(4):1-15. doi: 10.1145/3528223.3530127
Kerbl B, Meuleman A, Kopanas G, Wimmer M, Lanvin A, Drettakis G. A Hierarchical 3D Gaussian Representation for Real-Time Rendering of Very Large Datasets. ACM Trans Graph. 2024;43(4):1-15. doi: 10.1145/3658160
Ren K, Jiang L, Lu T, Yu M, Xu L, Ni Z, et al. Octree-GS: Towards Consistent Real-time Rendering with LOD-Structured 3D Gaussians. IEEE Trans Pattern Anal Mach Intell. 2025;PP:1-15. doi: 10.1109/tpami.2025.3568201
Barron JT, Mildenhall B, Verbin D, Srinivasan PP, Hedman P. Mip-NeRF 360: Unbounded Anti-Aliased Neural Radiance Fields. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022:5460-5469. doi: 10.1109/cvpr52688.2022.00539
Niemeyer M, Barron JT, Mildenhall B, Sajjadi MSM, Geiger A, Radwan N. RegNeRF: Regularizing Neural Radiance Fields for View Synthesis from Sparse Inputs. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022:5470-5480. doi: 10.1109/cvpr52688.2022.00540
Yuan YJ, Lai YK, Huang YH, Kobbelt L, Gao L. Neural Radiance Fields From Sparse RGB-D Images for High-Quality View Synthesis. IEEE Trans Pattern Anal Mach Intell. 2023;45(7):8713-8728. doi: 10.1109/tpami.2022.3232502
Wang S, Xu H, Li Y, Chen J, Tan G. IE-NeRF: Exploring transient mask inpainting to enhance neural radiance fields in the wild. Neurocomputing. 2025;618:129112. doi: 10.1016/j.neucom.2024.129112
Sabour S, Goli L, Kopanas G, Matthews M, Lagun D, Guibas L, et al. SpotLessSplats: Ignoring Distractors in 3D Gaussian Splatting. ACM Trans Graph. 2025;44(2):1-11. doi: 10.1145/3727143
Liu Y, Ye K, Shao T, Zhou K. TR-Gaussians: High-Fidelity Real-Time Rendering of Planar Transmission and Reflection With 3D Gaussian Splatting. IEEE Trans Visual Comput Graphics. 2026;32(7):5882-5894. doi: 10.1109/tvcg.2026.3675416
Reiser C, Szeliski R, Verbin D, Srinivasan P, Mildenhall B, Geiger A, et al. MERF: Memory-Efficient Radiance Fields for Real-time View Synthesis in Unbounded Scenes. ACM Trans Graph. 2023;42(4):1-12. doi: 10.1145/3592426
Chang K, Gan Y, Zhu W, Wang K, Chen Z, Cheng Y, et al. ScaleGS: Closing the Gap between Real-time 3D Gaussian Splatting and Real-time XR Rendering. ACM Trans Archit Code Optim. 2025;22(4):1-26. doi: 10.1145/3774425
Yu A, Ye V, Tancik M, Kanazawa A. pixelNeRF: Neural Radiance Fields from One or Few Images. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021:4576-4585. doi: 10.1109/cvpr46437.2021.00455
Nguyen-Ha P, Huynh L, Rahtu E, Matas J, Heikkilä J. Cascaded and Generalizable Neural Radiance Fields for Fast View Synthesis. IEEE Trans Pattern Anal Mach Intell. 2024;46(5):2758-2769. doi: 10.1109/tpami.2023.3335311
Jiang L, Mao Y, Xu L, Lu T, Ren K, Jin Y, et al. AnySplat: Feed-forward 3D Gaussian Splatting from Unconstrained Views. ACM Trans Graph. 2025;44(6):1-16. doi: 10.1145/3763326
Trevithick A, Chan M, Stengel M, Chan E, Liu C, Yu Z, et al. Real-Time Radiance Fields for Single-Image Portrait View Synthesis. ACM Trans Graph. 2023;42(4):1-15. doi: 10.1145/3592460
Marí R. Accelerating NeRF with the Visual Hull. Image Processing On Line. 2024;14:217-231. doi: 10.5201/ipol.2024.553
Wang C, Jiang R, Chai M, He M, Chen D, Liao J. NeRF-Art: Text-Driven Neural Radiance Fields Stylization. IEEE Trans Visual Comput Graphics. 2024;30(8):4983-4996. doi: 10.1109/tvcg.2023.3283400
Wu Q, Tan J, Xu K. PaletteNeRF: palette-based color editing for NeRFs. Communications in Information and

