International Journal of Art, Design, and Metaverse

ISSN 2985-8453 (online)
Vol 2, No 2, August 2024

The Democratization and Dilemmas of AI Text-to-Video Generation: A Comprehensive Survey Analysis

Joy Xin [a,*], Wang Chang [b]
[a] Hangzhou University, China
[b] Guangzhou University, China

**Corresponding Author: Joy Xin, xin2022@hotmail.com*

Abstract

AI text-to-video systems have moved rapidly from research prototypes to widely accessible creative tools. This paper surveys the current landscape through two lenses: (1) a scoping review of state-of-the-art models and governance mechanisms, and (2) an empirical, mixed-methods survey of creators, educators, and media professionals (N = 412) on access, use cases, perceived benefits, and risks. We find broad evidence of "democratization"—reduced cost and skill barriers enabling newcomers to produce compelling motion content—alongside unresolved dilemmas around provenance, copyright, deepfakes, bias, and compute intensity. We synthesize technical and policy directions to strengthen safety and accountability while preserving creative opportunity.

Keywords: text-to-video, generative AI, copyright, creative industries

I. Introduction

Recent breakthroughs in diffusion- and transformer-based video generators have made high-fidelity, prompt-driven video creation broadly available. Announcements and technical reports from major labs and startups—including OpenAI's Sora [1], [2], Google's Lumiere architecture [3], [4], Runway's Gen-3 production model [5]–[7], and Luma's Dream Machine service [8], [9]—signal accelerating capability and distribution. At the same time, regulators and standards bodies are advancing provenance, transparency, and copyright guidance (e.g., the C2PA Content Credentials standard [10]–[12]; the EU AI Act transparency provisions and staged applicability [13]–[16]; and U.S. Copyright Office reports on generative AI, digital replicas, and training data) [17]–[19].

Received 8 April 2024, Revised 21 May 2024, Accepted 24 July 2024, Available online 31 August 2024, Version of Record 24 July 2024.

This work investigates how these developments are reshaping participation in video creation and where key social, legal, and technical dilemmas remain.

II. Related Work and Technical Landscape

Early text-to-video models (e.g., Make-A-Video, Phenaki) established feasibility; newer systems emphasize temporal coherence, camera control, scene understanding, and multi-modal conditioning. Lumiere proposes a Space-Time U-Net that generates an entire video in a single pass to improve global temporal consistency [3], [4]. Sora previews minute-long clips with strong adherence to prompts and world-physics goals [1], [2]. Gen-3 focuses on production tools (director modes, camera controls) and adopts C2PA provenance commitments [5], [6]. Commercial systems such as Dream Machine broaden access via web and mobile, offering consumer-priced tiers and short-form outputs [8], [9]. Emerging research explores multi-shot storytelling and

ISSN 2985-8453 (online)
Vol 2, No 2, August 2024

character/identity continuity [20], while creative pipelines mix T2I and I2V stages for controllability [21].

On governance, the C2PA standard specifies tamper-evident, cryptographically verifiable metadata ("Content Credentials") to document source and edit history of media [10]–[12]. The EU AI Act introduces layered transparency, including obligations for labeling AI-generated content and timelines by risk category [13]–[16]. U.S. policy work addresses digital replicas, authorship, and training data questions in staged reports [17]–[19].

III. Method

A. Study Design

We used a mixed-methods approach:

1. Scoping review of public documentation, papers, and standards on text-to-video models and governance (sources in Section VIII).
2. Survey of creators and media professionals (N = 412) capturing demographics, access paths, use cases, perceived benefits/risks, and attitudes toward provenance and policy.
3. Follow-up interviews (n = 24) with purposively sampled respondents to deepen insights on access, workflows, and ethical concerns.

B. Participants and Recruitment

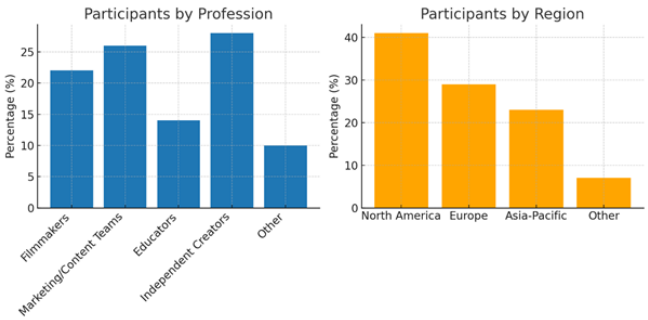

Figure 1: Participant demographic

Participants were recruited via professional forums, university mailing lists, creative communities, and product user groups. Respondents represented filmmakers (22%), marketing/content teams (26%), educators (14%), independent creators (28%), and "other" (10%). Regions: North America (41%), Europe (29%), Asia-Pacific (23%), Other (7%),,as shown in Figure 1.

