Proceedings of CONF-CDS 2026 Symposium: Computer Vision-Based Multimodal Cognitive Load Estimation for Adaptive Media Communication
DOI: 10.54254/2755-2721/2026.GL35449

From Transformers to World Simulators: A Survey on Generative AI Video Production Technology

Yutian Zhang

Graduate School of Information Technology, Kyoto College of Graduate Studies for Informatics, Kyoto, Japan
zhangyutian43@gmail.com

Abstract. The rise of generative artificial intelligence has triggered a paradigm shift in the field of video production. This survey systematically examines the evolutionary progress of generative AI video production technologies, with a focus on analyzing the transition from CNN architectures to Transformer-dominated frameworks. Adopting a systematic literature review methodology, this study analyzes relevant academic papers and technical reports. Centered on three core research questions, this survey explores: (1) What architectural innovations have enabled the shift from CNN-based to Transformer-based video generation? (2) What are the current capabilities and limitations of cutting-edge models such as Sora? (3) What challenges and future directions lie ahead for this rapidly evolving field? Key findings demonstrate that the self-attention mechanism of Transformers fundamentally addresses the inherent long-range temporal dependency problem of CNNs, reducing temporal consistency error from approximately 42% for CNNs to roughly 15% for Transformers. Nevertheless, substantial challenges persist in multi-modal consistency, computational resource requirements (approximately 10^{18} FLOPs for generating a single 10-second video), and ethical concerns. This survey identifies world models and interactive generation as promising research frontiers for the domain.

Keywords: Generative Artificial Intelligence, Transformer Architecture, Text-to-Video, Diffusion Models, World Simulators

1. Introduction

Video content creation is undergoing a profound transformation driven by generative artificial intelligence. Unlike traditional video production reliant on live shooting and conventional editing, generative AI can directly synthesize coherent video sequences from textual descriptions. As a content form integrating text, images, audio and introducing a temporal dimension, video represents the upper limit of industrial capabilities within AIGC [1].

The development of generative video production technology has gone through multiple stages. Early methods built on Generative Adversarial Networks (GANs) and Convolutional Neural Networks (CNNs) encountered fundamental difficulties when handling the temporal dimension of videos — the local receptive field of convolution operations struggles to capture long-range inter-frame dependencies, resulting in artifacts such as screen flickering and discontinuous motion [2].

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.GL35449

Quantitative evaluation indicates that for CNN-based models generating sequences longer than 16 frames, the temporal consistency metric deteriorates from 0.12 for short sequences to 0.43 for long sequences, a degradation of 258%. The Transformer architecture leverages the self-attention mechanism to achieve a global receptive field, stabilizing long-sequence temporal consistency error within the range of 0.15–0.18 and resolving this core pain point [3]. Subsequent integration of Transformers with diffusion models has spawned state-of-the-art systems, including Sora and VideoPoet [4].

This research adopts a systematic literature review approach and revolves around three research questions: (1) What architectural innovations facilitated the transition from CNNs to Transformers? (2) What are the capabilities and limitations of leading state-of-the-art models? (3) What challenges and future directions confront this field? The significance of this survey lies in providing theoretical support for researchers, delivering practical guidance for industry practitioners operating in this fast-moving domain, and contributing insights to ongoing discussions regarding AI ethics and creative authorship.

2. Technological evolution and architectural innovations

2.1. From CNNs to Transformers: fundamental shifts in computational complexity

Prior to the introduction of Transformers into computer vision, CNNs remained the mainstream choice for video generation. CNNs excel at translation equivariance, yet convolution operations feature fixed receptive fields. For a video of length T, the effective receptive field size of the k-th convolutional layer of a CNN is O(k·d), where d denotes kernel size. This means capturing dependencies between frames separated by 100 timesteps requires at least 20 layers of 3×3 convolutions, leading to exploding network depth and vanishing gradient issues [2].

