

Systematic Mapping Study on the Use of Generative Artificial Intelligence in Non-Interactive Audiovisual: Applications, Challenges, and Impacts

Guilherme de Queiroz Lima
Roth
Universidade Federal do Paraná
Curitiba, Paraná, Brazil
guilherme.roth@ufpr.br

Rafael de Andrade Pereira
Universidade Federal do Paraná
Curitiba, Paraná, Brazil
rafaelpontoandrade@gmail.com

Natasha Malveira Costa
Valentim
Universidade Federal do Paraná
Curitiba, Paraná, Brazil
natasha@inf.ufpr.br

ABSTRACT

Generative Artificial Intelligence (GAI) has transformed non-interactive audiovisual production, impacting creative workflows and media generation. This study maps the scientific literature on GAI in audiovisual production (2022-2025), identifying applications, challenges, and impacts. A Systematic Mapping Study was conducted using IEEE Xplore and ACM Digital Library, following PICOC and established SMS guidelines. The results highlight dominant technologies, application areas, evaluation practices, and ethical concerns, as well as methodological gaps and future research directions.

KEYWORDS

Generative Artificial Intelligence, Non-interactive Audiovisual Production, Systematic Mapping Study

1 INTRODUCTION

Audiovisual production has undergone transformations driven by the evolution of digital technologies and the increasing incorporation of Artificial Intelligence (AI) techniques in different stages of the creative pipeline, from narrative conception to editing and post-production processes. Intelligent systems have been used to support tasks involving perception, learning, and content generation, expanding possibilities for creation and automation [37].

Recent advances in generative models based on deep neural networks have significantly expanded the capabilities of automated digital media generation, including images, videos, and texts with high levels of quality and realism [9, 19, 25, 36, 42]. In parallel, the development of large language models (Large Language Models - LLMs) and generative systems has expanded creative automation practices and multimodal production, impacting contemporary audiovisual creation processes and their adoption in industrial and creative sectors [3, 14, 17, 20]. Market projections reinforce the urgency of this topic: the generative-AI market is projected to reach US\$109.37 billion by 2030, with an expected compound annual growth rate of 37.6% between 2025 and 2030 [20].

These technologies have altered traditional audiovisual workflows by accelerating iterative creative tasks and lowering the barrier to content generation, integrating GAI into professional and semi-professional practice through storyboard generation, synthetic voice, AI-assisted editing and prompt-based animation [5, 10, 17]. The same adoption has intensified debates over authorship, creative authenticity, labor transformation, algorithmic bias, deepfakes, dataset opacity and aesthetic homogenization, particularly where synthetic content becomes difficult to distinguish from human-created media [16, 43].

Despite the rapid expansion of this field, current research remains highly fragmented across technical, creative, industrial, and ethical perspectives. Different studies frequently focus on isolated aspects of audiovisual generation, such as visual fidelity, narrative automation, editing assistance, or ethical implications, without providing an integrated overview of how GAI technologies are being operationalized throughout non-interactive audiovisual production pipelines [15, 33, 43].

Given this scenario, it is necessary to understand how the scientific literature has investigated the use of Generative Artificial Intelligence (GAI) in non-interactive audiovisual production. Unlike immersive environments or electronic games, non-interactive audiovisual comprises works in which the narrative and information flow are linear and predefined, without the viewer exercising agency or control over the unfolding of the content, as occurs in movies, documentaries, short films, and traditional animations. Therefore, this paper presents a Systematic Mapping Study (SMS), conducted according to consolidated guidelines in the field, with the objective of identifying applications, challenges, impacts, and gaps present in recent scientific production on this theme [26, 35].

Searches were run in IEEE Xplore and the ACM Digital Library, yielding 60 primary studies from 667 records.

The work was guided by the following research question: *“What are the applications, challenges, and impacts of the use of GAI tools in the generation of non-interactive audiovisual content?”*. The importance of this study lies in the need to organize and systematize knowledge in a field that presents accelerated growth and profound transformations.

The results evidence the growth of research on GAI applied to audiovisual, the diversity of technical and creative approaches, and the presence of methodological limitations

and ethical challenges related to authorship and automated content production. Beyond reporting frequencies, this study contributes through: (i) a deliberately computing-centered mapping (IEEE Xplore and ACM Digital Library) to characterize how software and AI research is operationalizing GAI in non-interactive audiovisual pipelines; (ii) an integrated analysis that jointly considers technical choices, evaluation practices, and ethical/human impacts; and (iii) cross-variable synthesis to identify convergences and gaps between model families, application stages, and validation strategies. These contributions help consolidate the state of the art and support future investigations into the responsible adoption of GAI in audiovisual production.

2 THEORETICAL BACKGROUND

2.1 The Audiovisual Ecosystem

The term audiovisual designates any medium that simultaneously combines visual and sound elements to convey a message, narrative, or aesthetic experience [8], and over the last century the field has consolidated into a complex industry spanning cinema, television, advertising and digital media [43]. For the purposes of this mapping, its products are divided into two categories according to the agency granted to the viewer. In interactive audiovisual, exemplified by digital games and virtual reality, the engine renders images dynamically in response to the user's decisions, so the narrative has no fixed duration and the outcome may vary. In non-interactive audiovisual, the exclusive object of this research, the progression is linear [18]: films, television series, short films, music videos and advertising pieces, in which the viewer is receptive and creative agency belongs entirely to the director and the production team, who fix in advance each framing, the duration of each scene and the order of events.

This linearity imposes extreme rigor: because the viewer cannot control the camera, any continuity error, lighting failure or break in temporal coherence between consecutive frames destroys immersion. It is precisely this demand for temporal stability that makes integrating AI into the sector such a complex computational challenge [48].

Linear production follows three phases [18]: *pre-production*, covering scriptwriting, *concept art*, *storyboards* and *animatics*, which fix visual identity and pace before any camera rolls; *production*, the material-capture phase and the most expensive in personnel and resources; and *post-production*, which stitches the fragments into the linear narrative through editing, Visual Effects (VFX), color grading and sound design.

2.2 Fundamentals of GAI

While traditional audiovisual work relies on capturing the physical world or on manual modeling, GAI proposes a different paradigm. A subfield of machine learning [37], it is synthetic rather than analytical: instead of merely detecting patterns, it maps the statistical structure of large pre-existing corpora and uses that knowledge to create novel content from scratch [16].

Three architectural families underpin this evolution. **Generative Adversarial Networks (GANs)**, proposed in 2014, pit a generator against a discriminator in a competition that improved visual realism, though they remain hard to control and prone to instability and artifacts in complex sequences [19, 25]. **Transformers**, from 2017, rely on attention to capture long-range context and enabled few-shot, natural-language-driven interaction, becoming the basis of *prompt*-based tools [3, 14, 42]. **Diffusion models**, which iteratively denoise abstract static into an image, are now the state of the art for visual generation, surpassing GANs in quality and allowing style and composition to be steered via *guidance* without retraining [9, 36].

The public release of *ChatGPT* at the end of 2022 was the decisive inflection point for popularization, making Transformer-based models directly accessible to a general audience [3] and consolidating a *prompt*-centred paradigm of human-machine interaction that shaped the text-to-image and text-to-video tools that followed [10].

2.3 Applications of GAI in the Audiovisual Pipeline

With the maturation of diffusion models and transformers, generative tools ceased to be an experiment and began to infiltrate the traditional workflow of non-interactive production, though unevenly across phases [18].

Pre-production is where these technologies arrive first, with AI-assisted script analysis already informing greenlight decisions [43], and it absorbs them most easily because it is the experimental phase, where small visual inconsistencies are often welcomed as inspiration. Language models assist screenwriters in overcoming creative blocks, structuring narrative arcs and formatting texts [10], while image generators produce *concept art* of settings, costumes and colour palettes in seconds, and visual platforms automate *storyboards*, replacing hours of manual illustration with sketches that dictate the camera composition to be filmed later [1].

Shooting suffers the least direct interference, since it depends on the physical presence of light, actors and settings; its main intersection with GAI is neural rendering, which allows scenes to be relit computationally after capture, without a physical re-lighting setup [11].

Post-production concentrates both the greatest promise of automation and the greatest technical challenges [18]. GAI works well as point assistance: upscaling the resolution of old footage, removing unwanted objects and cleaning captured audio noise far more effectively than classic digital tools [22]; accelerating rotoscoping, one of the most exhausting manual tasks in cinema, through models that track the contours of complex subjects [12]; and cloning actors' voices so that mispronounced words can be corrected without recalling the cast to a dubbing studio [6]. Full automation, such as the autonomous generation of long scenes from text commands alone, still faces severe limitations in this phase [48].

3 RELATED WORK

Recent literature shows a growing body of studies on the use of GAI in audiovisual production, especially after the popularization of models capable of generating images, videos, and automated narrative content. Different reviews have sought to organize this field, focusing on narrative, operational, or ethical perspectives.

Tiwari et al. [41] present a review focused on the use of *deep learning* approaches for *deepfake* detection. The main objective was to analyze media security and integrity techniques and identify effective models for distinguishing synthetic content from real content. The results highlight the evolution of convolutional neural networks on this task, while the gap they foreground is generalizability: sustaining accuracy across datasets, not only the training one. Unlike this work, which focuses on the process of creating and applying GAI, the review concentrates exclusively on security and the containment of technological abuses.

