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

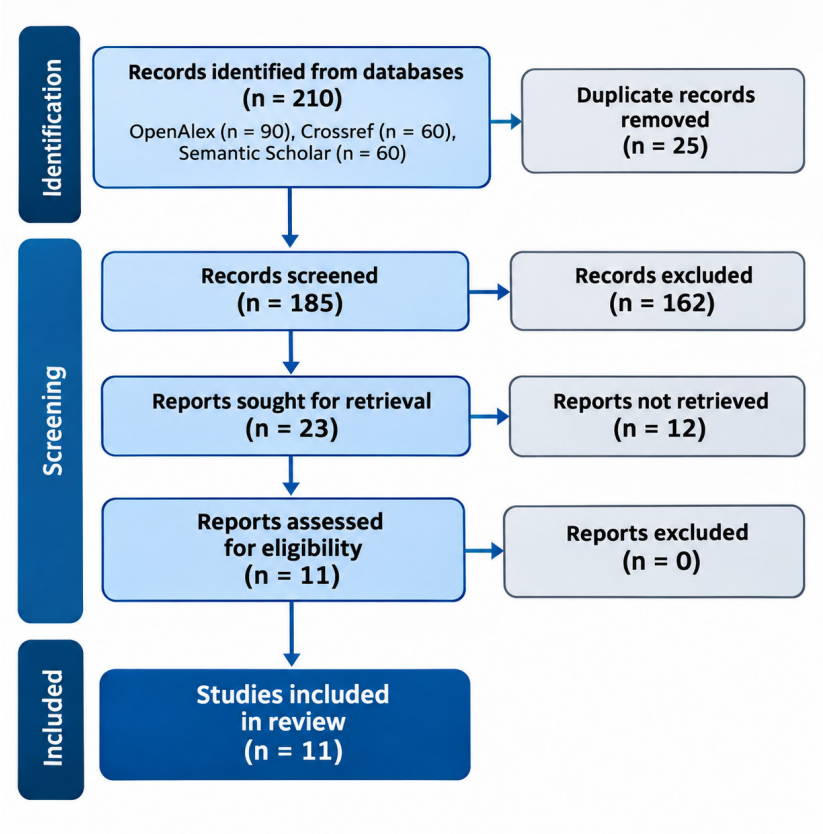
Risk of bias in the conventional sense, of a study's estimate of an effect, does not apply to a review of reporting practices, and no tool such as ROBINS-I or RoB 2 was therefore used. What plays the analogous role here is the completeness of reporting, and this was assessed uniformly rather than imputed: 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