Systems. 2023;23(4):447-475. doi: 10.4310/cis.2023.v23.n4.a4
Zhang J, Liu X, Ye X, Zhao F, Zhang Y, Wu M, et al. Editable free-viewpoint video using a layered neural representation. ACM Trans Graph. 2021;40(4):1-18. doi: 10.1145/3450626.3459756
Gao L, Yang J, Zhang BT, Sun JM, Yuan YJ, Fu H, et al. Real-time Large-scale Deformation of Gaussian Splatting. ACM Trans Graph. 2024;43(6):1-17. doi: 10.1145/3687756
Lyu X, Sun YT, Huang YH, Wu X, Yang Z, Chen Y, et al. 3DGSR: Implicit Surface Reconstruction with 3D Gaussian Splatting. ACM Trans Graph. 2024;43(6):1-12. doi: 10.1145/3687952
Tretschk E, Tewari A, Golyanik V, Zollhofer M, Lassner C, Theobalt C. Non-Rigid Neural Radiance Fields: Reconstruction and Novel View Synthesis of a Dynamic Scene From Monocular Video. 2021 IEEE/CVF International Conference on Computer Vision (ICCV). 2021:12939-12950. doi: 10.1109/iccv48922.2021.01272
Park K, Sinha U, Hedman P, Barron JT, Bouaziz S, Goldman DB, et al. HyperNeRF. ACM Trans Graph. 2021;40(6):1-12. doi: 10.1145/3478513.3480487
Song L, Chen A, Li Z, Chen Z, Chen L, Yuan J, et al. NeRFPlayer: A Streamable Dynamic Scene Representation with Decomposed Neural Radiance Fields. IEEE Trans Visual Comput Graphics. 2023;29(5):2732-2742. doi: 10.1109/tvcg.2023.3247082
Işık M, Rünz M, Georgopoulos M, Khakhulin T, Starck J, Agapito L, et al. HumanRF: High-Fidelity Neural Radiance Fields for Humans in Motion. ACM Trans Graph. 2023;42(4):1-12. doi: 10.1145/3592415
Zhao F, Jiang Y, Yao K, Zhang J, Wang L, Dai H, et al. Human Performance Modeling and Rendering via Neural Animated Mesh. ACM Trans Graph. 2022;41(6):1-17. doi: 10.1145/3550454.3555451
Gao X, Yang J, Kim J, Peng S, Liu Z, Tong X. MPS-NeRF: Generalizable 3D Human Rendering From Multiview Images. IEEE Trans Pattern Anal Mach Intell. 2025;47(8):6110-6121. doi: 10.1109/tpami.2022.3205910
Wu G, Yi T, Fang J, Xie L, Zhang X, Wei W, et al. 4D Gaussian Splatting for Real-Time Dynamic Scene Rendering. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2024:20310-20320. doi: 10.1109/cvpr52733.2024.01920
Guo Z, Zhou W, Li L, Wang M, Li H. Motion-Aware 3D Gaussian Splatting for Efficient Dynamic Scene Reconstruction. IEEE Trans Circuits Syst Video Technol. 2025;35(4):3119-3133. doi: 10.1109/tcsvt.2024.3502257
Zheng S, Zhou B, Shao R, Liu B, Zhang S, Nie L, et al. GPS-Gaussian: Generalizable Pixel-Wise 3D Gaussian Splatting for Real-Time Human Novel View Synthesis. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2024:19680-19690. doi: 10.1109/cvpr52733.2024.01861
Zhou B, Zheng S, Tu H, Shao R, Liu B, Zhang S, et al. GPS-Gaussian+: Generalizable Pixel-wise 3D Gaussian Splatting for Real-Time Human-Scene Rendering from Sparse Views. IEEE Trans Pattern Anal Mach Intell. 2025;PP:1-16. doi: 10.1109/tpami.2025.3561248
Hubert A, Elghazaly G, Frank R. Editing implicit and explicit representations of radiance fields: a survey. Machine Vision and Applications. 2025;36(6):119. doi: 10.1007/s00138-025-01742-4
Davaasambuu T. Film Production Artificial Intelligence Use. ADR. 2025;13(03):210-220. doi: 10.4236/adr.2025.133016
Lees D. Deepfakes in documentary film production: images of deception in the representation of the real. Studies in Documentary Film. 2024;18(2):108-129. doi: 10.1080/17503280.2023.2284680
Mazzacca G, Karami A, Rigon S, Farella EM, Trybala P, Remondino F. NERF FOR HERITAGE 3D RECONSTRUCTION. Int Arch Photogramm Remote Sens Spatial Inf Sci. 2023;XLVIII-M-2-2023:1051-1058. doi: 10.5194/isprs-archives-xlviii-m-2-2023-1051-2023
Quintana D, Rodríguez A, Sbert M, Silva Y, Rufino E, Boada I. Non-photorealistic rendering as a feedback strategy in virtual reality for rehabilitation. Virtual Reality. 2024;28(1):60. doi: 10.1007/s10055-024-00954-9
Chen T, Yang Q, Chen Y. Overview of NeRF Technology and Applications. J Comput-Aided Design Comput Graphics. 2025;37(1):51-74. doi: 10.3724/sp.j.1089.2023-00759

https://doi.org/10.62486/978-9915-9966-3-9.Ch09

Chapter 10 / Capítulo 10

SAP Superintelligence Series
ISSN: 3144-5624
ISBN (complete work): 978-9915-9851-0-7

Superintelligence Series Volume 5: Intelligent Visual Arts and Audiovisual Engineering: Generative AI, Computational Creativity and Superintelligence
ISBN (this volume): 978-9915-9966-5-3
DOI: 10.62486/978-9915-9966-3-9.Ch10
Pages: 121-131

Check for updates

© 2026 The authors. This is an open access book chapter distributed under the terms of the Creative Commons Attribution (CC BY) 4.0 License.

Machine learning in post-production editing, color grading, sound, dubbing and lip synchronization

Aprendizaje automático en la posproducción de edición, etalonaje, sonido, doblaje y sincronización labial

Emanuel Maldonado[1]

1 Maquet.ar. Buenos Aires, Argentina

ABSTRACT

Post-production is the stage at which a film or video becomes coherent, and machine learning has begun to change each of its core tasks in turn. This narrative review synthesizes how machine learning has advanced five post-production tasks, editing, color grading, sound design, dubbing and lip synchronization, drawing on the literature published between 2018 and 2026. Editing has moved toward assisted organization, shot selection and transcript-driven assembly; color grading toward deep colorization, exemplar-based transfer and temporally consistent stylization; sound design toward source separation, localization and generative audio; dubbing toward voice conversion, speech translation and multilingual systems; and lip synchronization toward diffusion-based and editing-based talking-head generation. The review finds that the five tasks have advanced unevenly, with lip synchronization the most active and color grading the least documented in the scholarly literature, and that a common set of limits, temporal consistency, identity preservation, creative control and evaluation, recurs across the tasks. It concludes that machine learning has matured into a reliable assistant for post-production but has not replaced the creative decisions that the discipline still reserves for human practitioners.