Franco Lazarte et al. [15] conducted a systematic review on the transformation of audiovisual storytelling under AI. Its central conclusion is a paradox: although the technology promises innovation, the reviewed literature shows its current use tending to reproduce nostalgic and retro aesthetics rather than create original ones. Although that review offers a profound analysis of narrative, this mapping differs by adopting a technical-industrial focus, prioritizing the characterization of technological architectures and operational challenges (such as computational cost and visual fidelity) in the specific domain of non-interactive production.

Orak and Turan [33] carried out a systematic review analyzing 21 studies retrieved from the Web of Science, ERIC, Taylor & Francis, Education Full Text EBSCO and ScienceDirect databases, with the objective of identifying the benefits and limitations of AI in digital video production. Its corpus is centred on AI-generated educational and instructional video, grouped into clusters such as efficiency and time saving, automatic editing and subtitling, and it concludes that the potential of AI is not yet efficiently harnessed. The present work instead uses IEEE Xplore and the ACM Digital Library and targets non-interactive audiovisual production, so the two are complementary.

Voinea [43] addresses the creative implications and ethical challenges of GAI, focusing on issues of authorship and copyright. Proposing the *Algorithmic Auteur* framework, it concludes on three dynamics: an efficiency/homogenization nexus, in which predictive logic favours derivative content; labour agreements as emerging algorithmic governance; and a rights-based European regulatory model diverging from a market-driven American one. While that study engages in a predominantly theoretical and ethical discussion, this SMS proposes a unified analytical structure that integrates these ethical challenges into technical applications and their practical impacts on the audiovisual production pipeline.

Other reviews add that, because generative systems are trained on large-scale historical corpora, optimizing for statistically recurrent patterns may reduce stylistic diversity and

favour derivative outputs over genuinely disruptive experimentation [15, 18, 43].

Within WebMedia itself, recent editions have addressed adjacent problems. Martins et al. [32] automate the generation of 3D sign-language animation from machine-translated input, a case of fully generated linear audiovisual output produced for accessibility, and Santos et al. [38] map LLM-based tools supporting usability, accessibility and user-experience activities, a secondary study methodologically parallel to ours. Neither addresses the audiovisual production pipeline as such, which is the gap this mapping occupies within the symposium's own agenda.

Taken together, prior studies provide relevant but still fragmented coverage: some emphasize detection and media-security issues, others focus on narrative and cultural/ethical implications, and others discuss digital video production in instructional contexts [15, 33, 41, 43]. In contrast, our study narrows the domain to non-interactive audiovisual production, uses a computing-centered corpus, and jointly analyzes six dimensions in a single mapping: application targets, model families, evaluation methods, technical limitations, ethical/creative concerns, and industrial/human impacts. This sharper comparative positioning clarifies that our contribution is not only cataloging studies but also connecting technical and socio-creative evidence within a single analytical frame.

4 RESEARCH METHOD

This SMS was conducted following the guidelines in [26, 35]. The objective of this study, defined according to the GQM (*Goal-Question-Metric*) paradigm [2], consists of analyzing the use of GAI, with the purpose of characterizing applications, challenges, and impacts, from the perspective of researchers and professionals in the field, in the context of non-interactive audiovisual production.

The main question is the one stated in Section 1. To answer it, seventeen sub-questions (SQ1-SQ17) were defined, two of which were decomposed into two conditional sub-items each (SQ4.1 and SQ4.2 under SQ4; SQ8.1 and SQ8.2 under SQ8), totalling twenty-one extraction fields, detailed in Table 1.

Residual aggregate categories are reported explicitly rather than suppressed, because in some sub-questions they are among the largest, as in the residual cluster of SQ7 with 15 studies. In SQ8 every study carries a recorded answer, so the eight non-affirmative cases are split into six negative and two partial rather than merged into a single negative class.

Regarding the data synthesis presented in Table 1, it is important to note that the total number of occurrences may deviate from the sample size of 60 papers. Where the sum exceeds 60, as in knowledge areas (SQ2, 88 occurrences), base models (SQ13, 71) and ethical challenges (SQ7, 70), the cause is multi-labelling: one paper may fall under several categories. Conversely, nine sub-questions tabulate fewer than 60 assignments, most markedly SQ9 (39), SQ5 (42), SQ14 (47) and SQ10 (48). This shortfall combines single-occurrence labels, kept apart when the residual option was filled in free

Table 1: Distribution of responses for the SMS research sub-questions. $N = 60$ primary studies. Except for SQ1, SQ4.2, SQ8, SQ8.2 and SQ17, which admit a single answer, all sub-questions are multi-label, so column percentages may exceed 100%. Percentages are computed over $N = 60$; each answer block reads left to right, top to bottom. All frequencies were rebuilt from the per-study identifier lists of the extraction dataset.

Research Subquestion	Answer	N	%	Answer	N	%
SQ1. Type of contribution	Solution Proposal	31	51.67	Evaluation Research	19	31.67
	Validation Research	7	11.67	Other (single occurrences)	3	5.00
SQ2. Area of knowledge	Media and Entertainment	34	56.67	Artificial Intelligence	31	51.67
	Digital Design	13	21.67	Education	7	11.67
	Social Sciences	3	5.00			
SQ3. Main focus of the application	Video generation/modification	19	31.67	Storytelling/narratives	12	20.00
	Other	10	16.67	Static-image creation	7	11.67
	Audiovisual effects	5	8.33	Editing-process automation	2	3.33
SQ4. GAI technology employed	Diffusion Models	29	48.33	Transformers	22	36.67
	Other	12	20.00	GANs	10	16.67
	Hybrid models	6	10.00	Autoencoders	3	5.00
	Recurrent networks	2	3.33			
SQ4.1. Criteria or metrics used	Visual fidelity (Realism)	26	43.33	Subjective eval. (MOS, Aesth.)	19	31.67
	Narrative coherence	8	13.33			
SQ4.2. Evaluator conclusions	Positive / Superior	35	58.33	Mixed or Limited	13	21.67
	Not applicable	12	20.00			
SQ5. Original tool or customization	Combination of existing tools	20	33.33	Partial customization	8	13.33
	Proprietary/extensive dev.	6	10.00	Original, no modifications	6	10.00
	Other	2	3.33			
SQ6. Technical challenges	Other specific challenges	22	36.67	Temporal coherence	16	26.67
	System stability	7	11.67	Latency / Response time	5	8.33
	Scalability	4	6.67	Resolution limits	4	6.67
	Audio-video synchronization	4	6.67			
SQ7. Ethical and creative challenges	Impact on creative industry	19	31.67	Authenticity and authorship	17	28.33
	Other	15	25.00	Algorithmic bias	10	16.67
	Privacy / Data protection	4	6.67	Unauthorized use of refs.	3	5.00
	Plagiarism / Copyright	2	3.33			
SQ8. Empirical analysis present?	Yes	52	86.67	No	6	10.00
	Partially	2	3.33			
SQ8.1. Empirical method	Controlled experiment	22	36.67	Survey / Questionnaire	15	25.00
	Case study	11	18.33	Interview	9	15.00
	Other	5	8.33	Heuristic evaluation	3	5.00
	Participant observation	2	3.33			
SQ8.2. Participants reported	Reported a sample	48	80.00	Did not report / N.A.	12	20.00
SQ9. Main purpose of use	Other	13	21.67	Post-production automation	9	15.00
	Script/narrative creation	9	15.00	Image/photography creation	3	5.00
	Special-effects generation	3	5.00	Interactive animations	2	3.33
SQ10. Content types generated	Static images	14	23.33	Animations	13	21.67
	Full videos	12	20.00	Other	9	15.00
SQ11. Devices and infrastructure	Computers / Notebooks	24	40.00	Cloud Interfaces	12	20.00
	Other	7	11.67	Not informed	5	8.33
	Specialized devices	2	3.33	Immersive environments	1	1.67
	Voice assistants	1	1.67	Mobile devices	1	1.67
SQ12. User interaction type	Textual Prompt	25	41.67	Graphical User Interface	12	20.00
	Multimodal	7	11.67	Textual Prompt (variant label)	4	6.67
	Other	2	3.33	Programmatic (API/SDK)	1	1.67
SQ13. Specific base model used	Stable Diffusion	23	38.33	GPT-4	12	20.00
	Other	11	18.33	MidJourney	8	13.33
	CLIP	7	11.67	DALL-E	7	11.67
	GPT-3	3	5.00			
SQ14. Validation references	Qualitative analyses (Experts)	15	25.00	Quant. metrics (FID, IS, PSNR)	13	21.67
	Industry standards	10	16.67	Other	5	8.33
	A/B tests with users	4	6.67			
SQ15. Potential industry impact	Productivity increase	24	40.00	New creative workflows	18	30.00
	Production democratization	12	20.00			
SQ16. Reproducibility challenges	Data dependency	22	36.67	Variability in results	17	28.33
	Other	8	13.33	Lack of std. benchmarks	7	11.67
	Technical validation diff.	3	5.00	Cross-environment instab.	3	5.00
	Data bias	3	5.00			
SQ17. Intended perceptibility of GAI	Not informed	29	48.33	Yes (Intentional/Stylized)	18	30.00
	No (Imperceptible/Realism)	12	20.00	Partially perceptible	1	1.67

text (as in the duplicated textual-prompt label of SQ12), with fields left unrecorded. Neither class was redistributed, so no percentage rests on imputation, but the shortfall bounds how far SQ5, SQ9 and SQ14 characterize the whole corpus.

