https://doi.org/10.2352/EI.2026.38.7.IMAGE-266

From Pixels to Worlds: A Survey on the New Wave of High-Fidelity Video Generation

Weijuan Xi
Purdue University

Abstract

The field of computer vision is currently undergoing a pivotal transformation, shifting its focus from discriminative to generative tasks. Over the past two decades, the discipline was primarily defined by the discriminative imperative, which sought to enable machines to perceive, classify, and segment the visual world. However, catalyzed by the development of the Diffusion Transformer (DiT), the years 2024 and 2025 marked a Generative Turn, where the benchmark of artificial visual intelligence has evolved from mere classification to controllable simulation. The ability to generate high-fidelity, physically consistent video has led to the development of advanced generative models capable of representing underlying physical dynamics and environmental causality through large-scale data and computation. This survey provides a comprehensive analysis of the recent emergence of high-fidelity video generation. It traces the evolution from the era of feature engineering to the current Diffusion Transformers (DiTs) based generation era, summarizes the present state of video generation and the technical advancements driving this period, and offers a guide detailing the architectures, data selection, and training methodologies essential for high-fidelity video generation.

Introduction

The trajectory of computer vision over the last twenty years moved from hand-crafted local features to deep semantic representations, and finally, to generative modeling of spatiotemporal dynamics. As shown in Fig. 1, it can be briefly divided into several eras:

From 2005 to 2012 is the Era of Feature Engineering, and computer vision was a discipline of intuition and manual design. The field was dominated by the extraction of features, which are the mathematical descriptions of local image patches that were invariant to scale, rotation, and illumination. Algorithms such as the Scale-Invariant Feature Transform (SIFT) [39] and Histogram of Oriented Gradients (HOG) [10] were the standard-bearers of this era. The researchers manually encoded the visual properties that they believed were important, such as edges, corners, and texture gradients. During this period, visual content synthesizing was limited to texture tiling or morphing techniques that blended existing pixels. The fundamental bottleneck was the semantic gap between the low-level pixel data a computer perceives and the high-level concepts that humans understand.

From 2012 to 2018 is the Deep Learning Revolution, which is dominated by the descriminative deep neural network. The pivotal moment arrived in 2012 with the ImageNet Large Scale Visual Recognition Challenge [50], where the AlexNet architecture [32] demonstrated the overwhelming superiority of Convolutional

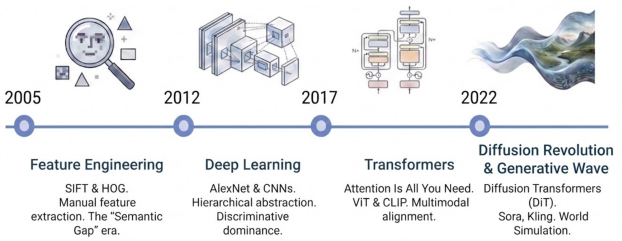


Figure 1. *Evolution of computer vision paradigms: From discriminative perception to high-fidelity generative synthesis.*

Neural Networks (CNNs) [33, 53, 21, 52, 61, 6]. This marked the end of manual feature engineering. The CNN learned its own features, stacking layers of convolutions to build a hierarchy of abstraction—from edges in the early layers to complex object parts in the deeper layers. For the next half-decade, the field was obsessed with discriminative tasks. Architectures like VGG [54], Inception [56, 57, 55], and the residual connections of ResNet [20, 21] pushed the classification accuracy beyond human levels. However, generative capabilities remained nascent. The primary innovation in generation during this era was the Generative Adversarial Network (GAN) [16], introduced by Ian Goodfellow in 2014.

From 2017 to 2022 is the Rise of Transformers [62] and the Pre-training Paradigm [46, 11, 12], and the seeds of the current generative wave were sown in Natural Language Processing (NLP) at this era. The introduction of the Transformer architecture in the landmark paper "Attention Is All You Need" (2017) fundamentally changed how AI models handled sequences. Unlike Recurrent Neural Networks (RNNs) [13, 5, 23, 8], which processed data sequentially and forgot long-term dependencies, Transformers used "self-attention" to weigh the relevance of every part of the input sequence simultaneously. The adaptation of this architecture to vision via Vision Transformers (ViT) [12] demonstrated that images could be treated as sequences of patches. This broke the inductive bias of convolutions, which assumed local connectivity. A Transformer could attend to pixels on opposite sides of an image (or video) with equal ease, theoretically solving the long-range dependency problem that had plagued video generation. Simultaneously, the concept of multimodal alignment matured. Models like CLIP (Contrastive Language-Image Pre-training) [45] learned to map text and images into a shared latent space. By understanding the semantic relationship between the text and an image, models could effectively be steered by natural language.