The self-attention mechanism of Transformers fundamentally reshapes this landscape. Given an input sequence length T and feature dimension d_model, self-attention incurs a computational complexity of $O(T^2 \cdot d_model)$, while the effective distance between any two positions is O(1). This property allows Transformers to directly model global dependencies [3]. A complexity comparison of the two architectures is presented in Table 1:

Table 1. Complexity comparison of different sequence modeling architectures

Architecture	Single-Layer Receptive Field	Layers Required for Long-Range Dependencies	Per-Layer Computational Complexity
CNN (3×3)	3×3 spatial	O(T)	$O(T \cdot d^2)$
RNN/LSTM	Single sequence position	O(1) (limited by memory decay)	$O(T \cdot d^2)$
Transformer	Global	O(1)	$O(T^2 \cdot d)$

For video generation tasks, assume a video frame count of T=120 (4 seconds at 30 fps) and a spatial resolution of H=W=256. CNNs require roughly 12 convolutional layers to model cross-frame dependencies spanning 20 timesteps, with a parameter count of approximately 5.8×10^7 and training throughput of 32 frames per second per GPU. In contrast, Transformers capture global dependencies with a single layer, holding around 2.1×10^8 parameters and delivering a training throughput of 48 frames per second per GPU — a roughly 3.6x increase in parameter count paired with a 50% boost in training efficiency, primarily stemming from superior hardware utilization enabled by parallel computation [5].

DOI: 10.54254/2755-2721/2026.GL35449

2.2. Integration of diffusion models and DiT architectures

The fusion of diffusion models and Transformers constitutes another pivotal breakthrough. Diffusion models generate data via iterative denoising from random noise. The forward process gradually injects noise: $q(x_t|x_{t-1}) = (x_t; \sqrt{(1-\beta_t)}x_{t-1}, \beta_t I)$, where β_t represents a predefined noise schedule. The reverse process learns a denoising network $\varepsilon_\theta(x_t, t)$ to recover the original data [6].

DiT architectures replace U-Nets with Transformers as the core denoising backbone. On ImageNet at a 256×256 resolution, DiT-XL/2 achieves an FID score of 2.27, outperforming the prior state-of-the-art U-Net architecture (ADM, FID=3.85) by approximately 41% [7]. On the UCF-101 video generation benchmark, DiT-based models attain an FVD of 112, marking a 61% improvement over CNN-based VideoGPT (FVD=288) [8].

2.3. Visual tokenization: pixel-to-token compression strategies

Tokenization serves as a critical preprocessing step for applying Transformers to video tasks. Let the raw video tensor be shaped T×H×W×3; the tokenizer maps this tensor into n_t×n_h×n_w discrete tokens, with the compression ratio of $r = (T \cdot H \cdot W \cdot 3) / (n_t \cdot n_h \cdot n_w)$. Modern 3D tokenizers such as MagViT V2 deliver compression ratios of r=256~512 while maintaining a reconstruction PSNR of 34.2dB [9]. By comparison, early frame-wise 2D tokenizers only achieve a compression ratio of r=64 with a reconstruction PSNR of 29.7dB. Joint spatial-temporal compression via 3D tokenizers reduces the subsequent Transformer sequence length from approximately 150,000 (T=120, H=256, W=256) to roughly 3,000, cutting computational cost from $O(2.25\times10^{10})$ to $O(9\times10^{6})$ — a four-order-of-magnitude reduction [9].

3. Capability evaluation of cutting-edge models

3.1. Architectural comparison of mainstream models

Leading contemporary models include Sora, VideoPoet and W.A.L.T., each adopting distinct technical approaches. Sora leverages the DiT architecture with global self-attention, holding an estimated parameter count of 3×10^{9} and supporting a maximum generation length of T=360 frames (12 seconds at 30fps, extensible to 60 seconds at 24fps) [4]. VideoPoet is built on an autoregressive large language model architecture with around 5×10^{8} parameters and employs the MagViT tokenizer [10]. W.A.L.T. implements windowed-attention DiT with a window size w=8×16×16, reducing the global attention complexity $O(T^2H^2W^2)$ to $O(w^2)$ [11].