Some labels linked to interactivity still appear in isolated cases because a subset of papers briefly references interactive contexts while keeping their main experiment and evaluation focused on linear outputs (e.g., standalone animation assets or non-interactive rendered sequences).

Each field offered a predefined set of answer options, and every answer had to be accompanied by a mandatory verbatim excerpt from the primary study. SQ1, SQ8 and SQ17 admitted a single answer, as did the conditional fields SQ4.2 and SQ8.2; all remaining fields were multi-label, since one study may report several technologies, challenges or content types at once. Three definitions are stated explicitly here because they govern how the results should be read. In SQ4.2, *evaluator* denotes whoever performed the evaluation reported in the primary study (its own authors, an expert panel, or user-study participants), and *content* denotes the audiovisual artifact generated with GAI, not the paper. In SQ5, *original tool without modifications* means a third-party tool used as-is, whereas *proprietary/extensive development* means a system built by the authors of the primary study. In SQ11 the classes are not mutually exclusive: a study whose interface runs on a desktop while the model is invoked through a remote API is recorded under both *computers/notebooks* and *cloud interfaces*, which is the case for 10 of the 24 studies in the former class. SQ2 records the field or fields in which the study positions itself, which is why a corpus entirely about GAI still shows only 31 studies under Artificial Intelligence: the remainder frame their contribution as media, design or education work that employs GAI rather than as AI research.

The taxonomy adopted for SQ1 follows Wieringa et al. [46]: *Solution Proposal* denotes a novel method or tool, or a concrete improvement of an existing one; *Evaluation Research* examines an existing approach in a real or realistic context with empirical evidence; *Validation Research* tests a proposal in controlled or prototype settings prior to broad deployment; and the *Philosophical*, *Opinion* and *Experience* categories cover conceptual reflection, author positioning and practice-based lessons.

The search strategy was based on the PICOC model [26]: *Population*: Generative Artificial Intelligence; *Intervention*: generation and manipulation of content; *Comparison*: not applicable, as the objective of this SMS is the characterization and survey of the state of the art of a broad and emerging area, and not the experimental comparison between specific techniques or algorithms; *Outcome*: impacts, ethics, quality, and creativity; *Context*: non-interactive audiovisual production.

The search string used was: ("Artificial intelligence" OR "AI") AND ("generate" OR "produce" OR "create" OR "manipulate" OR "write") AND ("impact" OR "ethics" OR "quality" OR "consequences" OR "perception" OR "realism" OR

"creativity") AND ("film" OR "cinema" OR "movie" OR "animation" OR "advertisement" OR "CGI" OR "special effects" OR "VFX" OR "script" OR "voice-over").

Searches were conducted in the IEEE Xplore (<https://ieeexplore.ieee.org>) and ACM Digital Library (<https://dl.acm.org>) digital libraries, due to their strong coverage of applied computing and software engineering research, consistent with the computing-centered scope of this SMS. Therefore, claims in this paper are bounded to evidence indexed in those venues and should not be interpreted as an exhaustive panorama of all media/design publication ecosystems. Before final extraction, we conducted pilot searches to calibrate the search string, testing alternative terms and checking whether known relevant papers were retrieved (e.g., [3, 19, 36, 42]). The final search string was adopted after this calibration stage, balanced recall and thematic precision for the non-interactive audiovisual context. The initial extraction of search results was exported to the Porifera platform at 12:53 on April 26, 2025. The defined temporal window (2022-2025) is justified by the fact that 2022 marks a technological inflection point with the launch and widespread adoption of modern GAI architectures, such as large language models and latent diffusion models [3, 36], which have significantly impacted audiovisual production.

No study published in 2022 satisfied the inclusion criteria, so the declared 2022-2025 window is the window searched, not the window observed: the search covered records from January 1, 2022 up to the export date of April 26, 2025, which is also why 2025 is represented only partially. The resulting distribution is reported in Section 5.

The selection criteria were fundamental to maintain the scope of the study. The inclusion criteria adopted were: IC1: Studies with AI applications in audiovisual content; and IC2: Presence of technical, creative, or operational evaluation. The exclusion criteria were EC1: Thematic irrelevance; EC2: Non-peer-reviewed works; EC3: Duplicates; EC4: Incomplete documents; EC5: Languages other than Portuguese or English; EC6: Literature reviews; and EC7: Studies focused on interactive media (games), biomedical or forensic applications.

Most exclusions in the initial stage were due to thematic irrelevance and non-peer-reviewed works, while full-text exclusions were mainly related to scope mismatch and lack of evaluation. One record was identified as a duplicate and removed under EC3 during the 667-to-162 transition. IC2 required some form of technical, creative, aesthetic or operational evaluation of the generated content, which is broader than a formal empirical study: this is why all 60 studies satisfy IC2 while SQ8 finds explicit empirical analysis in 52 of them. No snowballing was performed, since the mapping deliberately bounds its corpus to peer-reviewed records indexed in the two computing libraries in order to keep the search fully re-executable; the resulting coverage limitation is registered in Section 5.3.

The selection process was conducted in two stages: (i) screening by title, abstract, and keywords; and (ii) full reading of eligible studies. In both stages, the evaluation was

performed independently by four researchers, followed by consensus discussion.

Agreement was measured by the Fleiss Kappa index [13], obtaining values of 0.730 in the 1st stage and 0.645 in the 2nd stage, indicating substantial levels of agreement according to established interpretation criteria [27], which place the 0.61-0.80 range in the substantial band. The Porifera platform [4] was used to record decisions and maintain traceability in both stages.

To automate data extraction and analysis, custom Python scripts were developed to execute three main steps.

The extraction instrument was one .docx form per study, filled in by the researchers themselves and not by the evaluators of the primary studies. Each form held a structured table with the twenty-one fields, the predefined answer options marked as checkboxes, and a mandatory free-text field carrying the verbatim excerpt that justified each choice. First, the completed .docx extraction forms were converted to plain text (.txt), stripping visual formatting but preserving checkbox markers. Second, all individual files were merged into a single consolidated dataset. Finally, a quantitative analysis script used regular expressions and keyword matching to identify article IDs, extract responses from checkboxes and text delimiters, normalize the data against predefined options, and automatically calculate response frequencies to generate the final structured report. Because this scripted extraction could introduce construct validity risks, we added verification procedures: researchers manually inspected sampled outputs against original forms. This human-in-the-loop validation step reduced the chance of systematic misclassification caused by parsing or normalization rules.

To ensure transparency and reproducibility, the list of extracted studies and data synthesis are available at the following repository: <https://figshare.com/s/6b28d0b8c22d009b6af9>.

5 QUANTITATIVE RESULTS

The systematic mapping process followed a progressive filtering strategy, beginning with 667 studies retrieved from the selected digital libraries (500 from IEEE Xplore and 167 from ACM Digital Library). After title, abstract, and keyword screening, 162 studies were considered potentially relevant for full-text assessment (104 IEEE and 58 ACM), corresponding to a first-stage retention rate of 24.3% (162/667). The second-stage screening resulted in a final portfolio of 60 papers used for quantitative and qualitative synthesis (35 IEEE and 25 ACM),

equivalent to 9.0% of the initial set (60/667) and 37.0% of first-stage approvals (60/162). Figure 1 reports these counts per database and per stage.

The temporal distribution of the selected studies, shown in Figure 2, follows a rapid growth trend: 12 studies (20.0%) in 2023, 39 (65.0%) in 2024 and 9 (15.0%) up to the April 2025 cut-off, with none from 2022 meeting the criteria. The 2023-2024 biennium alone accounts for 51 of the 60 papers (85.0%), evidencing the accelerated academic consolidation of

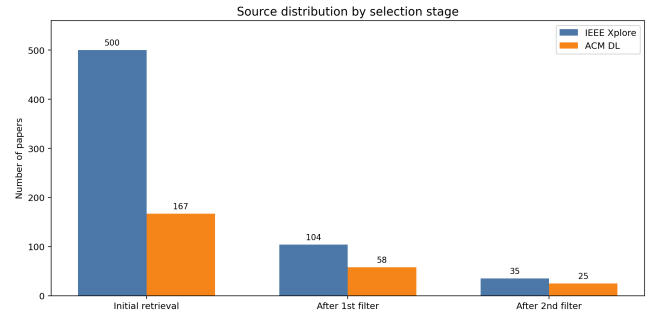


Figure 1: Number of records per database at each selection stage. IEEE Xplore remains the larger source across all filters, while the ACM Digital Library contributes a growing share of the corpus as screening advances, supplying 25 of the 60 studies finally selected.

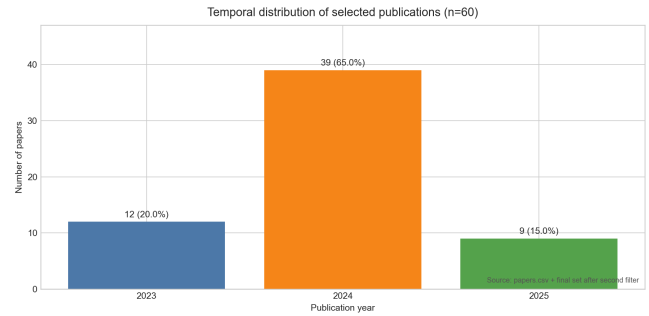


Figure 2: Temporal distribution of the 60 selected studies. No study published in 2022 met the inclusion criteria, and 2025 is covered only up to the search date of April 26, 2025.