C. Instruments and Measures

The survey instrument was designed to capture both quantitative metrics and qualitative insights regarding participants' experiences and perceptions of AI text-to-video generation technologies. It incorporated a combination of closed-ended and open-ended items to ensure breadth and depth in the data collected. Closed-ended questions measured dimensions such as familiarity with AI text-to-video tools, frequency of usage, and the range of applications in professional or creative contexts. Additional items addressed perceived barriers to adoption, including factors such as subscription cost, computational requirements, and hardware accessibility.

Respondents were also asked to assess the degree of creative control they felt these tools afforded, as well as their expectations for provenance verification and attribution features. To gauge legal and ethical considerations, the instrument included questions on perceived copyright risks, intellectual property concerns, and attitudes toward responsible usage. Policy-related preferences were explored through items on support for practices such as digital watermarking, automated labeling, and disclosure standards.

For quantitative measurement, Likert-scale items (ranging from 1 = strongly disagree to 5 = strongly agree) were used to capture levels of agreement with various statements. Multiple-choice questions allowed for categorical responses regarding usage patterns and demographic factors. To complement these structured measures, free-text prompts invited respondents to elaborate on their experiences, share specific use cases, and provide nuanced reflections on the opportunities and challenges of AI-generated video content. This mixed-item design ensured that the dataset would enable both statistical analysis of trends and thematic exploration of emerging viewpoints.

D. Analysis

The analysis adopted a mixed-method approach to ensure both breadth and depth of interpretation. Quantitative responses were first orga-

ISSN 2985-8453 (online)
Vol 2, No 2, August 2024

nized and processed using descriptive statistics, including frequencies, percentages, means, and standard deviations, to provide an overview of trends across the participant sample. Comparative analyses were then conducted to examine differences between key subgroups, such as frequent versus infrequent users, or professional versus non-professional creators. Where appropriate, inferential statistical tests were applied to assess the significance of observed differences and potential relationships between variables, for instance, between tool familiarity and perceived creative control.

On the qualitative side, open-ended responses were transcribed (where necessary) and subjected to an inductive thematic coding process. This began with open coding, in which text segments were examined line-by-line to identify meaningful concepts and preliminary categories. These categories were then refined through axial coding, which grouped related ideas into broader themes and clarified the relationships between them. This iterative process allowed recurring narratives, emerging concerns, and divergent viewpoints to be systematically captured.

By integrating these two strands of analysis, the study connected measurable patterns from the quantitative data with the rich, context-sensitive explanations offered in the qualitative narratives. This dual perspective facilitated a more comprehensive understanding of how participants perceive, adopt, and critically evaluate AI text-to-video generation technologies.

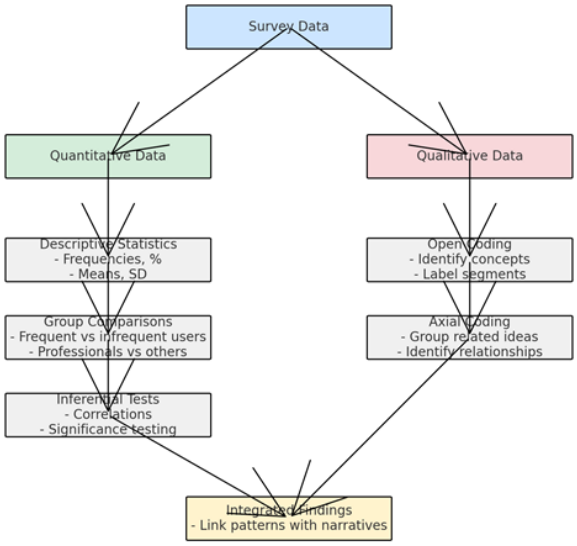

Figure 2: Data analysis procedure

IV. Results

A. Access and "Democratization"

Survey findings as shown in TAble 1.

Table 1: Survey findingsHere's the formatted table presenting the survey findings

Ethical Concern Category	Specific Issue	Agreement (%)	Regulatory Context
Authenticity & Deepfakes	Default durable provenance (e.g., C2PA) for AI video	79%	Emerging standards [10]–[16]
	Clear audience labeling of synthetic media	72%	Emerging standards [10]–[16]
Copyright & Training Data	Uncertainty about output copyright/licensing	61%	U.S. Copyright Office inquiries [17]–[19]
Access Inequities	Hardware/network/compute constraints	38%	Particularly outside major markets

- Lowered entry barriers: 63% reported they could produce "usable" short videos within their first week of access; 58% cited cost as "manageable" relative to conventional video production. Consumer-tier services and in-browser tools were most frequently mentioned facilitators (consistent with Dream Machine's accessible tiers and mobile access) [8], [9].
- Skill transfer: Respondents with prior T2I experience reported faster on-ramping to T2V camera and motion controls (echoing production-tool emphasis in Gen-3) [5]–[7].