From 2022 to 2025 is the Diffusion Revolution and the Gen-

erative Wave. The final piece of the puzzle was the Denoising Diffusion Probabilistic Model (DDPM) [22]. Unlike GANs, which relied on unstable adversarial training, diffusion models learned a more stable objective: reversing a gradual noising process. They learned to construct structure from chaos. The computational cost of diffusion was initially prohibitive for high-resolution video. The breakthrough was Latent Diffusion Models (LDMs) [48], exemplified by Stable Diffusion. By compressing images into a lower-dimensional latent space using Variational Autoencoders (VAEs) [31] and performing the diffusion there, researchers decoupled the computational cost from the pixel resolution. The convergence that Diffusion models operating in latent space, powered by scalable Transformer backbones ignited the video generation in 2024. No longer constrained by the instability of GANs or the locality of CNNs, models like Sora and Kling began to demonstrate emergent capabilities, simulating complex lighting, fluid dynamics, and object permanence purely from exposure to vast datasets.

This survey presents a comprehensive taxonomy of high-fidelity video generation. We introduce the pivotal architectural transition from Generative Adversarial Networks (GANs) to Diffusion Transformers (DiT) and provide a rigorous synthesis of the latest advancements in data-centric curation and multi-phase training paradigms. Furthermore, we summarize current state-of-the-art (SOTA) models and identify the emerging trends defining both proprietary and open-weights ecosystems. The paper concludes with an analysis of critical remaining challenges and the technological trajectory toward truly autonomous, agentic video environments.

Defining High-Fidelity Video Synthesis

High-fidelity video generation has been a huge challenge in the past two decades though researchers achieve main breakthrough. It is essential to establish a clear definition of the high-fidelity video generation before delving into the technical complexities. High fidelity in this context is comprised of three distinct dimensions: spatial fidelity, which governs visual quality through high resolution, sharpness, and photorealism; temporal fidelity, which ensures consistency by maintaining smooth motion, adhering to fundamental physical laws, and preserving the identity and structure of subjects over time; and semantic fidelity, which reflects how accurately and strictly the generated content adheres to input control signals or prompts.

The past several years have witnessed an explosion of research in high-fidelity video generation, encompassing a diverse array of methodologies such as Spatio-temporal Attention, Diffusion Transformers (DiT) [44], Flow Matching [27], and 3D-aware latent representations. However, rather than navigating the exhaustive technical minutiae of this rapidly expanding landscape, it is more instructive to first examine the three fundamental pillars that define the modern state-of-the-art: architectural innovation, data-centric curation, and advanced training strategies. These pillars collectively address the core challenges of visual sharpness, temporal consistency, and semantic adherence.

Architectural Foundations and Innovations

A suite of architectural and algorithmic innovations has emerged, successfully transitioning video generation from a nascent research frontier to a scalable, high-fidelity reality. Key among these are Diffusion Transformers (DiT) [44] for long-range dependency modeling, Pyramidal Flow Matching [27] for efficient high-resolution training, and 3D Causal VAEs [81] for dense spatio-temporal compression. While a single, unified architecture for video generation has not yet been universally codified, the frontier models, such as Sora 2 [42], Veo 3.1 [17], and Kling 3.0 [59], have largely converged toward a tripartite framework (Fig. 4): multimodal input encoding and alignment for precise semantic control, iterative latent processing centered on a DiT backbone for scalable generation, and high-fidelity spacetime decoding to reconstruct photorealistic pixels with temporal grace. In this section, we introduce the Causal VAE, DiT, and Flow Matching which are the three main techniques of the converged architecture.

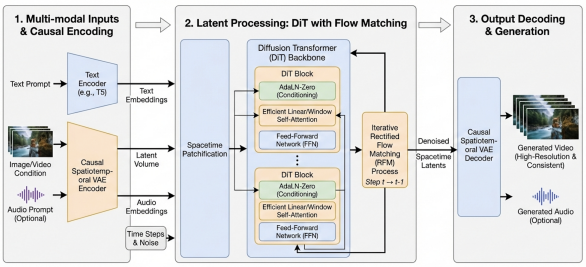


Figure 2. *Architectural convergence in video generation.*

Spatio-Temporal Compression: Causal VAEs

The high dimensionality of the video data, which is represented as $\mathscr{V} \in R^{T\times H\times W\times C}$, necessitates a robust compression mechanism. In this notation, T and C correspond to the temporal and channel dimensions, while H and W define the spatial resolution of each frame. Modern frameworks employ a Causal VAE [81] to map pixels into a compact latent manifold $\mathscr{Z}$. Unlike spatial-only VAEs [48, 1], the Causal VAE utilizes Causal 3D Convolutions where the receptive field for any frame t is strictly restricted to frames $\{\tau \mid \tau \leq t\}$. Formally, the encoding process is defined as:

$$z_t = \mathscr{E}(V_{\leq t}; \phi)$$

This temporal causality ensures that the latent representation preserves the unidirectional flow of time, preventing information leakage from the future and enabling autoregressive video extension. To mitigate memory constraints during training on long sequences, Temporal Tiling is often employed, processing the video in overlapping chunks while maintaining boundary consistency through hidden state propagation.