GAI in non-interactive audiovisual production after the dissemination of publicly accessible systems such as ChatGPT, Stable Diffusion and Midjourney.

Regarding publication venues, conference and workshop papers strongly predominate in the final corpus, totalling 57 studies (95.0%), while journal articles account for only 3 papers (5.0%). This predominance of conference publications is characteristic of rapidly evolving technological domains, where dissemination speed often precedes methodological consolidation and large-scale validation. Many studies present prototype-oriented systems, exploratory pipelines, or early-stage empirical analyses rather than mature industrial deployments.

To provide a venue-level perspective, we additionally normalized publication sources using DOI metadata for the final set of 60 studies. The three journal papers are concentrated in *IEEE Transactions on Visualization and Computer Graphics* (2 papers) and *IEEE Access* (1 paper). In the conference/workshop view by acronym, the *ACM Conference on Human Factors in Computing Systems* (CHI) is the most recurrent venue with 8 papers, followed by the *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR) with 5 papers, while the *Extended Abstracts of the*

CHI Conference on Human Factors in Computing Systems (CHI-EA), the *IEEE International Conference on Multimedia and Expo Workshops* (ICMEW), the *ACM International Conference on Multimedia* (MM), and the *International Symposium on Visual Information Communication and Interaction* (VINCI) appear with 2 papers each. Workshop proceedings are the least frequent venue type in the corpus: the most recurrent of them is ICMEW, already counted above, followed by isolated records in the *ACM International Conference on Interactive Media Experiences Workshops* (IMXW) and the *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (CVPRW).

These patterns indicate two complementary profiles: CHI and CHI-EA concentrate human-AI interaction and production-oriented design, while CVPR, MM, VINCI and ICMEW carry a computer-vision and multimedia-engineering profile, with archival journal publication confined to technically oriented outlets.

5.1 Results and Discussion by Sub-question

SQ1 - Type of contribution. In the Wieringa-based taxonomy defined in Section 4, solution proposals predominate with 31 studies (51.67%), followed by evaluation research with 19 (31.67%) and validation research with 7 (11.67%), plus three residual studies with one occurrence each: an opinion paper, an experience report and a study whose recorded label falls outside the taxonomy, which closes the 60 studies exactly as required by this single-answer field. *Anim-Director* [28] is paradigmatic of what a “solution proposal” currently looks like in this field. The authors argue that traditional animation pipelines depend on training generative models with human-labelled data, producing short, information-poor, and context-incoherent clips; their answer is an autonomous agent in which a large multimodal model (GPT-4 with vision) acts as a “director”, refining the input narrative, generating scene-level prompts, invoking off-the-shelf image and video generators, and then *evaluating and selecting* the candidates itself through a self-reflection pathway, without manual intervention. Crucially for SQ1, the proposal is not merely demonstrative: it is validated on narratives sampled from the TinyStories corpus and assessed with consistency and video-quality metrics adapted from VBench-style evaluation suites, and the authors’ own ablation shows that removing the LMM-based selection step degrades video quality and subject/background consistency scores. At the same time, they acknowledge that semantic text-video alignment remains only marginally better than baselines, and that maintaining stable backgrounds across clips remains problematic.

SQ2 - Area of knowledge. The largest concentrations are Media and Entertainment (34 studies, 56.67%), Artificial Intelligence (31 studies, 51.67%), and Digital Design (13 studies, 21.67%), with additional participation from Education (7 studies, 11.67%) and Social Sciences (3 studies, 5.00%). Because these are multi-label classifications, their joint frequency supports an interdisciplinary interpretation

rather than isolated-domain adoption. *ReelFramer* [44] illustrates how these areas overlap inside a single artifact: it addresses a journalism problem, since news outlets struggle to translate print articles into short social-media videos that young audiences consume; it borrows its central construct, narrative framing, from literary theory; and it operationalizes that construct with generative models that co-produce scripts and storyboards with the user. The authors identify three distinct narrative framings for news reels, each balancing information and entertainment differently, and show that two design decisions are decisive: letting the journalist explore multiple framings, and establishing the foundational details of the narrative before generation, which yields scripts measurably more relevant and coherent than direct article-to-script generation.

SQ3 - Main focus of the GAI application. The distribution spans video generation/modification (19 studies), storytelling/narratives (12 studies), a residual cluster (10 studies), static-image creation (7 studies), audiovisual effects (5 studies) and editing-process automation (2 studies). The co-dominance of video and narrative is well captured by *Intelligent Grimm* [30], which deliberately sits between the two: its task, open-ended visual storytelling, requires generating a coherent *sequence* of images that follows a storyline rather than isolated frames. Its contribution, StoryGen, is an autoregressive generator built on a latent diffusion backbone with a vision-language context module, so that each new frame is conditioned not only on its text prompt but also on the preceding image-caption pairs, an explicit architectural answer to the narrative-consistency problem that single-shot text-to-image models cannot address. Equally revealing of the field’s priorities is what the authors had to build first: because existing storytelling datasets covered only a few animations with fewer than ten recurring characters, they constructed StorySalon, a corpus of nearly 160,000 animation-style images with an average of 14 frames per story, precisely to make long-range semantic correspondence learnable.

SQ4 - GAI technology employed. Diffusion models (29 studies) and transformers (22 studies) are clearly dominant, while a residual cluster (12 studies), GANs (10 studies), hybrid models (6 studies), autoencoders (3 studies), and recurrent networks (2 studies) appear with lower frequencies. More interesting than the raw counts is that the two dominant families are also used *together*, as the hybrid class and the overlaps in Table 2 show, with the transformer acting as the planner and the diffusion model as the renderer. *GPT4Motion* [31] is the clearest expression of this division of labor. Its starting diagnosis is that end-to-end video-diffusion models suffer from high computational cost and physically incoherent motion because they are asked to hallucinate dynamics they were never explicitly taught. The framework answers with a strict separation of competences: GPT-4 translates the prompt into an executable Blender script; Blender’s physics engine simulates the scene dynamics across frames; and only then is a Stable Diffusion model conditioned on the simulated frames to render the final video, all of it training-free. The authors validate on three canonical classes of physical motion,

reporting that the method outperforms four text-to-video baselines on every quantitative metric and that residual flickering is significantly lower than in competing models, and they add a user study with 30 participants in which their videos received all the votes for physical accuracy, text-video alignment and lowest flickering.

SQ4.1 - Criteria and metrics used. Visual fidelity/realism appears in 26 studies (43.33%), subjective assessments in 19 studies (31.67%), and narrative coherence in only 8 studies (13.33%). The work of Cheng et al. on audio-driven gesture generation [7] shows what the dominant fidelity regime looks like when fully developed. Motivated by a precise trade-off, since VAE-based generators suffer from jitter and instability while diffusion-based ones produce smooth motion at prohibitive inference cost, the authors evaluate their conditional-GAN-enhanced diffusion model on two public datasets (BEAT and ZeroEGGs) with two complementary metrics: Fréchet Gesture Distance (FGD), the distance between the latent feature distributions of generated and real motion, and Beat Alignment, the synchrony between gesture and audio beats. On BEAT the model reaches an FGD of 84.0 against 100.3 for the strongest diffusion baseline, and on ZeroEGGs 54.7 against 78.4, while cutting per-frame inference to roughly 3-4 ms against tens of milliseconds for competing approaches. The broader pattern is that the field measures most reliably what it can quantify (realism, distributional distance, perceptual preference) while the narrative properties most specific to linear audiovisual production remain the least instrumented, for want of agreed, comparable metrics.

SQ4.2 - Evaluator conclusions about the content. Positive or superior results are reported in 35 studies (58.33%), mixed or limited results in 13 studies (21.67%), and 12 studies (20.00%) do not record an evaluator conclusion. This residual class is not an absence of any assessment, since IC2 required some form of evaluation of the generated content; it is the absence of a recorded conclusion. *TADA!* [29] typifies the dominant “positive” profile and also reveals its boundaries. It generates expressive 3D avatars from text by combining a 2D diffusion model with the SMPL-X parametric body model, explicitly targeting a documented failure of prior text-to-avatar methods, the misalignment between geometry and texture in the face region, which made earlier avatars plausible as stills but unusable once animated. In the user study, participants preferred *TADA* over all compared baselines on geometry, texture and consistency with the prompt, and because expression parameters deform the character during training, the avatars can be driven by standard animation sources. The opposite pole, the 13 studies reporting mixed or limited results, is exemplified by Grigis and De Angeli [21]: their seven professional playwrights judged LLMs unsuitable for dramaturgy yet useful for editorial tasks, the decisive obstacle being not output quality but the safety policy embedded in the models, which refused the taboos and conflicts dramaturgy is built on. Even so, the framing of the positive cases is symptomatic: superiority is superiority *relative to other generative baselines*, under prompts chosen

by the authors, rather than certified adequacy for professional character pipelines.