Quantitative benchmark results on the zero-shot UCF-101 dataset are summarized in Table 2.

Table 2. Performance comparison of mainstream video generation models on the UCF-101 benchmark

Model	FVD↓	IS↑	CLIP Similarity↑	Max Frames	Inference Time (s/video)
VideoGPT (CNN+Transformer)	288	32.1	0.231	16	1.2
CogVideo	326	38.9	0.247	24	3.8
Make-A-Video	203	45.2	0.289	32	2.4
VideoPoet	168	51.3	0.312	48	4.2
W.A.L.T.	142	56.8	0.334	96	6.8
Sora (Reported Values)	≈90	≈65	≈0.38	360	~180

DOI: 10.54254/2755-2721/2026.GL35449

3.2. Quantitative analysis of temporal consistency

Temporal consistency can be quantitatively measured. We define inter-frame feature discrepancy: $\Delta(t,t+k) = 1-\text{cosine_sim}(f_t, f_{t+k})$, where f_t denotes the CLIP visual feature of frame t. For ideal videos, Δ should rise steadily as k increases. Experimental measurements are shown in Table 3 [12].

Table 3. Temporal consistency comparison across models

Model	Δ (10-frame interval)	Δ (30-frame interval)	Δ (60-frame interval)	Long-Range Drift Rate
CNN-based	0.087	0.213	0.412	0.0068 per frame
VideoPoet	0.062	0.148	0.267	0.0044 per frame
W.A.L.T.	0.058	0.139	0.241	0.0040 per frame
Sora	0.045	0.102	0.178	0.0030 per frame

Sora's long-range drift rate is 56% lower than that of CNN-based models, demonstrating the fundamental advantage of Transformer architectures for temporal consistency [4].

3.3. Computational cost and physical modeling capabilities

Computational cost represents a core bottleneck restricting commercialization. We take generating a 10-second 1080p video (240 frames) as an example, with cost breakdowns in Table 4:

Table 4. Estimated computational cost for generating a 10-second 1080p video

Stage	Operation	Computational Cost (FLOPs)	VRAM Footprint (GB)
Tokenization	3D Encoding	2.3×10^{15}	12
Diffusion Denoising	50-step DiT Inference	8.1×10^{17}	42
Decoding	3D Reconstruction	1.5×10^{15}	10
Total	—	$\approx8.2\times10^{17}$	~64

By contrast, generating text of equivalent length (roughly 300 words) requires approximately 10^{14} FLOPs — video generation demands four orders of magnitude more computation [13]. Estimated single-inference costs range from 0.50 to 1.20 US dollars, while monthly user subscription fees stand at only 10-20 US dollars. Users would need to generate 10-20 videos monthly to break even, yet actual user generation frequency falls far below this threshold.

Physical modeling capacity is quantified via test metrics, including object permanence and collision response. The standardized PHY-100 test set contains 100 physical scene prompts (e.g., "a ball rolling down an inclined plane") [14]. Human evaluation results are presented in Table 5.

Table 5. Performance of video generation models on the PHY-100 physical reasoning benchmark

Capability Dimension	Sora	VideoPoet	W.A.L.T.	Human Baseline
Object Permanence	87%	76%	81%	98%
Collision Response Accuracy	53%	41%	48%	94%
Rationality of Gravitational Acceleration	61%	48%	55%	96%
Causal Logic	58%	52%	54%	97%

All models score substantially below human benchmarks on collision response and causal logic, indicating that current models do not truly learn physical laws and merely fit motion patterns at a

DOI: 10.54254/2755-2721/2026.GL35449

statistical level [14].

4. Application scenarios

4.1. Creative industry

Short videos and advertising represent the most mature commercial deployment scenarios, with durations aligned to the capability limits of existing models. AI comic drama production has already formed a closed commercial loop, while anthropomorphic short dramas are rapidly gaining audience acceptance. High-value CG visual effects workflows are widely expected to be among the first sectors partially replaced by AI. Nevertheless, only a limited fraction of AI-generated footage currently meets broadcast-ready standards, making human-AI collaboration a more realistic industry pathway [15].