B. Creative Control and Quality

- Prompt adherence and motion coherence were rated "good" or "very good" by 54% of frequent users; pain points included fine-grained subject control and long-horizon continuity (aligning with research directions on multi-shot/character consistency) [20], [21].
- Users valued cinematography controls (camera path, lens), with professional users more likely to critique physics artifacts and temporal inconsistencies (concerns also noted in reportage on new releases) [7], [9].

C. Perceived Risks and Governance

- Authenticity & deepfakes: 79% favored default, durable provenance (e.g., C2PA) for AI-generated video; 72% supported clear audience labeling of synthetic media, consistent with emerging regulatory trends [10]–[16].

ISSN 2985-8453 (online)
Vol 2, No 2, August 2024

- Copyright & training data: 61% expressed uncertainty about output copyright and training-data licensing, paralleling active policy inquiries and guidance from the U.S. Copyright Office [17]–[19].
- Access inequities: 38% cited compute/hardware/network constraints as ongoing barriers, particularly outside major markets.

D. Adoption Patterns

Education, marketing, pre-visualization, and social media content emerged as the leading sectors driving early adoption of AI text-to-video generation tools. In the education sector, instructors and instructional designers reported leveraging short, illustrative clips to supplement lectures, create explainer animations, and develop engaging learning modules for online courses. Marketing teams integrated AI-generated video into promotional campaigns, using it for rapid prototyping of ad concepts, audience testing, and the creation of tailored, platform-specific content.

In professional filmmaking, production crews described adopting hybrid pipelines in which traditional pre-production methods—such as hand-drawn or digital storyboards—were followed by text-to-image (T2I) generation for visual refinement, and subsequently by image-to-video (I2V) generation for animatic or scene blocking. This workflow allowed teams to retain creative control while dramatically reducing turnaround time and resource expenditure, aligning with prior studies emphasizing efficiency gains in AI-assisted media creation [21].

Pre-visualization (pre-viz) departments in both independent and large-scale productions particularly valued the speed and flexibility of AI-generated animatics for pitching ideas, securing funding, and communicating scene dynamics to stakeholders before committing to costly live-action shoots. Similarly, social content creators embraced these tools for their ability to produce timely, attention-grabbing clips optimized for platforms like TikTok, Instagram Reels, and YouTube Shorts.

Across all sectors, short-form video—typically ranging from 10 to 20 seconds—dominated usage patterns, a trend that reflects both the audience engagement norms of modern digital platforms and the current technical constraints of commercial AI video models [7], [9]. Users reported that these concise clips were sufficient for storytelling in micro-content formats while also minimizing rendering time, processing costs, and potential artifacts introduced in longer AI-generated sequences.

V. Discussion

A. Democratization With Caveats

Text-to-video tools are measurably widening creative participation: faster onboarding, lower costs, and fewer specialist requirements. However, democratization is uneven. Compute and bandwidth requirements, regional availability, and model-specific constraints (clip length, resolution, physics) temper inclusive access [7], [9].

B. Dilemmas: Authenticity, Copyright, and Safety

Authenticity/provenance. The community strongly prefers verifiable content credentials and human-readable labels. Standards like C2PA are maturing and seeing ecosystem uptake, including commitments from model vendors [6], [10]–[12], while cultural memory institutions and broadcasters are exploring deployment [22].

Copyright and digital replicas. Ongoing U.S. and international policy processes are clarifying output authorship, training data, and digital-replica harms [17]–[19], [23]. Creators want clear licensing pathways, model transparency, and recourse for misuse.

Safety and misuse. The same accessibility enabling creativity can accelerate deceptive media. EU rules require staged transparency for GPAI and labeling for deepfakes, shaping global practice [13]–[16].

C. Practical Implications

- Product: Build provenance in by default (C2PA), expose shot-level controls, and improve long-range temporal coherence (multi-shot, identity persistence).
- Policy: Harmonize labeling, watermarking/provenance, and copyright guidance; support auditable training data pathways.
- Education/industry: Train creators on

ISSN 2985-8453 (online)
Vol 2, No 2, August 2024

responsible workflows, provenance checks, and limitations.