SQ5 - Original tool versus customization. Twenty studies adopt combinations of existing tools, 8 report partial customization, 6 present proprietary/extensive development, 6 use original tools without modifications, and 2 fall under other strategies. The integration-oriented categories hold 34 of the 42 recorded labels (20+8+6), so the field prioritizes orchestration over pure off-the-shelf use; SQ5 being multi-label, these are labels, not distinct studies. *Anim-Director* [28], analysed under SQ1, is the corpus’s clearest specimen of this orchestration paradigm: none of its generative components is novel, and the contribution lies entirely in the control loop that generates candidate prompts, inspects the resulting images, scores candidate videos and iterates. The authors’ ablation shows that removing the LMM-driven selection step measurably degrades video quality and consistency scores, so a substantial share of the reported superiority comes from the orchestration rather than from the generators themselves.

SQ6 - Technical challenges identified. The challenges are led by an “other” cluster of specific bottlenecks (22 studies, 36.67%), followed by temporal coherence (16 studies, 26.67%), system stability (7 studies, 11.67%), latency/response time (5 studies, 8.33%), scalability (4 studies, 6.67%), resolution limits (4 studies, 6.67%), and audio-video synchronization (4 studies, 6.67%). The residual cluster deserves the same treatment given to its counterpart in SQ9: it aggregates bottlenecks that are genuinely study-specific and resist a shared vocabulary, the largest recurring one being computational cost (2 studies), with the remainder appearing once each. Its size is therefore a finding about the immaturity of the field’s technical vocabulary rather than a defect of the classification scheme, and it does not displace temporal coherence as the leading named obstacle. *Text2NeRF* [49] makes these abstract categories concrete by documenting its failure modes with unusual candor. The system generates navigable 3D scenes from a single text prompt by using a pre-trained diffusion model as content prior and monocular depth estimation as geometric prior, with a progressive inpainting-and-updating strategy to keep texture and geometry consistent as the virtual camera moves, the 3D analogue of temporal coherence. The reported limitations map directly onto the SQ6 taxonomy: when the scene contains large occlusions, inaccurate depth estimates produce visible geometric distortions in the warped views, and the distortion grows with camera offset, eventually defeating the inpainting stage (a coherence/stability failure); and each scene requires on the order of 1.5 hours of NeRF optimization (a latency/scalability failure), far from interactive production timescales. The general pattern is that the binding constraints are no longer per-frame quality, which diffusion models largely solved, but *cross-frame and cross-view consistency* and the computational budget required to enforce it. Within this corpus, then, the leading named obstacle to integrating GAI into linear audiovisual production is the lack of temporal coherence: diffusion-based tools struggle to maintain the consistency of an object’s identity over time, so that a generated striped shirt may change its number of

stripes from one frame to the next, a phenomenon known in the industry as “flickering”.

SQ7 - Ethical and creative challenges. The most frequent categories are impact on the creative industry (19 studies, 31.67%), authenticity/authorship (17 studies, 28.33%), a residual cluster (15 studies, 25.00%) and algorithmic bias (10 studies, 16.67%), followed by privacy/data protection (4 studies, 6.67%), unauthorized use of references (3 studies, 5.00%), and plagiarism/copyright (2 studies, 3.33%). *wr-AI-ter* [45] speaks directly to the leading category, but is notable because it treats authorship not as a discussion-section concern but as the dependent variable of the experiment. Motivated by prior evidence that AI co-writing tools erode writers’ psychological ownership of the resulting text, the authors structure screenwriting into four stages in which the human retains decision authority at each step, and measure the effect with 23 participants of heterogeneous expertise using Likert-scale instruments. Their central finding is that AI assistance accelerated screenwriting and improved perceived script quality *without* compromising the participants’ sense of ownership, in contrast with earlier systems where ownership scored among the lowest dimensions. For SQ7 this reframes the dominant category as a design problem rather than only a policy problem: whether GAI displaces or empowers creative identity depends measurably on where the interaction places human decisions. The corpus, however, contains few studies of this kind; most register ethical concerns without operationalizing them, which is why authorship remains a contested terrain rather than an engineering requirement.

SQ8 - Presence of empirical analysis. Empirical validation is a strong characteristic of the field, with 52 studies (86.67%) reporting empirical analysis, against 6 non-empirical studies and 2 partially empirical cases, with no study left without a recorded answer. The audio-description tool *EasyAD* [39] stands out because its empirical evaluation extends to real deployment. Blind and visually impaired audiences access the visual layer of films through audio descriptions inserted into speech gaps, but producing them is laborious, involving gap identification, script authoring, dubbing and integration. *EasyAD* automates the full chain, detecting the speech gaps, generating candidate description text through a multimodal large language model and synthesizing the narration via text-to-speech, while also correcting a recurring defect of earlier tools, the misidentification of background music as speech. Decisively for SQ8, the system is not a laboratory prototype: it is operational at the China Braille Library, and the user study was conducted with six professional audio-description authors working there, in their actual production context, where processing time for a medium-difficulty movie fell by nearly 50%. This is empirical evidence of a different order from laboratory benchmarks, since it measures workload reduction for real professionals on the content they were going to produce anyway.

SQ8.1 - Empirical method adopted. Controlled experiments lead (22 studies), followed by surveys/questionnaires (15 studies), case studies (11 studies), interviews (9 studies), a residual

cluster (5 studies), heuristic evaluation (3 studies) and participant observation (2 studies). The case-study tradition is exemplified by Song and Yip’s account of the short film *Giant* [40]. Rather than benchmarking a model, the authors propose a pre-production pipeline and guidelines for generating mood boards from text using text-to-image models, and validate the proposal by producing a pre-visualization of a film project. Their starting premise is itself a methodological position: text-to-image generation is a stochastic process that requires human intervention to assist creation, so unguided prompting yields inconsistent material and the research value lies in the curated workflow around the model rather than in its raw outputs. The case study concludes that following the guidelines allowed the filmmakers to generate mood boards that conveyed the intended atmosphere of the project, a claim about fitness within a real creative practice rather than about measurable superiority.

SQ8.2 - Participants reported. Forty-eight studies (80.00%) reported sample information and 12 studies (20.00%) did not. That class is “did not report or not applicable” and subsumes the eight non-affirmative SQ8 cases. Among the reporting studies, samples are typically small and purposive rather than population-representative. *ReelFramer* [44], analysed under SQ2, is characteristic: its summative study involved five participants, but recruitment was sharply targeted at the tool’s intended users, journalism graduate students who consume short-form video, and the ten resulting reels were additionally judged by domain experts. Such a sample supports no population-level inference, and the authors claim none; it is calibrated to observe how the co-creation process unfolds rather than to estimate an effect.

SQ9 - Main purpose of GAI use. The most frequent labels are “other” (13 studies, 21.67%), post-production automation (9 studies, 15.00%), and script/narrative creation (9 studies, 15.00%), followed by image/photography creation and special-effects generation (3 studies each) and interactive animations (2 studies). The size of the “other” cluster is itself a finding, and the intelligent teaching platform of Wu et al. [47] shows why it should not be dismissed as classification noise. The platform is built for education, not cinema: it combines an LLM-based tutoring assistant with a video-generation feature through which teachers upload background images and lecture scripts and the system produces digital-human teaching videos. Nothing in this purpose fits the classical film-pipeline categories, and the “author” is a teacher with no audiovisual training; yet the output is unambiguously linear, non-interactive audiovisual content, produced end-to-end by generative models and consumed exactly as a filmed lecture would be. Cases like this indicate that GAI is not merely automating existing audiovisual workflows but creating *new demand* for linear video in domains such as education, institutional communication and accessibility that previously could not afford production at all.

SQ10 - Types of audiovisual content generated. The outputs are relatively balanced between static images (14 studies), animations (13 studies), and full videos (12 studies), with a residual cluster (9 studies). *MAS* [24] illustrates the

animation-asset profile and, simultaneously, why modular outputs dominate over finished films. The method generates 3D motion sequences using only a 2D diffusion model trained on poses extracted from in-the-wild videos by off-the-shelf estimators, with no motion-capture data at any point. Its mechanism denoises several 2D sequences representing different camera views of the same motion simultaneously, and at every diffusion step a consistency block triangulates them into a single unified 3D sequence, which is reprojected onto the views before the next iteration, so that 3D consistency is enforced throughout generation rather than patched afterward. The authors deliberately choose domains where motion capture is impractical but footage is abundant: professional basketball manoeuvres from broadcast recordings, rhythmic gymnastics and horse races. The output, however, is not a film: it is a reusable *asset*, a skeletal motion intended to be consumed downstream by character animation and rendering stages.

SQ11 - Devices and infrastructure. Computers/notebooks are the dominant setup (24 studies), followed by cloud interfaces (12 studies), with marginal frequencies for specialized devices (2 studies), immersive environments, voice assistants and mobile devices (1 study each), plus “other” (7 studies) and not-informed cases (5 studies). As stated in Section 4, these classes are not exclusive: 10 of the 24 desktop studies also appear under cloud interfaces, because the interface runs locally while the model is invoked remotely. The near absence of mobile and edge deployment becomes understandable when computational requirements are examined: *Text2NeRF* [49], discussed under SQ6, needs roughly 1.5 hours of per-scene GPU optimization before a single navigable result exists. Workloads of this magnitude structurally bind the research to desktop GPUs and cloud back-ends: a phone cannot run the optimization, and a browser can at most stream its result. The infrastructure profile of SQ11 is thus less a design choice than a symptom of the field’s computational economics, and it also shapes *who* can participate in the research itself: groups without sustained GPU access can study prompting and interaction (SQ12) but not generation methods.