Keywords: video post-production; video editing; color grading; sound design; dubbing; lip synchronization

RESUMEN

La posproducción es la etapa en la que una película o un video cobra coherencia, y el aprendizaje automático ha empezado a cambiar cada una de sus tareas centrales por turno. Esta revisión narrativa sintetiza cómo el aprendizaje automático ha hecho avanzar cinco tareas de posproducción, edición, etalonaje, diseño sonoro, doblaje y sincronización labial, a partir de la bibliografía publicada entre 2018 y 2026. La edición ha avanzado hacia la organización asistida, la selección de planos y el montaje guiado por el texto; el etalonaje hacia la colorización profunda, la transferencia por ejemplos y la estilización temporalmente consistente; el diseño sonoro hacia la separación de fuentes, la localización y el audio generativo; el doblaje hacia la conversión de voz, la traducción de habla y los sistemas multilingües; y la sincronización labial hacia la generación de cabezas parlantes basada en difusión y en edición. La revisión constata que las cinco tareas han avanzado de forma desigual, con la sincronización labial como la más activa y el etalonaje como el menos documentado en la bibliografía académica, y que un conjunto común de límites, consistencia temporal, preservación de la identidad, control creativo y evaluación, se repite entre las tareas. Concluye que el aprendizaje automático ha madurado como un asistente fiable de la posproducción, pero no ha sustituido las decisiones creativas que la disciplina sigue reservando a los profesionales humanos.

Palabras clave: posproducción de video; edición de video; etalonaje; diseño sonoro; doblaje; sincronización labial

INTRODUCTION

Post-production is the stage of audiovisual work that transforms raw material into a finished piece, and it is the stage at which the craft of film and video is most clearly an act of judgment rather than of capture. Editing assembles shots into a rhythm, color grading sets the image's tone and continuity, sound design builds the world the audience hears, dubbing carries the dialogue across languages, and lip synchronization binds a voice to a face. Each of these tasks has historically been performed by a specialist whose expertise is tacit, slow to acquire and hard to automate, which is why the arrival of machine learning in post-production has been watched with a mixture of anticipation and resistance across the industry.[1,2]

Machine learning has reached the five tasks at different speeds and through different routes, and that unevenness is itself the first fact this review records, a fact that a purely technical account would miss but that matters to anyone who must decide where to invest, what to adopt and what still requires a specialist's hand. Editing was among the first to be touched, because shot selection and assembly can be framed as ranking and retrieval problems, and the literature on automatic and assisted editing now spans text-driven, summary-driven and shot-classification approaches. (3-5) Color grading arrived through a different door, the colorization and style-transfer methods of computer vision, which give an artist the ability to propagate a grade across frames rather than paint it frame by frame.[6,7] Sound design was transformed by the deep-learning revolution in audio signal processing, from source separation to generative synthesis.[2,8] Dubbing draws on the neighboring fields of voice conversion and speech translation, and lip synchronization has become the most active of the five, propelled by generative and diffusion models.[9,10] The five tasks are thus five different stories, and the review tells each in turn before asking what they share, a structure that respects the unevenness rather than flattening it into a single, falsely uniform account.

The significance of the topic is not only technical but economic and professional, because post-production is where the cost, the schedule and the creative control of a production converge. A method that automates a repetitive part of post-production changes who does the work and how it is credited, and the literature that accompanies the technology has already begun to register those consequences, from the voice actors whose likeness can now be cloned to the editors whose judgment a benchmark now attempts to measure.[1,11] The review therefore treats post-production not as a neutral pipeline but as a set of practices with stakes, and it reads the machine-learning literature against those stakes.

The objective of this review is to synthesize how machine learning has advanced the five post-production tasks of editing, color grading, sound design, dubbing and lip synchronization. Three specific objectives organize the account: to synthesize the methods applied to each task, to report the capabilities and limits the literature claims for them, and to identify the cross-cutting gaps that separate the current state of the art from routine professional use.

DEVELOPMENT

SYNTHESIS

Table 1 provides a comparative overview of the five tasks, highlighting the principal methodological approaches and representative findings discussed in the review.

Table 1. Machine-learning methods and representative findings across the five post-production tasks.

Task	Representative methods	Representative finding
Editing	Text-based editing, summarization, shot classification	Transcript-driven and summary-driven assembly maturing; benchmarks lag visual effects
Color grading	Deep colorization, exemplar transfer, interactive stylization	Temporally consistent colorization and artist-guided propagation
Sound design	Source separation and localization, generative audio	Full audio chain now covered, from separation to text-to-audio generation
Dubbing	Voice conversion, speech translation	Multilingual systems maturing; emotion and oversight remain open
Lip synchronization	Diffusion and editing-based talking-head generation	Most active axis; identity and naturalness remain open

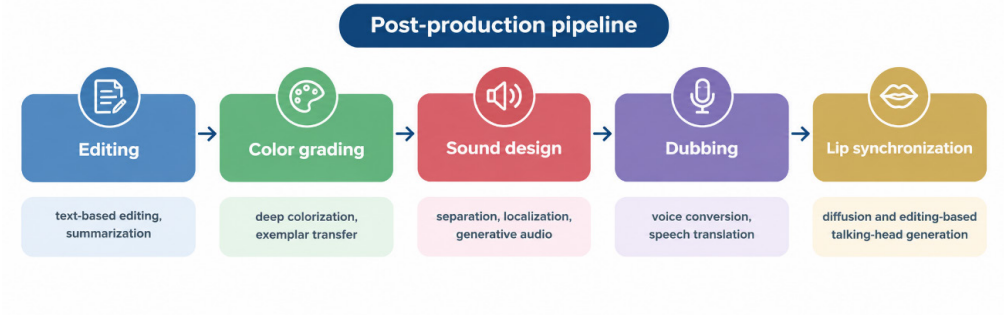
https://doi.org/10.62486/978-9915-9966-3-9.Ch10

Table 2 identifies common research gaps and challenges across these areas, while Figure 1 illustrates their integration into the post-production workflow.

Table 2. Cross-cutting gaps across the five tasks and their implications.

Gap	Where it appears	Implication
Temporal consistency	Color, lip synchronization	Flicker and drift between frames limit broadcast use
Identity and naturalness	Dubbing, lip synchronization	Voice and face must survive editing without artifacts
Creative control	Editing, color grading, sound	Automation must remain directable by an artist
Evaluation	All five tasks	Benchmarks lag the visual-effects tasks and human judgments

Figure 1. The five post-production tasks and the machine-learning capabilities mapped onto them.



Given the diversity of evaluation datasets, performance metrics, and experimental conditions, the findings are synthesized narratively without statistical pooling. Quantitative results are presented in relation to their original studies to preserve methodological context and avoid misleading comparisons.

