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

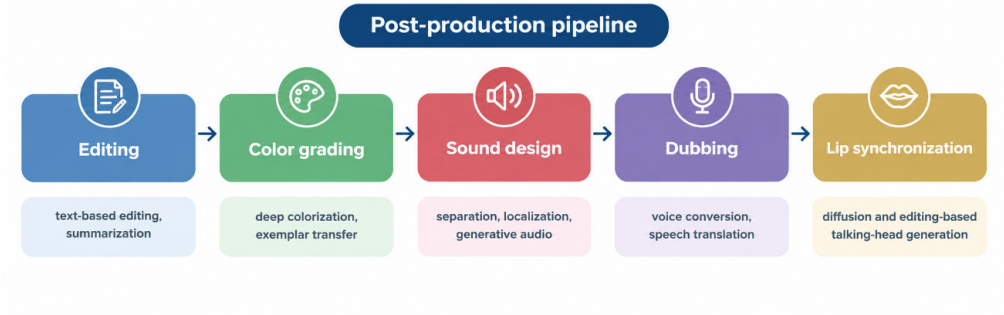
https://doi.org/10.62486/978-9915-9966-3-9.Ch10

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