SQ12 - Type of user interaction. Textual prompting predominates (25 studies), followed by graphical interfaces (12 studies), multimodal interaction (7 studies), “other” (2 studies) and programmatic interaction (1 study). A further 4 studies were recorded under a variant spelling of the textual-prompt label produced by the extraction normalization; merging them raises textual interaction to 29 studies (48.33%). *GPT4Motion* [31], analysed under SQ4, exposes the hidden structure of this paradigm: what the user types is a single sentence, but it is inserted into a prompt template authored by the system designers that instructs GPT-4 to emit a syntactically valid Blender script, and the whole reliability of the system rests on that template, since a malformed script produces no video at all rather than a worse one. The hegemony of textual prompts therefore does not mean interaction design has become trivial, but that its complexity has been displaced from GUIs into prompt scaffolding, a layer that is rarely documented, versioned, or evaluated, with direct

consequences for the reproducibility problems registered in SQ16.

SQ13 - Specific base model used. Usage concentrates on Stable Diffusion (23 studies), GPT-4 (12 studies), MidJourney (8 studies), CLIP (7 studies), DALL-E (7 studies), and GPT-3 (3 studies), with 11 “other” occurrences. The dominance of Stable Diffusion follows from its access regime rather than from quality alone: because its weights are open, contributions can be grafted onto the pre-trained backbone, as *Intelligent Grimm* [30] does under SQ3 by introducing a vision-language context module on top of the existing cross-attention, inheriting the backbone’s visual competence for free. Closed models such as MidJourney can be prompted but never modified, which confines them to the integration strategies of SQ5 rather than the architectural ones. The 23-to-8 ratio in SQ13 is therefore also a map of what kind of science each access regime permits.

SQ14 - Reference models used in validation. Validation combines qualitative expert analyses (15 studies), quantitative metrics such as FID/IS/PSNR (13 studies), industry-standard comparison (10 studies), and user A/B tests (4 studies), with additional “other” strategies (5 studies). The strongest studies hedge across these categories instead of choosing one, as *TADA!* [29] does under SQ4.2 by combining qualitative renders, a user study and a functional animation test. That last layer is the most production-relevant and the least standardized: there is no shared benchmark for animatability the way FID benchmarks image realism, so every paper that cares about it improvises its own demonstration, and no two demonstrations are comparable.

SQ15 - Potential impact on the audiovisual industry. Expected impacts concentrate on productivity increase (24 studies, 40.00%), new creative workflows (18 studies, 30.00%), and production democratization (12 studies, 20.00%). The text-to-multimedia pipeline of Jadhav et al. [23] is representative of the productivity-centered framing: its declared contribution is explicitly the acceleration of content creation rather than any single algorithmic advance. The framing is symptomatic: impact is argued from capability rather than measured in production economics, with cost, throughput and quality deltas in a working studio left unquantified. Only deployment-grade studies such as EasyAD (SQ8) close that gap, and they are rare. The democratization category requires the same scrutiny: as SQ11 showed, access has been democratized at the interface level while remaining concentrated at the compute and platform level, so that lowered entry barriers coexist with new dependencies on infrastructure owned by a small number of providers.

SQ16 - Reproducibility challenges. The barriers concentrate on training-data dependency (22 studies), result variability (17 studies), an “other” cluster (8 studies) and lack of standardized benchmarks (7 studies), followed by validation difficulties, cross-environment instability and data bias (3 studies each). *MAS* [24], analysed under SQ10, is instructive because data dependency is simultaneously its motivation and its residual weakness: the method exists precisely because 3D motion-capture data is scarce and expensive, and

Table 2: Co-occurrence of application focus (SQ3) and model family (SQ4), from the per-study identifier lists. Both dimensions are multi-label, so neither row nor column sums reproduce the marginals of Table 1; the SQ3 residual cluster (10 studies) is not shown.

Focus (SQ3)	Diff.	Trans.	GAN	Hyb.	AE	RNN	Studies
Video gen./modification	12	12	5	3	1	0	19
Storytelling/narratives	8	10	0	3	0	0	12
Static-image creation	6	4	2	1	0	0	7
Audiovisual effects	4	3	2	0	1	0	5
Editing automation	1	2	0	0	0	0	2

it substitutes abundant 2D footage processed by off-the-shelf pose estimators. That substitution relocates the dependency rather than removing it, since the learned motion prior is now bounded by the accuracy of those upstream estimators and by the idiosyncrasies of the harvested footage, and the evaluation itself must lean on further pose-estimation models. A replication with different estimators, footage sources or estimator versions would therefore not guarantee reproduction of the reported behaviour, even with the authors’ code in hand.

SQ17 - Intended perceptibility of GAI use. The largest class is silence: 29 studies (48.33%) do not state the intention, followed by 18 studies (30.00%) that make GAI use intentionally visible (“Yes”, intentional/stylized), 12 (20.00%) that pursue imperceptibility (“No”), and one “Partially” case. The silent class aggregates every form that declares no intention either way; its composition is detailed in Section 5.3. The stylized pole is well represented by the *Giant* case study [40], analysed under SQ8.1: because the system’s role is confined to pre-production mood boards, perfect realism is neither required nor desired, and disclosure is structurally guaranteed because the artifact never reaches the audience. Contrast this with the 12 studies (20.00%) pursuing imperceptibility, where success is defined precisely as the viewer’s inability to detect synthesis: in that regime, every technical advance in coherence and fidelity simultaneously raises the stakes of non-disclosure, concentrating the ethical exposure mapped in SQ7 (deepfake misuse, undisclosed synthetic media, authorship erasure) on exactly the studies that are technically most successful. The most telling number, however, is the largest one: nearly half of the corpus (29 studies, 48.33%) does not report its perceptibility intention at all, at a time when regulators, platforms, and industry bodies are beginning to demand provenance disclosure from commercial productions.

5.2 Cross-analysis and Emerging Patterns

Table 2 shows that video- and VFX-centered studies rely predominantly on diffusion and transformer pipelines. Read against SQ4.1, a further asymmetry emerges, which the corpus illustrates rather than tabulates: fidelity-oriented evaluation concentrates in those targets, while narrative work falls back on small expert panels and qualitative judgment. Within that table, diffusion and transformers tie on video (12 studies each), but transformers lead on storytelling (10

against 8) and on editing automation, while diffusion leads on static images (6 against 4) and on effects, and GANs survive almost exclusively in the pixel-oriented targets, being entirely absent from narrative work. Where standardized distributional metrics exist, as in the gesture-generation work of Cheng et al. [7], evaluation becomes numerically comparable; where they do not, as in ReelFramer [44], it cannot, which yields uneven evaluation maturity across application targets [15, 43].

Connecting SQ6, SQ7, SQ15, and SQ17 suggests that technical limits do not stay technical: improvements in coherence and realism of the kind pursued by GPT4Motion [31] are precisely what turn the openly stylized outputs of the *Giant* workflow [40] into potentially deceptive media, intensifying the authorship and disclosure tensions that wr-AI-ter [45] addresses at the interaction-design level. Ethical debate thus intensifies in high-realism scenarios and recedes in intentionally stylized ones, and that boundary moves as the limits of SQ6 are overcome.

Finally, SQ8, SQ8.1, SQ8.2, and SQ16 expose a methodological gap: empirical validation is frequent and the controlled experiment is its modal form (22 of the 67 method labels), yet the absence of standardized benchmarks is itself a reported reproducibility barrier (7 studies), and the samples in the exemplars above are small and purposive. EasyAD [39] shows that ecologically valid measurement is feasible (a near 50% reduction in processing time with professional describers in a real institution), which makes its rarity a choice of the field rather than an impossibility. In practice, the field advances quickly in prototyping but still lacks the open, comparable evidence needed to transfer results from laboratory to production.

5.3 Threats to Validity

Construct validity. Extraction was automated with Python scripts over the completed forms, so a parsing or normalization rule could misclassify an answer. Mitigations: the pipeline is idempotent, verified by SHA-256 hash of the final report across repeated runs; test files exercised checkbox variants and quoted excerpts; and a manual audit, first on 10% and later on 50% of the forms, yielded 100% agreement on study identifiers, 98.3% on selected options and 95.8% on identified technologies (computed over audited fields, not whole forms), the divergences tracing to nomenclature variants such as “GPT-4” against “GPT4”. The label variant recorded in SQ12 is a visible instance of that error class. The search string is a second construct threat, since a fast-moving vocabulary may escape terms fixed in April 2025; pilot searches checking retrieval of known relevant papers were used to calibrate it.

Internal validity. Classification into the extraction categories is interpretative. Screening was independent across four researchers, with $\kappa = 0.730$ and $\kappa = 0.645$, in the substantial band but not perfect; borderline studies, typically those mentioning AI and audiovisual only tangentially, were reassessed by at least two researchers and settled by consensus recorded in Porifera. In SQ17, nine studies answered the

perceptibility field in free text as not explicitly declared, and three recorded it under a transparency label without stating an intention; all twelve are counted as not informed rather than as negative answers, since none declares an intention either way.