EDITING

Editing is the task that decides what the audience sees and in what order, and it is the task whose automation raises the most direct question about creative authorship, since the cut is the signature of the editor's judgment. Machine learning has addressed it through three complementary framings. The first is transcript-driven editing, in which the cut is specified by editing the text of the dialogue rather than the timeline, a capability demonstrated on talking-head video and one that maps naturally onto the interview and documentary work that dominates much of post-production.[3] The second is summarization, in which a model selects the shots that carry a long recording's content, a problem the literature frames as a ranking and selection task over candidate shots.[4] The third is shot classification, in which a model labels the shots, such as field-sport footage by play type, so that an editor or an automated system can find the material it needs.[5] The three framings share a common shape, that editing is reduced from a continuous act of judgment to a set of discrete selection decisions, which is what makes it tractable to learning but also what defines the limit of the automation.

The survey literature describes the state of the field as one in which semi-automatic and fully automatic solutions have reduced the workload of editing but have not absorbed its judgment, a state the surveys present as an accomplishment rather than as a failure of ambition. The solutions that work best are those that assist a human editor, organizing footage, proposing summaries and assembling rough cuts, rather than those that aim to replace the editor entirely, and the surveys are candid that the complex, context-dependent guidelines a skilled editor follows are the part that learning has not captured.(1) The finding that recurs is that the bottleneck is not the learning method but the definition of the task, since an edit is a creative decision whose quality is judged against a director's intent, and that intent is not yet in the training signal. The axis therefore ends where the others do, with automation that is genuinely useful and genuinely partial.

Two adjacent capabilities extend the editing axis beyond the cut itself. Video description, the automatic captioning of what a shot contains, gives an editor and an automated system the index

https://doi.org/10.62486/978-9915-9966-3-9.Ch10

by which material is found, and the curation of sports highlights shows the same description-and-ranking machinery turned toward a finished product, a reel assembled by selecting the moments a multimodal model scores as exciting.[12,13] Procedural content generation contributes the underlying idea that content can be produced by a learned model against a set of constraints, which, though developed for games, frames the more ambitious question of whether an edit itself can be generated rather than merely assisted.(14) These adjacent capabilities do not change the boundary the axis draws, but they widen the set of tasks that sit on the automated side of it.

The axis is also where the question of evaluation is most acute, because an edit has no natural ground truth, and the same footage admits an indefinite number of defensible cuts. The datasets and benchmarks that have been proposed for AI-assisted editing accordingly frame the problem as assistance rather than as the prediction of a single correct edit, measuring whether a system organizes, selects and assembles in ways a human editor finds useful rather than whether it reproduces a reference cut.(1) This is a more honest formulation than the error metrics used elsewhere in the review, and it is one of the places where the editing literature, by accepting that its object is a judgment, has come closest to the boundary the review keeps returning to. The absence of a shared benchmark for that judgment, however, means the axis's progress is harder to compare than its ambition deserves.

COLOR GRADING

Color grading arrived in the machine-learning literature through colorization and style transfer, and the axis is best understood as the application of those methods to a professional practice, an application that has succeeded on the underlying technique while lagging on the professional integration. Deep exemplar-based colorization showed that a network could transfer color from a reference image to a grayscale target, which is the technical core of a grade, and the review literature traces the line from early analogy-based methods to the deep models that now dominate.[6,15] The step that matters for post-production is temporal consistency, since a grade that flickers between frames is unusable, and the video colorization methods that propagate a deep feature representation across frames and regularize themselves address exactly that problem.(7) Interactive stylization completes the picture by giving the artist a sparse set of patches from which the network learns and then propagates a style across the footage, which is the closest the axis comes to the artist-in-the-loop workflow that grading actually uses.(16)

The color-grading axis is nevertheless the thinnest of the five in the scholarly literature, and that thinness is itself a finding. The work that exists is strong on the underlying methods, colorization, transfer and consistency, but the studies that connect those methods to the professional practice of grading, its tonal decisions, its continuity requirements and its creative intent, are comparatively few and largely qualitative.[17,18] The analyses that do address grading as a practice describe the same tension from the practitioner's side, the automated tools that change the workflows of photography and cinema while the authorial intent behind the image must still be protected, and the studies of color management in new-media films describe a practice whose aesthetic goals the automated methods do not yet encode.[19,20] The axis therefore records a gap between the maturity of the technique and the maturity of its integration, a gap that the discussion returns to, and it is the clearest case in the review of a task whose academic treatment has not kept pace with its professional importance.

SOUND DESIGN

Sound design is the axis on which deep learning has made the most sweeping claims, because the audio signal-processing literature was transformed by learning before the other tasks were touched. The foundational survey of the period describes a field in which deep networks had come to cover the audio chain, from classification and detection to enhancement and separation, and the specific surveys of source localization and of audio-visual enhancement and separation document the same coverage in more detail.[2,21,22] For post-production, the immediate consequence is the ability to separate a dialogue track from music and effects, to localize a sound in a scene, and to enhance speech against noise, which are among the most time-consuming parts of the sound editor's work.(22) The audio-visual form of these methods is the one that matters most to post-production, because it binds the sound to the picture, enhancing speech by reference to the lips that produced it and separating a source by reference to where it is on

https://doi.org/10.62486/978-9915-9966-3-9.Ch10

screen, a coupling that has no counterpart in the earlier single-modality methods.[22,23]

The generative turn is the axis's second and more recent story, and it is the turn that has changed what sound design means as a task. Language-model and self-supervised approaches generate coherent audio and music from text or from a conditioning signal, and the surveys of symbolic music generation document the same capability for score.(8,24,25) The practical promise for post-production is direct, a text prompt that yields an ambient bed, a music cue or a sound effect, and the limitation the literature reports is equally direct, that generation is not yet authoring, and the coherent audio a model produces still has to be shaped, placed and mixed by a sound designer. The axis therefore reaches further than the others, covering the full chain from separation to generation, while reproducing the same boundary between the automated and the authored.