Scalable Backbones: Diffusion Transformers

The transition from CNN-based U-Nets to Diffusion Transformers (DiT) marks a fundamental shift toward the neural scaling law [29, 82, 80] paradigm. The DiT architecture synthesizes the patch-based sequence processing of Vision Transformers (ViT) with the iterative denoising process characteristic of diffusion frameworks. By treating video latents as a sequence of spatio-temporal visual tokens, DiT provides a highly scalable alternative to traditional convolutional backbones. Following the patchification process, the latent z is divided into $p \times p \times q$ patches. The architecture leverages the global self-attention mechanism to model

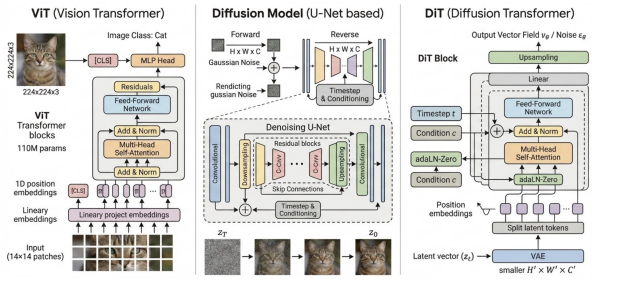


Figure 3. *Video generation architecture: From Vision Transformers (ViT) and Convolutional Diffusion to the Diffusion Transformer (DiT).*

long-range dependencies across the entire volume. The standard training objective minimizes the reweighted MSE loss:

$$\mathscr{L}_{DiT} = E_{z_0,\varepsilon\sim\mathscr{N}(0,\mathbf{I}),t}\left[\|\varepsilon - \varepsilon_\theta(z_t,t,\mathbf{c})\|^2\right]$$

where $\mathbf{c}$ represents the conditioning vector (e.g., CLIP text embeddings). The primary advantage of DiT lies in its Inductive Bias Neutrality; by removing the local constraints of convolutions, DiT can capture complex, non-local physical interactions and globally coherent compositions that are characteristic of high-fidelity scenes.

Efficient Sampling: Flow Matching Methodologies

While traditional Diffusion Models rely on Stochastic Differential Equations (SDEs) with curved sampling paths, Flow Matching (FM) optimizes a deterministic Probability Flow ODE. FM aims to learn a vector field $v_\theta(x_t,t)$ that generates a linear probability path between noise x_0 and data x_1:

$$x_t = (1-t)x_0 + tx_1$$

The corresponding objective function is:

$$\mathscr{L}_{FM} = E_{t\sim\mathscr{U}(0,1),x_0,x_1}\left[\|v_\theta(x_t,t) - (x_1 - x_0)\|^2\right]$$

Because the learned trajectory is a straight line in the latent space, the solver requires significantly fewer steps to reach the data manifold. This efficiency is critical for producing high-fidelity, high-resolution outputs (1080p+) in commercially viable timeframes.

Dataset

The trajectory of high-fidelity video generation is inextricably linked to the evolution of the underlying training data. As the objective of generative models shifts from merely predicting the next sequence of pixels to simulating temporally coherent, multimodal worlds, the requirements for training datasets have fundamentally changed. The reliance on noisy, web-scraped video-text pairs has been largely superseded by sophisticated data curation pipelines emphasizing high resolution, long-take temporal consistency, and dense semantic grounding powered by Multimodal Large Language Models (MLLMs).

Early diffusion and autoregressive video models were heavily reliant on massive, broadly scraped datasets, such as WebVid-10M [3] and HD-VG-130M [67]. While these corpora provided the raw volume necessary to establish baseline text-to-video alignment, they suffered from low resolutions, short clip durations, and sparse, noisy alt-text. Recognizing that visual quality and temporal stability are bottlenecked by training data, recent efforts have pivoted toward rigorous desirability filtering. Datasets like Panda-70M [7] mark a significant transitional phase, where massive raw collections (derived from sources like HD-VILA-100M [77]) are processed using multiple cross-modality teacher models. By splitting long videos into semantically coherent clips and utilizing fine-grained retrieval models to select optimal captions, the field began moving toward automated, high-quality data refinement.