External validity. The corpus is deliberately computing-centred, drawn from IEEE Xplore and the ACM Digital Library, and therefore under-represents communication, media, design and arts venues where part of this literature is published. No snowballing was performed. Findings should be read as a characterization of how software and AI research operationalizes GAI in linear audiovisual production, not as an exhaustive panorama of the field.

Conclusion validity. In two sub-questions the residual aggregate category is the largest single class: technical challenges (SQ6, 22 studies) and purpose of use (SQ9, 13 studies). Both are discussed as substantive findings about the absence of a shared vocabulary in the field, but they do limit how far the corresponding distributions can be interpreted. The coverage shortfall reported in Section 4, which Table 2 shares, is a related limit. The fast pace of the field also bounds the temporal validity of every frequency reported here.

6 CONCLUSION AND FUTURE WORK

The mapped evidence indicates a field in transition from technical experimentation toward broader deployment contexts, with 19 studies (31.67%) already classified as evaluation research in real or realistic settings, while still presenting methodological fragilities. The predominance of solution proposals (51.67%) suggests that current effort is concentrated on immediate engineering bottlenecks (e.g., temporal stability and semantic controllability). The corpus's own distribution is legible here: temporal coherence leads the named bottlenecks (16 studies), while narrative coherence, the property most specific to linear audiovisual production, is assessed in only 8. At the same time, recurring reports of limited narrative/emotional depth indicate that visual quality gains do not automatically translate into stronger expressive outcomes.

The corpus also reveals structural dependence on external model ecosystems and limited standardization of evaluation practices, which together constrain reproducibility and cross-study comparability. These patterns suggest that methodological consolidation remains an important challenge as GAI adoption expands in audiovisual production.

From an ethical and labor perspective, the mapped literature points to concerns that go beyond automation efficiency, including authorship boundaries, accountability, and changes in professional identity. Our interpretation, derived from these patterns, is that near-term impacts are more consistent with role reconfiguration (e.g., curation, orchestration, and supervision of generative systems) than with total substitution of audiovisual pipelines.

A broader paradox emerges: the same technologies that expand creative access may also deepen dependency on proprietary infrastructures, narrow stylistic diversity and shift

decision power away from distributed creative teams. The central issue is therefore no longer whether GAI will transform non-interactive audiovisual production, but which governance model will shape it, concentration-oriented or plural and accountable. Our position is that technical progress should be paired with minimum safeguards for transparency, attribution and reproducibility: systematically reporting datasets, prompts and configurations, combining quantitative metrics with narrative evaluation, and validating in real production environments.

The threats to the validity of these findings, and the mitigations adopted for each, are reported in Section 5.3.

Future research should extend the temporal scope to the 2025-2026 biennium and develop a formal taxonomy to standardize the classification of tools and generated content, facilitating comparative benchmarking. A complementary direction, suggested by the computing-centred bias of this corpus, is to replicate the mapping over film studies, arts and communication venues, so that the technical reading offered here can be confronted with the analytical vocabulary of the fields that actually produce linear audiovisual work. Future replications may also adopt the PRISMA 2020 reporting guideline [34] alongside the guidelines followed here, as a way of standardizing the reporting of the selection process.

7 ETHICAL CONSIDERATIONS

This work is a Systematic Mapping Study analyzing only previously published scientific articles. It involved neither human participants nor the collection or analysis of personal data; therefore, no Research Ethics Committee submission was required, in accordance with Brazilian National Health Council Resolution CNS No. 510/2016.

The dataset holds only bibliographic metadata of the 60 primary studies, their classification against the twenty-one extraction fields, and excerpts from the published articles. No personal data of third parties were collected or stored, so no anonymization applied. It is openly available in the repository cited in Section 4; the full texts are not redistributed, respecting IEEE Xplore and ACM Digital Library licensing terms.

ACKNOWLEDGMENTS

The authors thank the Graduate Program in Informatics (PPGInf) of the Federal University of Paraná for the WebMedia 2026 travel grant via PROEX. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

The authors used generative AI tools for language revision and for a verification pass over the reference list, checked against Crossref and the original articles, and over the internal consistency of the reported frequencies, and take full responsibility for the final content. No generative tool took part in the search, screening, data extraction or classification of the primary studies, all carried out by the four researchers.

REFERENCES

- [1] Govind A, Ahad Anzar, Aiswarya Anil Nair, and Rohith Syam. 2024. GenAI Empowered Script to Storyboard Generator. In *2024 IEEE International Conference on Future Machine Learning and Data Science (FMLDS)*. 451–456. doi:10.1109/FMLDS638.05.2024.00085
- [2] Victor R Basili and H Dieter Rombach. 1988. The TAME project: Towards improvement-oriented software environments. *IEEE Transactions on Software Engineering* 14, 6 (1988), 758–773. doi:10.1109/32.6156
- [3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language Models are Few-Shot Learners. *arXiv preprint arXiv:2005.14165* (2020). doi:10.48550/arXiv.2005.14165
- [4] Thiago Prado de Campos, Eduardo Filgueiras Damasceno, and Natasha Malveira Costa Valentim. 2022. Porifera: A Collaborative Tool to Support Systematic Literature Review and Systematic Mapping Study. In *Proceedings of the XXXVI Brazilian Symposium on Software Engineering*. ACM, New York, NY, USA, 452–457. doi:10.1145/3555228.3555273
- [5] Valerio Capraro, Austin Lentsch, Daron Acemoglu, Selin Akgun, Aisel Akhmedova, Ennio Bilancini, Jean-Francois Bonnefon, Pablo Branas-Garza, Luigi Butera, Karen M. Douglas, et al. 2024. The impact of generative artificial intelligence on socioeconomic inequalities and policy making. *PNAS Nexus* 3, 6 (2024), pgae191. doi:10.1093/pnasnexus/pgae191
- [6] Qi Chen, Minghui Tan, Yuankai Qi, Jiaqi Zhou, Yuanqing Li, and Qi Wu. 2022. V2C: Visual Voice Cloning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 21210–21219. doi:10.1109/CVPR52688.2022.02056
- [7] Yongkang Cheng, Mingjiang Liang, Shaoli Huang, Gao Han, Jifeng Ning, and Wei Liu. 2025. Conditional GAN for Enhancing Diffusion Models in Efficient and Authentic Global Gesture Generation from Audios. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2164–2173. doi:10.1109/WACV61041.2025.00217
- [8] Michel Chion. 2019. *Audio-Vision: Sound on Screen* (2nd ed.). Columbia University Press, New York. doi:10.7312/chio18588
- [9] Prafulla Dhariwal and Alex Nichol. 2021. Diffusion Models Beat GANs on Image Synthesis. *arXiv preprint arXiv:2105.05233* (2021). doi:10.48550/arXiv.2105.05233
- [10] Yogesh K. Dwivedi, Nir Kshetri, Laurie O. Hughes, Emma Louise Slade, Anand Jeyaraj, Arpan Kumar Kar, Abdullah M. Baabdullah, Alex Koohang, Vishnu Priya Raghavan, Manju Ahuja, Hanaa Albanna, Mousa Ahmad Albashrawi, Adil S. Al-Busaidi, Janarthanan Balakrishnan, Yves Barlette, Somprakash Basu, Indranil Bose, Laurence Brooks, Dimitrios Buhalis, Lemuria Carter, Rahul De, Yanqing Duan, Denis Dennehy, Rameshwar Dubey, Rohita Dwivedi, John S. Edwards, Carlos Flaviano, et al. 2023. Opinion Paper: “So what if ChatGPT wrote it?” Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management* 71 (2023), 102642. doi:10.1016/j.ijinfomgt.2023.102642
- [11] Farshad Einabadi, Jean-Yves Guillemaut, and Adrian Hilton. 2021. Deep Neural Models for Illumination Estimation and Relighting: A Survey. *Computer Graphics Forum* 40, 6 (2021), 315–331. doi:10.1111/cgf.14283
- [12] Oscar Estrada Torrejon, Nicholas Peretti, and Ricardo Figueroa. 2020. Rotoscope Automation with Deep Learning. *SMPTE Motion Imaging Journal* 129, 2 (2020), 16–26. doi:10.5594/JMI.2019.2959967
- [13] Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76, 5 (1971), 378–382. doi:10.1037/h0031619
- [14] Luciano Floridi and Massimo Chiriatti. 2020. GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds and Machines* 30, 4 (2020), 681–694. doi:10.1007/s11023-020-09548-1
- [15] Erick Giovanni Franco Lazarte, Maria Trinidad Juárez Pacco-taipe, Rosa Candelaria Ramirez Heredia, and Teresa Marlene Vela Loyola. 2025. Audiovisual Narrative in the Age of Artificial Intelligence: Advances, Trends and Challenges: A Systematic Review. *Journal of Educational and Social Research* 15, 1 (2025), 1. doi:10.36941/jesr-2025-0001
- [16] Jorge Franganillo. 2022. Contenido generado por inteligencia artificial: oportunidades y amenazas. *Anuario ThinkEPI* 16 (2022). doi:10.3145/thinkepi.2022.e16a24
- [17] Leonardo Gambacorta, Tullio Jappelli, and Tommaso Oliviero. 2025. *Exploring Household Adoption and Usage of Generative AI: New Evidence from Italy*. BIS Working Paper 1298. Bank for International Settlements. <https://www.bis.org/publ/work1298.htm>
- [18] Iryna Gavran, Serhii Honcharuk, Volodymyr Mykhalov, Kateryna Stepanenko, and Natalia Tsimokh. 2025. The Impact of Artificial Intelligence on the Production and Editing of Audiovisual Content. *PDT&C* 54, 3 (July 2025), 223–235. doi:10.1515/pdtc-2025-0022
- [19] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, Vol. 27. 2672–2680. doi:10.48550/arXiv.1406.2661
- [20] Grand View Research. 2025. Generative AI Market Size And Share | Industry Report, 2030. Industry Report, Grand View Research. <https://www.grandviewresearch.com/industry-analysis/generative-ai-market-report> Accessed November 2025.
- [21] Paolo Grigis and Antonella De Angeli. 2024. Playwriting with Large Language Models: Perceived Features, Interaction Strategies and Outcomes. In *Proceedings of the 2024 International Conference on Advanced Visual Interfaces (AVI)*. 1–9. doi:10.1145/3656650.3656688
- [22] Satoshi Iizuka and Edgar Simo-Serra. 2019. DeepRemaster: Temporal Source-Reference Attention Networks for Comprehensive Video Enhancement. *ACM Transactions on Graphics* 38, 6 (2019), 1–13. doi:10.1145/3355089.3356570
- [23] Dhiraj Jadhav, Sakshee Agrawal, Sakshi Jagdale, Pranav Salunkhe, and Raj Salunkhe. 2024. AI-Driven Text-to-Multimedia Content Generation: Enhancing Modern Content Creation. In *2024 8th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*. 1610–1615. doi:10.1109/I-SMAC61858.2024.10714771
- [24] Roy Kapon, Guy Tevet, Daniel Cohen-Or, and Amit H. Bermano. 2024. MAS: Multi-view Ancestral Sampling for 3D Motion Generation Using 2D Diffusion. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1965–1974. doi:10.1109/CVPR52733.2024.00192
- [25] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and Improving the Image Quality of StyleGAN. *arXiv preprint arXiv:1912.04958* (2020). doi:10.48550/arXiv.1912.04958
- [26] Barbara Kitchenham and Shari Charters. 2007. *Guidelines for Performing Systematic Literature Reviews in Software Engineering*. Technical Report TR/SE-0401. Keele University and Empirical Software Engineering, National ICT Australia Ltd., Keele, UK and Sydney, Australia. <https://www.cs.keele.ac.uk/research/se/Resources/KitchenhamCharters2007.pdf> Joint Technical Report.
- [27] J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33, 1 (1977), 159–174. doi:10.2307/2529310
- [28] Yunxin Li, Haoyuan Shi, Baotian Hu, Longyue Wang, Jiashun Zhu, Jinyi Xu, Zhen Zhao, and Min Zhang. 2024. Anim-Director: A Large Multimodal Model Powered Agent for Controllable Animation Video Generation. In *SIGGRAPH Asia 2024 Conference Papers*. 1–11. doi:10.1145/3680528.3687688
- [29] Tingting Liao, Hongwei Yi, Yuliang Xiu, Jiaxiang Tang, Yangyi Huang, Justus Thies, and Michael J. Black. 2024. TADA! Text to Animatable Digital Avatars. In *2024 International Conference on 3D Vision (3DV)*. 1508–1519. doi:10.1109/3DV62453.2024.00150
- [30] Chang Liu, Haoning Wu, Yujie Zhong, Xiaoyun Zhang, Yanfeng Wang, and Weidi Xie. 2024. Intelligent Grimm - Open-ended Visual Storytelling via Latent Diffusion Models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 6190–6200. doi:10.1109/CVPR52733.2024.00592
- [31] Jiayi Lv, Yi Huang, Mingfu Yan, Jiancheng Huang, Jianzhuang Liu, Yifan Liu, Yafei Wen, Xiaoxin Chen, and Shifeng Chen. 2024. GPT4Motion: Scripting Physical Motions in Text-to-Video Generation via Blender-Oriented GPT Planning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 1430–1440. doi:10.1109/CVPRW63382.2024.00150
- [32] Luisa Martins, Stênio Ferreira, and Tiago Maritan. 2025. Automatic 3D animation generation for Sign Language: A case study