Two further capabilities round out the sound-design picture and show its breadth. Binaural generation synthesizes the spatial audio that positions a sound in a scene, a task that was once the province of specialized recording and is now approached as a learning problem, and cross-modal retrieval between audio and lyrics supplies the ability to find a cue by its content rather than its label.[26,27] The reinforcement-learning survey documents the same spread from the control side, showing that audio-based systems can now be trained to make sequential decisions, which is the direction in which an automated mix or an automated placement of sound would plausibly go.(28) These capabilities do not close the gap between generation and authoring, but they show that the sound axis has pushed automation further across the chain than the other axes have.

DUBBING

Dubbing is the task that carries a performance across a language boundary, and machine learning has approached it through voice conversion and speech translation, the two capabilities that together allow a speaker's words to be re-voiced in another language while retaining the speaker's own voice. The overview of voice conversion traces the field from statistical models to deep networks, establishing the technical core, that a source voice can be transformed into a target voice while preserving the linguistic content, and the emotional-voice-conversion work extends that core to the expressive dimensions that dubbing actually requires.[9,29] The multilingual dubbing review assembles these pieces into the system view, describing the pipeline from transcription through translation and synthesis to the re-voicing of a video, and it is candid about the current limits of the pipeline as a whole.(30)

The axis is distinguished by the salience of its human and institutional questions, which the technical literature registers more than the other axes do. The studies of automatic dubbing for educational content identify the practical challenges, terminology, timing and speaker consistency, that a deployed system must solve, and the analyses of the role of language specialists argue that the pipeline still requires human oversight at points the technology cannot yet automate.[31,32] Alongside these, the work on voice cloning and on the situation of voice actors makes explicit what the technical papers leave implicit, that dubbing is not only a signal-processing problem but a labor and identity problem, and that a cloned voice is a person's instrument and their livelihood.[11,33] The axis therefore couples a maturing technical core to an unresolved set of human questions, and that coupling is its defining feature.

The technical core itself is broader than the conversion-and-translation framing alone, and its breadth matters for how the axis is read, since a narrow reading would mistake a maturing pipeline for a solved one. Noise-robust voice conversion addresses the real-world condition in which the source audio is not clean, which is the condition of most production material, and the multilingual dubbing systems that assemble transcription, translation and synthesis into a single pipeline show that the pieces are being integrated into deployable tools rather than remaining separate demonstrations.[34,35] The translation scholarship adds the constraint that the technology alone cannot supply, that dubbing must respect the audiovisual constraints of timing and the pragmatics of politeness that a literal translation would break, which is precisely the judgment a language specialist contributes.[36] The axis is therefore not only a maturing technical core coupled to human questions, but a core whose remaining failures are located exactly where the human questions are.

https://doi.org/10.62486/978-9915-9966-3-9.Ch10

LIP SYNCHRONIZATION

Lip synchronization is the most active of the five axes, and its activity is the story the axis tells, a story of a field that has moved through several paradigms within the space of the review window. The surveys of audio-driven facial animation and of audio synthesis with lip-syncing map a field that has moved quickly through several paradigms, and the reviews of audio-driven talking-face generation organize the same material around the problem of producing a face whose mouth matches a voice while the rest of the face stays natural.(10,37,38) The editing-based approach reframes synchronization as a post-production task in the narrowest sense, taking an existing talking head and modifying its expression, lips and realism in sequence so that a new audio track can be laid onto an old performance.(39) The perceptual and generative foundations come from the work on speech-lip synchronization and on audio-driven three-dimensional facial animation, which supply the constraints and the realism the axis is judged on.[40,41]

The axis's open problems are the most precisely stated in the review, and they are the same problems that appear, in different forms, across the other four axes, which is what makes the axis a useful summary of the review as a whole. Identity must survive the edit, so that the talking head remains recognizably the same person, and naturalness must survive it too, so that the mouth moves like speech rather than like an animation, and the literature reports both as active targets rather than solved.[38,39] The diffusion and generative methods that now dominate the axis have raised its quality while raising its cost and its training complexity, and the axis ends, like the others, with a technique that is impressive in isolation and still being integrated into a workflow.

The foundations beneath the axis are more extensive than the headline methods suggest, and they explain why the axis has moved so fast relative to the other four. Visual speech analysis, the broader field of reading speech from faces, supplies the perceptual understanding of how lips, jaw and phonemes correspond that the generative methods exploit, and the single-image talking-face methods show that the capability now extends to animating a face from a single photograph rather than a video.[42,43] High-resolution systems target the specific defect that had held the axis back, that a synchronized mouth at low resolution is not yet a usable deliverable, and their work is the clearest sign that the axis is being pushed toward production rather than left at demonstration quality.(44) The axis is therefore not only the most active but the most clearly aimed at the standard that post-production actually requires.

DISCUSSION

The five axes, read together rather than one at a time, describe a single uneven advance rather than five separate ones. Lip synchronization has moved furthest and fastest, propelled by the generative models that arrived in the window, while color grading, the oldest of the tasks, has the thinnest scholarly treatment despite mature underlying methods, and the other three occupy the middle ground, each with a maturing technical core and an unresolved boundary between the automated and the authored.(2,6,10) The unevenness is explained less by the difficulty of the tasks than by where the neighboring literatures happened to be, since lip synchronization inherited the momentum of generative vision, and color grading had no equivalent neighboring field pushing it forward, a point the review offers as the key to the pattern.[15,38]

The most consistent finding across the axes is the boundary between what the methods automate and what they leave to the practitioner, and it is the same boundary in every case, whether the task is cutting, grading, mixing, re-voicing or synchronizing a face. Editing automates selection but not judgment, color grading propagates a grade but does not choose it, sound generation produces material but does not place or mix it, dubbing converts a voice but does not yet carry its expression, and lip synchronization moves a mouth but must preserve the identity and the naturalness that a performance requires.(1,8,16,29,39) The review reads this boundary not as a failure of the methods but as a stable description of what post-production is, a discipline in which the creative decision is the product and the repetitive labor is the cost, and in which machine learning has so far reduced the cost without supplying the product. This is the single conclusion that the five axes share.