A critical challenge in building realistic video models is the morphing artifact and the breakdown of physical laws over extended generations. To address these limitations, data curation strategies have increasingly prioritized long-take filtering. By systematically removing scene cuts, fade-ins, and synthetic text overlays, these pipelines compel models to internalize sustained physical motion and long-range temporal consistency. A prime example is LVD-2M (2024) [75], the first large-scale dataset specifically curated for long-take generation. Containing 2 million videos spanning over 10 seconds with significant motion dynamics, it pairs continuous footage with temporally-dense hierarchical captions. Similarly, the push for native high-definition generation is exemplified by OpenVid-1M (2025) [40], which meticulously filters for high aesthetics and clarity, notably including a specialized subset (OpenVidHD-0.4M) dedicated entirely to 1080p resolution. These datasets ensure that models learn structural consistency over time rather than just static spatial aesthetics.

The most transformative trend in the current landscape is the deprecation of human-written or metadata-derived captions in favor of dense, synthetic descriptions generated by advanced MLLMs. This approach bridges the gap between simple visual concepts and the complex spatiotemporal reasoning required for world simulation. Dataset curation in 2025 has become highly specialized to address specific generative failures and real-world usage patterns: Datasets like VideoUFO [66] explicitly address the distribution gap between academic training data and real-world text-to-video prompts. By clustering over a million real user prompts into focused topics and scraping YouTube specifically for those domains, VideoUFO provides dual-density captions tailored strictly to what users actually attempt to generate; Current models often fail at fine-grained physical contact. The introduction of HOIGen-1M [38] addresses this by providing over a million manually verified video clips specifically focused on accurate Human-Object Interaction (HOI), complete with dense interaction captions; Moving beyond isolated subjects, datasets like Sekai [37] introduce thousands of hours of first-person and drone footage paired with high-precision camera trajectory annotations, pushing models to understand 3D spatial navigation and geographical consistency.

Training Strategy

The high-fidelity video generation models rely on a phased training approach. This hierarchy is designed to manage the quadratic complexity of spatio-temporal attention while progressively improving the model from static images to high-resolution, production-quality video.

The first phase establishes the model's basic understanding of the world. Rather than training on video from scratch, models

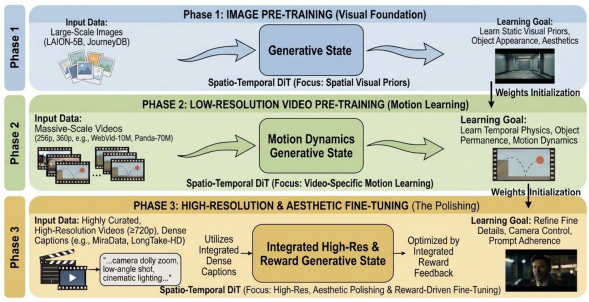


Figure 4. *Hierarchical spatio-temporal DiT training for video generation: From foundational learning to controlled aesthetic polishing.*

are initialized with weights from large-scale text-to-image (T2I) models trained on billions of diverse images to encode strong visual priors, spatial composition, and aesthetic quality. This ensures that the model can render complex textures, lighting, and semantic concepts before it ever sees a moving frame. This phase leverages massive datasets like LAION [51] or internal high-resolution image repositories to build a robust latent space.

Once a visual foundation is established, the model is introduced to the temporal dimension. This is where the model learns to maintain consistency across the latent spacetime patches. To keep VRAM costs manageable, training is conducted on massive volumes of video at lower resolutions (e.g., 256p or 360p) to learn temporal physics, such as, how objects move, fluid dynamics, and maintaining object permanence across frames. This stage focuses on motion diversity rather than pixel-perfect clarity. Training at this stage utilizes millions of varied video clips (such as, WebVid-10M [3], Panda-70M [7], InternVid [69]). While these clips may have lower aesthetic quality, they must provide high motion complexity to prevent the model from simply learning static or sliding motions. By training at lower resolutions, the model can process longer sequences or higher frame rates, which is critical for understanding long-range temporal dependencies.

The final phase, often referred to as the polishing step which transitions the model to high-resolution outputs (720p, 1080p, or 4K) and variable aspect ratios. The phase refines fine-grained details and ensures that the output is production ready. This stage bridges the gap between a technically correct motion and a cinematically pleasing video. Unlike the previous phases, this stage uses a significantly smaller but highly curated dataset. These videos are filtered for high aesthetic scores, professional lighting, and minimal watermarks. For this phase, MiraData [28] is often cited as a benchmark for high-fidelity video generation because of its focus on long-range temporal consistency and structured captions; LongTake-HD [78] is used for scenario coherence and ensures that a video feels like a single professionally shot take rather than a series of disconnected frames.