- of Sign Language Machine Translation. In *Proceedings of the 31st Brazilian Symposium on Multimedia and the Web (WebMedia)*. Sociedade Brasileira de Computação, Rio de Janeiro, Brazil, 67–76. doi:10.5753/webmedia.2025.16135
- [33] Cihan Orak and Zeynep Turan. 2024. Using artificial intelligence in digital video production: A systematic review study. *Journal of Educational Technology & Online Learning* 7, 3 (Sept. 2024), 286–307. doi:10.31681/jetol.1459434
- [34] Matthew J. Page, Joanne E. McKenzie, Patrick M. Bossuyt, Isabelle Boutron, Tammy C. Hoffmann, Cynthia D. Mulrow, et al. 2021. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 372 (2021), n71. doi:10.1136/bmj.n71
- [35] Kai Petersen, Robert Feldt, Shahid Mujtaba, and Michael Mattsson. 2008. Systematic Mapping Studies in Software Engineering. In *12th International Conference on Evaluation and Assessment in Software Engineering (EASE)*. 68–77. doi:10.14236/ewic/EA SE2008.8
- [36] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. doi:10.48550/arXiv.2112.10752
- [37] Stuart J. Russell and Peter Norvig. 2016. *Artificial Intelligence: A Modern Approach* (3rd ed.). Pearson Education, Upper Saddle River, NJ. <https://aima.cs.berkeley.edu/>
- [38] Simone de Oliveira Santos, Robson de Moraes Alves, Francisco Carlos de Mattos Brito Oliveira, and Francisco Ivanilso Soares Araujo. 2025. Investigating LLM-based tools to support Usability, Accessibility, User Experience in HCI activities: A Systematic Literature Mapping. In *Proceedings of the 31st Brazilian Symposium on Multimedia and the Web (WebMedia)*. Sociedade Brasileira de Computação, Rio de Janeiro, Brazil, 631–645. doi:10.5753/webmedia.2025.16160
- [39] Ming Shen, Gang Huang, Yuxuan Wu, Shuyi Song, Sheng Zhou, Liangcheng Li, Zhi Yu, Wei Wang, and Jiajun Bu. 2024. Making Accessible Movies Easily: An Intelligent Tool for Authoring and Integrating Audio Descriptions to Movies. In *Proceedings of the 21st International Web for All Conference (W4A '24)*. 160–164. doi:10.1145/3677846.3677855
- [40] Junrong Song and David Yip. 2023. Exploring the Intersection of AI Art and Film: A Case Study of Giant. In *2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*. 347–352. doi:10.1109/ICMEW59549.2023.00066
- [41] Aniruddha Tiwari, Rushit Dave, and Mounika Vanamala. 2023. Leveraging Deep Learning Approaches for Deepfake Detection: A Review. *arXiv preprint* (April 2023). arXiv:2304.01908 [cs.LG]
- [42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems*. doi:10.48550/arXiv.1706.03762
- [43] Dan Valeriu Voinea. 2025. The Algorithmic Auteur: AI, Cultural Production, and the Reconfiguration of Audiovisual Media. *Social Sciences and Education Research Review* 12, 1 (July 2025), 256–268. doi:10.5281/zenodo.15804554
- [44] Sitong Wang, Samia Menon, Tao Long, Keren Henderson, Dingzeyu Li, Kevin Crowston, Mark Hansen, Jeffrey V Nickerson, and Lydia B Chilton. 2024. ReelFramer: Human-AI Co-Creation for News-to-Video Translation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–20. doi:10.1145/3613904.3642868
- [45] Christoph Johannes Weber, Sebastian Burgkart, and Sylvia Rothe. 2024. wr-AI-ter: Enhancing Ownership Perception in AI-Driven Script Writing. In *Proceedings of the 2024 ACM International Conference on Interactive Media Experiences (IMX '24)*. 145–156. doi:10.1145/3639701.3656325
- [46] Roel J. Wieringa, Neil A. M. Maiden, Nancy R. Mead, and Colette Rolland. 2006. Requirements engineering paper classification and evaluation criteria: a proposal and a discussion. *Requirements Engineering* 11, 1 (2006), 102–107. doi:10.1007/s00766-005-0021-6
- [47] Weimei Wu, Jianhua Hu, Yingjie Huang, Wenying Zeng, and Hui Shao. 2024. Intelligent Teaching Platform Based on Large Models. In *Proceedings of the 2024 International Symposium on Artificial Intelligence for Education*. 409–415. doi:10.1145/3700297.3700367
- [48] Zhiyu Yin, Kehai Chen, Xuefeng Bai, Ruili Jiang, Juntao Li, Hongdong Li, Jin Liu, Yang Xiang, Jun Yu, and Min Zhang. 2025. A Survey: Spatiotemporal Consistency in Video Generation. *arXiv preprint arXiv:2502.17863* (2025). doi:10.48550/arXiv.2502.17863
- [49] Jingbo Zhang, Xiaoyu Li, Ziyu Wan, Can Wang, and Jing Liao. 2024. Text2NeRF: Text-Driven 3D Scene Generation With Neural Radiance Fields. *IEEE Transactions on Visualization and Computer Graphics* 30, 12 (2024), 7749–7762. doi:10.1109/TVCG.2024.3361502