Temporal consistency is the cross-cutting technical gap, and it is worth stating precisely because it appears under different names in each axis. In color grading it is the flicker of a grade between frames, in lip synchronization it is the drift of a mouth or a face across a sequence, and in each case the literature treats it as an active problem rather than a solved one, addressed by regularization, propagation and explicit temporal modeling but not yet eliminated.[7,38] Identity

https://doi.org/10.62486/978-9915-9966-3-9.Ch10

and naturalness are the human face of the same gap, since a method that fails to preserve a face or a voice across frames fails the first requirement of a post-production tool, that it must not break the performance it is asked to improve.[29,40] The recurrence of the same gap across tasks is the strongest evidence the review has that it is a property of the problem rather than of any single method.

Creative control is the gap that separates the technical literature from the professional practice, and it is the one the review most wants to foreground. The methods are increasingly powerful, but a powerful method that cannot be directed is a demonstration rather than a tool, and the studies that do connect the methods to practice, the interactive stylization, the language-specialist workflow, the editor-assisting benchmark, are the exceptions rather than the rule.(1,16,32) The production-oriented literature that sits alongside the technical work makes the same point from the practitioner's side, that automation is useful precisely to the extent that it remains under the control of the person whose intent the work is meant to realize.[11,45] The review therefore identifies creative control, not raw capability, as the property that will decide whether these methods enter routine use.

The evaluation gap is the last of the cross-cutting findings, and it conditions all the others. The tasks are judged on heterogeneous benchmarks, and the benchmarks lag the tasks, since they measure the visual-effects problems that are easy to score, such as super-resolution and colorization error, rather than the editing, grading and sound decisions that are hard to score, such as whether a cut is right or a grade serves a scene.[1,46] The consequence is that the review can report capabilities and limits but rarely magnitudes, and the field cannot yet say, in a comparable way, how much better one method is than another on the tasks that actually matter. A benchmark for post-production judgment, against human editors and designers, is the single most useful instrument the field lacks.

The professional implications of the advance follow from the boundary the review has drawn, and they should be stated without either alarmism or complacency. Where a task is repetitive, machine learning changes who does it and how fast, and the literature already registers the consequence for voice actors, whose instrument can now be cloned, and for editors and colorists, whose judgment is now partly encoded in tools that did not exist when the window opened.[11,19] The defensible reading is not that the craft disappears but that it moves, from the repetitive labor that the methods have absorbed toward the supervisory, directorial and curatorial decisions that they have not, a movement that the practitioner-oriented literature describes from inside the workflows rather than against them.[1,32] A discipline that redefines its center of gravity in this way is not being replaced; it is being reorganized, and the reorganization is the practical stake of the technical advance this review documents.

The window itself explains why the five tasks advanced so unevenly, and the explanation is the arrival of generative models in its middle. The diffusion and language-model methods that now dominate lip synchronization and sound generation were not available at the start of the period, and the tasks that could be reached through the earlier computer-vision methods, editing and color, were the ones that matured first and have since been overtaken by the tasks the generative models unlocked.[8,10] The trajectory is therefore a case of the technology pulling the field, rather than the field pulling the technology, and a review that dates the advance to the middle of the window is dating it to a specific technical event whose consequences are still working through the five tasks at the window's end. This is the sense in which the review is, among other things, a record of what a single class of models did to a whole discipline in a few years.

Read against the larger advance of artificial intelligence, post-production is a microcosm of the pattern that recurs wherever a learned model meets a skilled practice. The model absorbs the parts of the practice that can be specified, the selection, the transfer, the separation, the conversion, the synchronization, and it stalls at the parts that cannot, the judgment, the intent, the expression, the control, and the history of the five tasks over the window is the history of that boundary moving outward without disappearing.[2,39] The implication for the study of advanced systems is that the decisive question is not capability but directability, since a system that produces without being directed is a demonstration, and a system that can be directed is a colleague. Post-production, where the directed human decision is the entire point of the work, is therefore the clearest available test of where that boundary actually sits, and the review's five axes are five measurements of it.

The agenda that follows is specific to the gaps the review found. First, temporal consistency and identity preservation are the highest-value technical targets, because they are the properties that separate a research demonstration from a broadcast deliverable, and they recur across color and lip synchronization.[7,38] Second, the field needs the artist-in-the-loop to be a design principle

https://doi.org/10.62486/978-9915-9966-3-9.Ch10

rather than an afterthought, so that editing, grading and sound methods expose the controls a practitioner needs rather than a black-box output.[16,32] Third, evaluation must move from visual-effects metrics toward judgment, with benchmarks built against human editors and designers, since the current metrics cannot capture the decisions that post-production is actually about.(1) A fourth item follows from the review's own method, that a cross-task and cross-lingual synthesis would test whether the five stories the review told separately are in fact one story told five times. These four items are the routes by which the five axes would converge from a set of impressive techniques into an integrated, directable post-production practice.

The review has limitations that follow from its design, and they are stated here so that the reader can weigh the synthesis accordingly. The search was restricted to three bibliographic sources and two languages, so the synthesis favors the mainstream English-language computer-vision and audio literature and may under-represent work in industry venues and in other languages, and the selection by a single reviewer introduces a selection bias that a dual review would reduce. The thematic grouping forced each study into a single task, where some methods span tasks, such as the audio-visual methods that serve both sound and lip synchronization, and the assignment is therefore a simplification.[22,23] The synthesis is narrative by necessity, because the tasks are evaluated on heterogeneous metrics, and no quantitative comparison across tasks was attempted. Each limitation is stated, and each suggests a concrete extension: a broader multilingual search, a cross-task coding of methods, and a systematic benchmark synthesis.

The agenda that follows is specific to the gaps the review found. First, temporal consistency and identity preservation are the highest-value technical targets, because they are the properties that separate a research demonstration from a broadcast deliverable, and they recur across color and lip synchronization.[7,38] Second, the field needs the artist-in-the-loop to be a design principle rather than an afterthought, so that editing, grading and sound methods expose the controls a practitioner needs rather than a black-box output.[16,32] Third, evaluation must move from visual-effects metrics toward judgment, with benchmarks built against human editors and designers, since the current metrics cannot capture the decisions that post-production is actually about.(1) These three items are the routes by which the five axes would converge from a set of impressive techniques into an integrated, directable post-production practice.