Landscape of Video Generation

The field of video generation has transitioned from producing short, stochastic flickering clips to creating coherent, high-fidelity world simulations. As of early 2026, the landscape is defined by two parallel tracks: proprietary closed-source ecosystems prioritizing extreme scale and production features, and a rapidly maturing open-weights community focusing on architectural efficiency.

Table 1: SOTA video model specifications.

Model	Development	Resolution	Duration
Closed-Source			
Sora 2	OpenAI	1080p	25s
Veo 3.1	Google	4K	30s
Kling 3.0	Kuaishou	4K	15s*
Gen-4.5	Runway	1080p	10s
Ray 3	Luma AI	4K	11s
Open-Source / Open-Weights			
Wan 2	Alibaba	1080p	15s
LTX-2	Lightricks	4K	20s
Hunyuan 1.5	Tencent	1080p	15s
Mochi 1	Genmo	480p	5.4s

**Supports temporal chaining up to 180s.*

Frontier Models

Table 1 shows the open and closed-source video generation models. Closed-source models currently lead in high-end cinematic physics, long-duration coherence, and native multi-modal integration. Building on the original DiT framework, *Sora 2* [42] has moved beyond simple pixel prediction to physics-grounded generation. It excels in complex world dynamics and introduced Character Cameos for consistent identity across scenes; *Kling 3.0* [59] is currently the industry leader in native resolution and duration, which supports native 4K at 60fps and generates clips up to 120 seconds. Its standout "Storyboard" feature allows for multi-shot generation with zero identity drift between cuts; Positioned as a cinematographic assistant, *Veo 3.1* [17] integrates spatial audio natively and focuses on cinematic lighting and camera controls. It leverages a massive dataset of high-fidelity film content to achieve a distinct film-like aesthetic compared to more synthetic-looking models; *Runway* [49] remains the primary choice for professional VFX pipelines, offering fine-grained control over specific object motions and character performances.

The open-weights community has significantly narrowed the performance gap, often introducing architectural innovations such as Mixture-of-Experts (MoE) [14] before their closed-source counterparts. *Wan 2* [63] released by Alibaba represents the first successful implementation of an MoE architecture in video generation. By using specialized high-noise and low-noise experts, Wan 2 expands the model capacity (14B+ parameters) while maintaining inference efficiency, allowing it to compete with closed models in esthetic quality; *Hunyuan1.5* [71] is a robust 13B-parameter DiT that utilizes a Multimodal Large Language Model (MLLM) as its text encoder. This allows for superior instruction adherence and semantic understanding, particularly in complex, multi-subject scenes; *Mochi 1* [58] , which focuses on fluid and photorealistic motion, is distinguished by its Asymmetric Diffusion Transformer (AsymmDiT) architecture, Mochi 1. It has become a favorite for the developer community due to its extensive LoRA [25] fine-tuning support; *LTX-2* [18] is an efficiency-first model optimized for local execution. Despite its smaller footprint, it supports native audio generation and 4K output, making it the current SOTA for lightweight on-device video synthesis.

Control Methodology

The evolution of generative video has been marked by a transition from linguistic prompting to precise multimodal condition-

ing. While early models relied on the inherent creativity of latent space, modern high-fidelity systems prioritize the ability of a user to enforce structural, temporal, and semantic constraints. We categorize these methodologies into four primary domains: geometric grounding, cinematographic steering, audio-visual synchrony, and programmatic orchestration.

Geometric control methods incorporate low-level structural priors into the denoising process to enforce spatial consistency. In the context of video DiTs, structural signals such as depth maps, surface normals, and Canny edges are often integrated via either a sidecar network or additional input channels to the transformer blocks. For example, MotionBrush utilizes a regional motion-masking branch to guide local pixel displacement; DragAnything [73] employs an entity-aware sidecar to map trajectory coordinates to latent features; and Boximator [65] integrates bounding-box constraints through dedicated attention layers to ensure precise object-level localization.

Early video synthesis frameworks often conflated global camera motion with intrinsic scene dynamics, leading to significant perspective distortions and a lack of geometric consistency. Modern cinematographic steering mitigates these artifacts by conditioning the generative process on explicit camera extrinsics. Key methodologies, such as CameraCtrl [19] and MotionCtrl [70], leverage camera pose sequences to modulate temporal attention layers. By mapping these poses into high-dimensional embeddings, the model learns to warp the background and foreground consistently with the physical laws of parallax. Frameworks like GEN3C [47] and I2VControl-Camera [15] allow users to define custom rotation, focal length, and movement distance, effectively decoupling global camera motion from local object dynamics. RealCam-I2V [35] and ReCamMaster [2] provide interfaces for interactive camera trajectory adjustment, allowing for the re-rendering of existing scenes from entirely new perspectives while maintaining temporal consistency.