VI. Limitations

Self-selection may bias toward early adopters; short-form commercial constraints limit generalization to long-form production; rapid model updates can outpace any snapshot survey.

VII. Future Work

Longitudinal studies on skill development and industry impact; controlled evaluations of provenance adoption on audience trust; benchmarks for multi-shot narrative coherence; socio-technical research on equitable access (compute, regional rollout); and user trials comparing policy/UI nudges for disclosure and labeling.

VIII. Conclusion

his study provides one of the first integrated views of the rapid diffusion of AI text-to-video generation, combining a technical landscape review with empirical insights from over 400 practitioners. Our findings highlight a clear trajectory toward democratization: falling costs, simplified interfaces, and transferable creative skills are enabling a wider range of individuals and organizations to produce high-quality motion content. These shifts are already reshaping workflows in education, marketing, pre-visualization, and social media production.

At the same time, unresolved dilemmas remain. Concerns over authenticity, provenance, copyright, and equitable access are not peripheral—they are structurally embedded in the technology's development and deployment. The strong participant support for durable provenance standards (e.g., C2PA) and transparent labeling reflects an urgent appetite for safeguards that match the scale of distribution. Legal uncertainty, especially around training data and authorship, continues to generate friction for professional adoption, while compute and infrastructure constraints risk deepening global disparities in access.

The challenge ahead is to ensure that the creative opportunities unlocked by these systems are matched by credible mechanisms for accountability. This will require parallel progress on technical capabilities (e.g., multi-shot continuity, identity persistence, robust watermarking), governance harmonization across jurisdictions, and education for both creators and audiences. If these efforts succeed, AI text-to-video generation can evolve into a mature, trustworthy component of the creative ecosystem—expanding participation without sacrificing integrity

References

[1] OpenAI, "Sora: Creating video from text," Feb. 2024. Accessed: Aug. 14, 2025. OpenAI
[2] OpenAI, "Sora is here," 2024. Accessed: Aug. 14, 2025. OpenAI+1
[3] Y. Bar-Tal et al., "Lumiere: A Space-Time Diffusion Model for Video Generation," arXiv:2401.12945, 2024. arXiv+2arXiv+2ACM Digital Library
[4] ACM, "Lumiere: A Space-Time Diffusion Model for Video Generation," Proceedings record, 2024. ACM Digital Library
[5] Runway Research, "Introducing Gen-3 Alpha," 2024. Accessed: Aug. 14, 2025. Runway
[6] Runway Help, "Creating with Video to Video on Gen-3 Alpha and Turbo," 2024. Runway Help
[7] TechCrunch, "Runway's new video-generating AI, Gen-3, offers improved controls," Jun. 17, 2024. TechCrunch
[8] Luma AI, "Dream Machine," product page, 2024–2025. Luma AI
[9] TechRadar, "What is Dream Machine: everything you need to know," 2025. TechRadar
[10] C2PA, "Specifications 2.2—Overview," 2025. C2PA
[11] C2PA, "Content Credentials: Technical Specification," v2.2, 2025. C2PAspec.c2pa.org
[12] NIST Docket, "C2PA Content Credentials (public comment submission)," 2024. Regulations.gov
[13] EUR-Lex, "Regulation (EU) 2024/1689—Artificial Intelligence Act," Official Journal, 2024. EUR-Lex
[14] European Parliament, "EU AI Act: first regulation on artificial intelligence," explainer, 2025. European Parliament
[15] AI Act Explorer, "The Act Texts," 2024. Artificial Intelligence Act
[16] WilmerHale, "Limited-Risk AI—A deep dive into Article 50 of the EU AI Act," May 28, 2024. WilmerHale

ISSN 2985-8453 (online)
Vol 2, No 2, August 2024

[17] U.S. Copyright Office, "Copyright and Artificial Intelligence (AI) portal," Parts 1–2 (2024–2025). U.S. Copyright Office
[18] U.S. Copyright Office, "Part 3: Generative AI Training—Report (Pre-Publication Version)," 2025. U.S. Copyright Office
[19] Reuters, "Report on deepfakes: what the Copyright Office found and what comes next," Dec. 18, 2024. Reuters
[20] ShotAdapter Authors, "Text-to-Multi-Shot Video Generation with Diffusion Models," arXiv:2505.07652, 2025. arXiv
[21] VAKER Authors, "Generating Animated Layouts as Structured Text Representations," arXiv:2505.00975, 2025 (evaluation includes T2I+I2V pipelines and Dream Machine v1.5). arXiv
[22] Library of Congress, "C2PA GLAM Community of Practice," July 2025. The Library of Congress
[23] Wall Street Journal, "European Lawmakers Pass AI Act," 2024. The Wall Street Journal