CONCLUSIONS

Machine learning has advanced the five core post-production tasks of editing, color grading, sound design, dubbing and lip synchronization, but it has advanced them unevenly and left the same boundary in each. Editing gained assisted organization and transcript-driven assembly, color grading gained deep colorization and artist-guided propagation, sound gained source separation and generative audio, dubbing gained voice conversion and multilingual systems, and lip synchronization gained diffusion-based and editing-based talking-head generation. The advance is real and the boundary is consistent, since each task automates the repetitive labor while leaving the creative decision to the practitioner. The cross-cutting gaps, temporal consistency, identity and naturalness, creative control and evaluation, recur across the tasks and mark the distance between the current state of the art and routine professional use. The review concludes that machine learning has matured into a reliable assistant for post-production, and that its entry into routine practice depends less on further capability than on making that capability directable, consistent and measurable against the judgments it is meant to serve. The five tasks, read together, describe a single advance that has reduced the repetitive labor of post-production while leaving its creative decisions intact, and the discipline's next step is not a more powerful model but a more directable one, whose temporal consistency, creative control and evaluation are as developed as its generation.

https://doi.org/10.62486/978-9915-9966-3-9.Ch10

REFERENCES

Zhang X, Li Y, Han Y, Wen J. AI Video Editing: a Survey [Internet]. MDPI AG; 2022. Available from:
Purwins H, Li B, Virtanen T, Schluter J, Chang SY, Sainath T. Deep Learning for Audio Signal Processing. IEEE J Sel Top Signal Process. 2019;13(2):206-219. doi: 10.1109/jstsp.2019.2908700
Fried O, Tewari A, Zollhöfer M, Finkelstein A, Shechtman E, Goldman DB, et al. Text-based editing of talking-head video. ACM Trans Graph. 2019;38(4):1-14. doi: 10.1145/3306346.3323028
Saini P, Kumar K, Kashid S, Saini A, Negi A. Video summarization using deep learning techniques: a detailed analysis and investigation. Artif Intell Rev. 2023;56(11):12347-12385. doi: 10.1007/s10462-023-10444-0
Minhas RA, Javed A, Irtaza A, Mahmood MT, Joo YB. Shot Classification of Field Sports Videos Using AlexNet Convolutional Neural Network. Applied Sciences. 2019;9(3):483. doi: 10.3390/app9030483
He M, Chen D, Liao J, Sander PV, Yuan L. Deep exemplar-based colorization. ACM Trans Graph. 2018;37(4):1-16. doi: 10.1145/3197517.3201365
Liu Y, Zhao H, Chan KCK, Wang X, Loy CC, Qiao Y, et al. Temporally consistent video colorization with deep feature propagation and self-regularization learning. Comp Visual Med. 2024;10(2):375-395. doi: 10.1007/s41095-023-0342-8
Borsos Z, Marinier R, Vincent D, Kharitonov E, Pietquin O, Sharifi M, et al. AudioLM: A Language Modeling Approach to Audio Generation. IEEE/ACM Trans Audio Speech Lang Process. 2023;31:2523-2533. doi: 10.1109/taslp.2023.3288409
Sisman B, Yamagishi J, King S, Li H. An Overview of Voice Conversion and Its Challenges: From Statistical Modeling to Deep Learning. IEEE/ACM Trans Audio Speech Lang Process. 2021;29:132-157. doi: 10.1109/taslp.2020.3038524
Jiang D, Chang J, You L, Bian S, Kosk R, Maguire G. Audio-Driven Facial Animation with Deep Learning: A Survey. Information. 2024;15(11):675. doi: 10.3390/info15110675
Paul A. The Voice Actor and Their Double: Working as a voice actor and teaching voice acting in the age of AI voice cloning. Tradition Innovations in Arts, Design, and Media Higher Education. 2024;1(1). doi: 10.9741/2996-4873.1017
Aafaq N, Mian A, Liu W, Gilani SZ, Shah M. Video Description. ACM ComputSurv. 2020;52(6):1-37. doi: 10.1145/3355390
Merler M, Mac KNC, Joshi D, Nguyen QB, Hammer S, Kent J, et al. Automatic Curation of Sports Highlights Using Multimodal Excitement Features. IEEE Trans Multimedia. 2019;21(5):1147-1160. doi: 10.1109/tmm.2018.2876046
Liu J, Snodgrass S, Khalifa A, Risi S, Yannakakis GN, Togelius J. Deep learning for procedural content generation. Neural Comput & Applic. 2021;33(1):19-37. doi: 10.1007/s00521-020-05383-8
Chen SY, Zhang JQ, Zhao YY, Rosin PL, Lai YK, Gao L. A review of image and video colorization: From analogies to deep learning. Visual Informatics. 2022;6(3):51-68. doi: 10.1016/j.visinf.2022.05.003
Texler O, Futschik D, kučera M, jamriška O, Sochorová Š, Chai M, et al. Interactive video stylization using few-shot patch-based training. ACM Trans Graph. 2020;39(4). doi: 10.1145/3386569.3392453
Wang X, Wu Y. An Analysis of the Precise Application of AI in Color Management of Film and Television. Research on Literary and Art Development. 2025;6(2):68-71. doi: 10.47297/wsprolaadwsp2634-786514.20250602
Zibzibadze M. AI-Driven Color Correction Workflows and the Control of Creative Image Integrity. ULAHU. 2025;02(04):11-16. doi: 10.70315/uloap.ulahu.2025.0204002
Chigozie Emmanuel Benson, Chinelo Harriet Okolo, Olatunji Oke. Automating Media Production Workflows: The Role of AI in Streamlining Post-Production, Editing, and Distribution. Int J Sci Res Civil Engg. 2024;8(5):168-176. doi: 10.32628/ijsrce248514
Ge W. Research on Color Management and Color Grading Applications in New Media Movies. EHSS. 2024;41:1-6. doi: 10.54097/yrgecs28
Grumiaux PA, Kitić S, Girin L, Guérin A. A survey of sound source localization with deep learning methods. The Journal of the Acoustical Society of America. 2022;152(1):107-151. doi: 10.1121/10.0011809
Michelsanti D, Tan ZH, Zhang SX, Xu Y, Yu M, Yu D, et al. An Overview of Deep-Learning-Based Audio-Visual Speech Enhancement and Separation. IEEE/ACM Trans Audio Speech Lang Process. 2021;29:1368-1396. doi: 10.1109/taslp.2021.3066303
Zhu H, Luo MD, Wang R, Zheng AH, He R. Deep Audio-visual Learning: A Survey. Int J Autom Comput. 2021;18(3):351-376. doi: 10.1007/s11633-021-1293-0
Liu H, Yuan Y, Liu X, Mei X, Kong Q, Tian Q, et al. AudioLDM 2: Learning Holistic Audio Generation With Self-Supervised Pretraining. IEEE/ACM Trans Audio Speech Lang Process. 2024;32:2871-2883. doi: 10.1109/taslp.2024.3399607
Ji S, Yang X, Luo J. A Survey on Deep Learning for Symbolic Music Generation: Representations, Algorithms, Evaluations, and Challenges. ACM ComputSurv. 2024;56(1):1-39. doi: 10.1145/3597493
Li S, Liu S, Manocha D. Binaural audio generation via multi-task learning. ACM Trans Graph. 2021;40(6):1-13. doi: 10.1145/3478513.3480560
Yu Y, Tang S, Raposo F, Chen L. Deep Cross-Modal Correlation Learning for Audio and Lyrics in Music Retrieval. ACM Trans Multimedia Comput Commun Appl. 2019;15(1):1-16. doi: 10.1145/3281746
Latif S, Cuayáhuitl H, Pervez F, Shamshad F, Ali HS, Cambria E. A survey on deep reinforcement learning for audio-based applications. Artif Intell Rev. 2023;56(3):2193-2240. doi: 10.1007/s10462-022-10224-2
Zhou K, Sisman B, Rana R, Schuller BW, Li H. Emotion Intensity and its Control for Emotional Voice Conversion. IEEE Trans Affective Comput. 2023;14(1):31-48. doi: 10.1109/taffc.2022.3175578
Bigioi D, Corcoran P. Multilingual video dubbing—a technology review and current challenges. Front Signal