Video is an inherently multimodal medium where the temporal axis is often dictated by audio cues. Lip-sync and expressive portrait animation [60, 9] utilize audio embeddings to drive the deformation of facial latents. The current frontier involves "cross-modal attention" where phonemes directly influence the generation of micro-expressions and gaze. Models like MF-ETalk [68] utilize decoupled Action Units (AUs) to synthesize facial expressions that match the emotional cadence of the input audio.

The final tier of control is the shift from manual prompting to deterministic, logic-based pipelines. Frameworks like ComfyUI [76] have popularized directed acyclic graphs (DAGs) for video generation. This allows researchers to chain disparate models into a single, reproducible pipeline. Tools like Boximator [65] allow for programmatic object placement. By defining $[x, y, w, h]$ coordinates across a timeline, users can enforce hard constraints on object trajectories, ensuring that agentic entities within the video follow a predefined narrative path.

Challenges and Research Directions

As the field of video generation matures from producing pixel-perfect frames to simulating coherent world dynamics, the research frontier has shifted toward addressing fundamental limitations in physical grounding and temporal reasoning. While architectural convergence on Diffusion Transformers (DiTs) and scaling laws have largely solved short-term visual fidelity for 5–10 second clips, extending this performance into stable, long-horizon world simulations remains a significant challenge. In this section, we analyze the critical bottlenecks and emerging methodologies that define the research landscape.

Physical Grounding: Beyond Intuitive Simulation

Current state-of-the-art models, including Sora 2 and Veo 3.1, operate primarily as Intuitive Physics Engines that learn visual correlations rather than explicit Newtonian laws. This leads to frequent violations of physical causality, such as hallucinations of objects during complex collisions. To bridge this gap, the research community is pivoting from purely data-driven diffusion to Physics-Informed Generative Models.

Recent breakthroughs, such as Neural Gaussian Force Fields (NGFF) [34] and PhysCtrl [64], suggest that integrating differentiable physics simulators directly into the latent denoising process can enforce structural integrity. By modeling physical dynamics as 3D point trajectories within a generative physics network, these models move beyond 2D pixel prediction toward true 4D world modeling. Furthermore, the transition from stochastic sampling in Diffusion Models to deterministic paths in Flow Matching (FM) provides a more stable framework for learning these physical trajectories, potentially reducing the straight-line error in complex motion synthesis.

Action-Conditioned World Models and Robotic Embodiment

For agentic video environments, a primary challenge is the disconnect between visual observations and the motor commands required for physical interaction. Traditional text-to-video (T2V) models lack a structural understanding of how specific actions affect a 3D scene. This has given rise to the Action-Conditioned Video Model (ACVM) [36] paradigm, exemplified by recent foundations like NVIDIA Cosmos 3 [41] and Genie 3 [4].

These models utilize a World Action Model architecture, which helps robots and autonomous agents mentalize the future by predicting next-frame observations conditioned on a specific action vector. Current research, such as the DreamZero [79] framework, demonstrates that models pre-trained on massive video datasets can achieve high zero-shot success rates in pick-and-place tasks when augmented with even a few hours of unlabeled robot data. This indicates that the latent knowledge within high-fidelity video generators is a viable dream simulator for scalable reinforcement learning.

Long-Horizon Consistency and Memory Architectures

While models like Kling 3.0 and Sora 2 can technically generate minutes of footage via temporal chaining, the temporal logic of these sequences often remains partial, suffering from accumulated distributional shifts and semantic drift. Maintaining character identity, set geometry, and narrative logic over long horizons is currently bottlenecked by the limited context windows of Transformers.

The most promising direction in 2026 involves Cycle-Consistency Objectives and explicit Memory Mechanisms [72, 24]. Frameworks such as LIVE [26] introduce a reverse-generation process to reconstruct the initial state from a forward rollout, effectively bounding the error accumulation that typically

plagues autoregressive generation. Simultaneously, the introduction of persistent memory banks, which is referred to as WorldMem [74] or WorldPack [43], allows models to store and retrieve 3D scene scaffolds, ensuring that rooms and subjects look the same even when revisiting after extended periods.

The Convergence of Vision and Graphics: Interactive 4D Spaces

The ultimate evolution of pixels to worlds lies in the transition from 2D video projections to interactive 4D realities. This shift represents a merging of generative vision with real-time computer graphics. Instead of generating static pixel streams, future systems are moving toward Video-to-4D Gaussian Splatting [30], where the model generates a dynamic 3D world that users can navigate and manipulate in real-time.