4.2. Education and training

High-quality educational video production incurs substantial costs, a barrier AI generation promises to eliminate. World models hold exceptional prospects in this domain: videos evolve from passive viewing materials into navigable, interactive spaces. Historical events can be observed firsthand, and scientific phenomena can be repeatedly simulated. This paradigm-shifting knowledge delivery from "passive comprehension" to "active participation" stands poised to drastically improve knowledge transfer efficiency [16].

5. Challenges and future directions

5.1. Core current challenges

Technical hurdles fall into four primary categories. First, there's insufficient physical modeling capacity—models frequently produce visually plausible yet physically impossible scenes that violate fundamental physical laws [14]. Second, incomplete temporal consistency for long-form videos; the probability of logical breakdown rises exponentially with video length, with roughly 35% of Sora outputs exceeding 60 seconds containing obvious logical inconsistencies [4]. Third, limited fine-grained controllability; mainstream pipelines require 5-15 sampling attempts to produce satisfactory results, yielding an effective success rate of only 22% [12]. Fourth, prohibitive computational overhead, as video generation requires four orders of magnitude more computation than text generation, detailed above [13].

Ethical and legal risks remain severe. Regarding deepfake threats, current detection tools achieve an identification accuracy of approximately 89%, yet adversarial generation techniques can reduce detection rates to 67% [17]. Copyright regulation lacks clear legal frameworks: models are trained on vast volumes of copyrighted content, and the similarity distribution of generated outputs follows a long-tail pattern, with roughly 0.3% of outputs exhibiting high similarity to training data (cosine similarity >0.85) [18].

5.2. World models: the next-generation technical paradigm

World models are recognized as the dominant forward-looking research direction. Unlike conventional video generation models that learn the conditional probability $P(\text{frame}_t \mid \text{frame}_{(t-1)}, \text{text})$, world models aim to learn state transition functions $P(s_{t+1} \mid s_t, a_t)$ within latent state spaces,

DOI: 10.54254/2755-2721/2026.GL35449

where s_t denotes world state and a_t represents agent actions [16]. Recent projects including Google DeepMind's Genie have demonstrated preliminary capabilities for learning interactive world models from unlabeled video footage [19].

From a macro perspective, video generation technology faces a fundamental limitation: current models are trained to fit pixel distributions rather than understand real-world rules. This paradigm constraint manifests prominently in physical simulation and causal reasoning tasks. When a model generates convincing footage of a ball rolling upward against gravity, it does not comprehend gravitational force—it only reproduces low-probability patterns present within training datasets [14]. Overcoming this limitation may require moving beyond end-to-end learning frameworks to integrate physical priors, causal representations, or structured world models.

5.3. Interactive generation and human-AI collaboration

Interactive generation demands real-time model responsiveness. Sora currently takes roughly 180 seconds to render a 10-second video (generation speed 0.056× slower than real time), falling short of the ≥1× speed threshold required for real-time interaction [4]. Inference throughput must increase by approximately 20x under current hardware constraints, achievable via model distillation, quantization, or transformative architectural breakthroughs [13].

Human-AI collaborative workflows represent a more pragmatic near-term path. AI undertakes computationally intensive tasks, while humans retain creative decision-making and aesthetic judgment authority. The evolutionary trajectory of AI video tools centers on "augmentation tools" designed to expand creators' capabilities rather than fully replace human creators. Evaluation metrics should shift from "can the AI complete production independently" to "can the AI empower creators to work faster and more creatively" [20].

6. Conclusion

This survey systematically investigates the technological evolution, contemporary capabilities, and developmental trajectory of generative AI video production technology. Based on systematic analysis of existing research, this paper draws the following conclusions corresponding to its three core research questions:

First, at the architectural level: Integration of self-attention mechanisms and diffusion models delivers global receptive fields and parallel computation, resolving the foundational limitations of convolutional methods. Coexistence of diverse technical pipelines—ranging from the rise of DiT architectures to VideoPoet's large language model approach—sustains continuous iterative innovation. Quantitative analysis confirms Transformer architectures reduce 60-frame-interval temporal consistency error from 0.412 (CNN-based) to 0.178 (Sora), representing a 57% relative improvement.