Process. 2023;3:1230755. doi: 10.3389/frsip.2023.1230755
Baños R. Key challenges in using automatic dubbing to translate educational YouTube videos. LANS-TTS. 2023;22. doi: 10.52034/lans-tts.v22i.763
Spiteri Miggiani G. The role of language specialists in AI-dubbing workflows. DT. 2026;13(2):154-180. doi: 10.1075/dt.25024.spi
Pérez A, Díaz-Munío GG, Giménez A, Silvestre-Cerdà JA, Sanchis A, Civera J, et al. Towards cross-lingual voice cloning in higher education. Engineering Applications of Artificial Intelligence. 2021;105:104413. doi: 10.1016/j.engappai.2021.104413
Du H, Xie L, Li H. Noise-robust voice conversion with domain adversarial training. Neural Networks. 2022;148:74-84. doi: 10.1016/j.neunet.2022.01.003
Kannojia R, Singh AK, Sharma I, Gupta S. Gen AI driven multilingual audio dubbing and synthesis system for cross-language video platforms. Results in Engineering. 2025;27:106241. doi: 10.1016/j.rineng.2025.106241
Napoli V. Speech Act (Im)Politeness and Audiovisual Constraints in Translation for Dubbing. JAT. 2020;3(2):29-46. doi: 10.47476/jat.v3i2.2020.119
Kadam A, Rane S, Mishra A, Sahu S, Singh S, Pathak S. A Survey of Audio Synthesis and Lip-syncing for Synthetic Video Generation. EAI Endorsed Transactions on Creative Technologies. 2021;8(28):169187. doi: 10.4108/eai.14-4-2021.169187
Liu S. Audio-Driven Talking Face Generation: A Review. J Audio Eng Soc. 2023;71(7/8):408-419. doi: 10.17743/jaes.2022.0081
Cheng K, Cun X, Zhang Y, Xia M, Yin F, Zhu M, et al. VideoReTalking: Audio-based Lip Synchronization for Talking Head Video Editing In the Wild. SIGGRAPH Asia 2022 Conference Papers. 2022:1-9. doi: 10.1145/3550469.3555399
Pan Z, Tao R, Xu C, Li H. Selective Listening by Synchronizing Speech With Lips. IEEE/ACM Trans Audio Speech Lang Process. 2022;30:1650-1664. doi: 10.1109/taslp.2022.3153258
Pan Y, Liu C, Xu S, Tan S, Yang J. VASA-Rig: Audio-Driven 3D Facial Animation with 'Live' Mood Dynamics in Virtual Reality. IEEE Trans Visual Comput Graphics. 2025;31(5):2416-2425. doi: 10.1109/tvcg.2025.3549168
Sheng C, Kuang G, Bai L, Hou C, Guo Y, Xu X, et al. Deep Learning for Visual Speech Analysis: A Survey. IEEE Trans Pattern Anal Mach Intell. 2024;46(9):6001-6022. doi: 10.1109/tpami.2024.3376710
Li Y, Shen X. Audio-driven single image talking face animation with transformers. Sci Rep. 2026;16(1):3796. doi: 10.1038/s41598-025-33928-1
G D M. Wav2Lip-HQ High-Resolution Audio-Driven Lip Synchronization for Realistic Virtual Avatars. IJRASET. 2025;13(7):388-393. doi: 10.22214/ijraset.2025.72934
Oluwaseun Abiola Ajiva, Onyinye Gift Ejike, Angela Omozele Abhulimen. Innovative approaches in high-end photo retouching and color grading techniques for enhanced marketing and visual storytelling, including for SMEs. Int J Front Sci Technol Res. 2024;7(1):057-065. doi: 10.53294/ijfstr.2024.7.1.0048
Liu H, Ruan Z, Zhao P, Dong C, Shang F, Liu Y, et al. Video super-resolution based on deep learning: a comprehensive survey. Artif Intell Rev. 2022;55(8):5981-6035. doi: 10.1007/s10462-022-10147-y

https://doi.org/10.62486/978-9915-9966-3-9.Ch10