Conclusion

The field of computer vision has undergone a pivotal transformation, shifting from the discriminative imperative of classification and segmentation toward a Generative Turn where intelligence is defined by sophisticated simulation. This survey has traced the historical evolution from manual feature engineering to the current era of world simulation, a transition catalyzed by the convergence of Causal 3D VAEs, Diffusion Transformers (DiT), and Flow Matching.

Central to the success of these high-fidelity systems is the advancement of multimodal controllability, which allows for precise geometric, cinematographic, and programmatic steering of generated content. By moving beyond intuitive correlations toward Physics-Informed and Action-Conditioned logic, the research frontier is now bridging the gap between pixel prediction and true physical causality. As these models evolve into interactive 4D realities, the distinction between a recorded reality and a generated simulation will effectively vanish, marking the full realization of the journey from pixels to worlds.

References

[1] Haleh Akrami, Anand A. Joshi, Jian Li, Paul M. Thompson, and Richard M. Leahy. Spatial variational auto-encoding via matrix-variate normal distributions. In *Proceedings of the 2019 SIAM International Conference on Data Mining (SDM)*, pages 441–449, 2019.

[2] Jianhong Bai, Menghan Xia, Xiao Fu, et al. Recammaster: Camera-controlled generative rendering from a single video, 2025.

[3] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval, 2022.

[4] Phil Ball, Jakob Bauer, et al. Genie 3: A new frontier for world models. Technical report, Google DeepMind, August 2025. Experimental release via Project Genie, January 2026.

[5] Yoshua Bengio, Patrice Simard, and Paolo Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166, 1994.

[6] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[7] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers, 2024.

[8] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, et al. Learning phrase representations using rnn encoder–decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[9] Jiahao Cui, Hui Li, Yun Zhan, et al. Hallo3: Highly dynamic and realistic portrait image animation with video diffusion transformer, 2025.

[10] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection, 2005.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the NAACL-HLT*, pages 4171–4186, 2019.

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

[13] Jeffrey L. Elman. Finding structure in time. *Cognitive Science*, 14(2):179–211, 1990.

[14] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity, 2022.

[15] Wanquan Feng, Jiawei Liu, Pengqi Tu, et al. I2vcontrol-camera: Precise video camera control with adjustable motion strength, 2025.

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, et al. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NIPS)*, volume 27, pages 2672–2680, 2014.

[17] Google. Veo 3.1: More consistency, creativity, and control with ingredients to video. https://blog.google/innovation-and-ai/technology/ai/veo-3-1-ingredients-to-video/, 2026. Accessed: 2026-03-22.

[18] Yoav HaCohen, Benny Brazowski, Nisan Chiprut, et al. Ltx-2: Efficient joint audio-visual foundation model, 2026.

[19] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation, 2025.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 6840–6851, 2020.

[23] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1797.

[24] Yicong Hong, Yiqun Mei, Chongjian Ge, et al. Relic: Interactive video world model with long-horizon memory, 2025.

[25] Edward J. Hu, Yelong Shen, Phillip Wallis, et al. Lora: Low-rank adaptation of large language models, 2021.