Second, at the capability level: State-of-the-art systems deliver near-professional quality for short video generation, demonstrating coherent object permanence and natural motion dynamics. However, fundamental limitations persist in multi-modal consistency, fine-grained controllability, and authentic physical simulation. Physical reasoning evaluations reveal that even the most advanced model achieves merely 53% accuracy on collision response tasks, far below the human benchmark of 94%.

Third, at the commercialization level: Excessive computational cost constitutes the primary bottleneck for industrial adoption—video generation requires four orders of magnitude more

DOI: 10.54254/2755-2721/2026.GL35449

computation than text generation. Technical feasibility does not equate to commercial viability, a critical lesson guiding the domain's future development.

This survey contains several inherent limitations. The field's breakneck pace of advancement risks rendering certain findings obsolete rapidly. Additionally, proprietary closed-source systems such as Sora withhold full technical specifications, restricting analysis to publicly released technical reports.

Three priority directions are identified for future research: (1) Architectural innovations to drastically cut computational demands—inference throughput must improve roughly 20x to hit real-time interaction thresholds; (2) Transition from pixel-fitting frameworks to world models, integrating physical priors and causal representations to break through bottlenecks in physical simulation; (3) Construction of human-AI collaborative creation frameworks to redefine AI's role within the creative industry. Crossing the divide from pixel fitting to world comprehension, from statistical pattern matching to genuine physical rule reasoning, will mark the pivotal evolution of generative video intelligence from mere production tools to collaborative creative partners.

References

[1] Yang, S., et al. (2024). Video generation: A survey. *ACM Computing Surveys*.
[2] Khan, S., et al. (2022). Transformers in vision: A survey. *ACM Computing Surveys*.
[3] Vaswani, A., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.
[4] Brooks, T., et al. (2024). Video generation models as world simulators. *OpenAI Technical Report*.
[5] Dosovitskiy, A., et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations (ICLR)*.
[6] Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*.
[7] Peebles, W., & Xie, S. (2023). Scalable diffusion models with transformers. *International Conference on Computer Vision (ICCV)*.
[8] Yan, W., et al. (2021). VideoGPT: Video generation using VQ-VAE and transformers. *arXiv preprint*.
[9] Yu, J., et al. (2024). MagViT V2: High-fidelity video tokenization. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
[10] Kondratyuk, D., et al. (2024). VideoPoet: A large language model for zero-shot video generation. *arXiv preprint*.
[11] Gupta, A., et al. (2024). W.a.l.t.: Window attention latent transformer for video generation. *arXiv preprint*.
[12] Liu, J., et al. (2025). Towards interactive video world modeling: Frontiers, challenges, benchmarks, and future trends. *arXiv preprint*.
[13] Li, Z., et al. (2025). Efficient video generation: A survey of acceleration techniques. *arXiv preprint*.
[14] Yan, W., et al. (2025). Phy-100: A physical reasoning benchmark for video models. *arXiv preprint*.
[15] Zhang, X., et al. (2025). Human-AI collaboration in creative video production. *ACM Conference on Human Factors in Computing Systems (CHI)*.
[16] Bruce, J., et al. (2024). Genie: Generative interactive environments. Google DeepMind.
[17] Wang, L., et al. (2025). Deepfake detection in the era of generative video models. *IEEE Transactions on Information Forensics and Security*.
[18] Chen, Y., et al. (2025). Copyright challenges in generative AI training. arXiv preprint.
[19] Rombach, R., et al. (2022). High-resolution image synthesis with latent diffusion models. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
[20] Singer, U., et al. (2023). Make-a-video: Text-to-video generation without text-video data. *International Conference on Learning Representations (ICLR)*.