[26] Junchao Huang, Ziyang Ye, Xinting Hu, et al. Live: Long-horizon interactive video world modeling, 2026.
[27] Yang Jin and et al. Pyramidal flow matching for efficient video generative modeling, 2025.
[28] Xuan Ju, Yiming Gao, Zhaoyang Zhang, et al. Miradata: A large-scale video dataset with long durations and structured captions, 2024.
[29] Jared Kaplan, Sam McCandlish, Tom Henighan, et al. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.
[30] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (TOG)*, 42(4):1–14, 2023.
[31] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*, 2014.
[32] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks, 2012.
[33] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
[34] Shiqian Li, Ruihong Shen, Junfeng Ni, et al. Learning physics-grounded 4d dynamics with neural gaussian force fields, 2026.
[35] Teng Li, Guangcong Zheng, Rui Jiang, et al. Realcam-i2v: Real-world image-to-video generation with interactive complex camera control, 2025.
[36] Yichen Li and Antonio Torralba. Multimodal action conditioned video generation, 2025.
[37] Zhen Li, Chuanhao Li, Xiaofeng Mao, et al. Sekai: A video dataset towards world exploration, 2025.
[38] Kun Liu, Qi Liu, Xinchen Liu, et al. Hoigen-1m: A large-scale dataset for human-object interaction video generation, 2025.
[39] David G. Lowe. Object recognition from local scale-invariant features, 1999.
[40] Kepan Nan, Rui Xie, Penghao Zhou, et al. Openvid-1m: A large-scale high-quality dataset for text-to-video generation, 2025.
[41] NVIDIA. Cosmos world foundation model platform for physical ai, 2025.
[42] OpenAI. Sora 2 is here. https://openai.com/index/sora-2/, 2025. Accessed: 2026-03-22.
[43] Yuta Oshima, Yusuke Iwasawa, Masahiro Suzuki, et al. Worldpack: Compressed memory improves spatial consistency in video world modeling, 2025.
[44] William Peebles and Saining Xie. Scalable diffusion models with transformers, 2023.
[45] Alec Radford, Jong Wook Kim, Chris Hallacy, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR, 2021.
[46] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *OpenAI Technical Report*, 2018.
[47] Xuanchi Ren, Tianchang Shen, Jiahui Huang, et al. Gen3c: 3d-informed world-consistent video generation with precise camera control, 2025.
[48] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022.
[49] Runway AI, Inc. Runway Gen-4.5: General World Models for High-Fidelity Video Generation, December 2025. Accessed: 2026-03-23.
[50] Olga Russakovsky, Jia Deng, Hao Su, et al. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.
[51] Christoph Schuhmann, Romain Beaumont, Richard Vencu, et al. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022.
[52] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*, pages 568–576, 2014.
[53] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
[54] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015.
[55] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.
[56] Christian Szegedy, Wei Liu, Yangqing Jia, et al. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015.
[57] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016.
[58] Genmo Team. Mochi 1. https://github.com/genmoai/models, 2024.
[59] Kuaishou Technology. Kling ai launches 3.0 model: Ushering in an era where everyone can be a director. https://ir.kuaishou.com/news-releases/news-release-details/kling-ai-launches-30-model-ushering-era-where-everyone-can-, 2026. Accessed: 2026-03-22.
[60] Linrui Tian, Siqi Hu, Qi Wang, Bang Zhang, and Liefeng Bo. Emo2: End-effector guided audio-driven avatar video generation, 2025.
[61] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4489–4497, 2015.
[62] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neu-*

ral Information Processing Systems (NeurIPS), pages 5998–6008, 2017.

[63] Team Wan, Ang Wang, Baole Ai, et al. Wan: Open and advanced large-scale video generative models, 2025.

[64] Chen Wang, Chuhao Chen, Yiming Huang, et al. Physctrl: Generative physics for controllable and physics-grounded video generation, 2025.

[65] Jiawei Wang, Yuchen Zhang, Jiaxin Zou, Yan Zeng, Guoqiang Wei, Liping Yuan, and Hang Li. Boximator: Generating rich and controllable motions for video synthesis, 2024.

[66] Wenhao Wang and Yi Yang. Videoufo: A million-scale user-focused dataset for text-to-video generation, 2025.

[67] Wenjing Wang, Huan Yang, Zixi Tuo, Huiguo He, Junchen Zhu, Jianlong Fu, and Jiaying Liu. Swap attention in spatiotemporal diffusions for text-to-video generation, 2024.

[68] Xueping Wang, Yuemeng Huo, Yanan Liu, Xueni Guo, Feihu Yan, and Guangzhe Zhao. Multimodal feature-guided audio-driven emotional talking face generation. *Electronics*, 14(13), 2025.

[69] Yi Wang, Yinan He, Yizhuo Li, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation, 2024.

[70] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation, 2024.

[71] Bing Wu, Chang Zou, Changlin Li, et al. Hunyuanvideo 1.5 technical report, 2025.

[72] Tong Wu, Shuai Yang, Ryan Po, et al. Video world models with long-term spatial memory, 2025.

[73] Weijia Wu, Zhuang Li, Yuchao Gu, et al. Draganything: Motion control for anything using entity representation, 2024.

[74] Zeqi Xiao, Yushi Lan, Yifan Zhou, et al. Worldmem: Long-term consistent world simulation with memory, 2026.

[75] Tianwei Xiong, Yuqing Wang, Daquan Zhou, Zhijie Lin, Jiashi Feng, and Xihui Liu. Lvd-2m: A long-take video dataset with temporally dense captions, 2024.

[76] Zhenran Xu, Xue Yang, Yiyu Wang, et al. Comfyui-copilot: An intelligent assistant for automated workflow development, 2025.

[77] Hongwei Xue, Tiankai Hang, Yanhong Zeng, et al. Advancing high-resolution video-language representation with large-scale video transcriptions, 2022.

[78] Xin Yan, Yuxuan Cai, Qiuyue Wang, et al. Long video diffusion generation with segmented cross-attention and content-rich video data curation, 2025.

[79] Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, et al. World action models are zero-shot policies, 2026.

[80] Yuanyang Yin, Yaqi Zhao, Mingwu Zheng, et al. Towards precise scaling laws for video diffusion transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[81] Lijun Yu, José Lezama, Nitesh B. Gundavarapu, et al. Language model beats diffusion – tokenizer is key to visual generation, 2024.

[82] Xiaua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

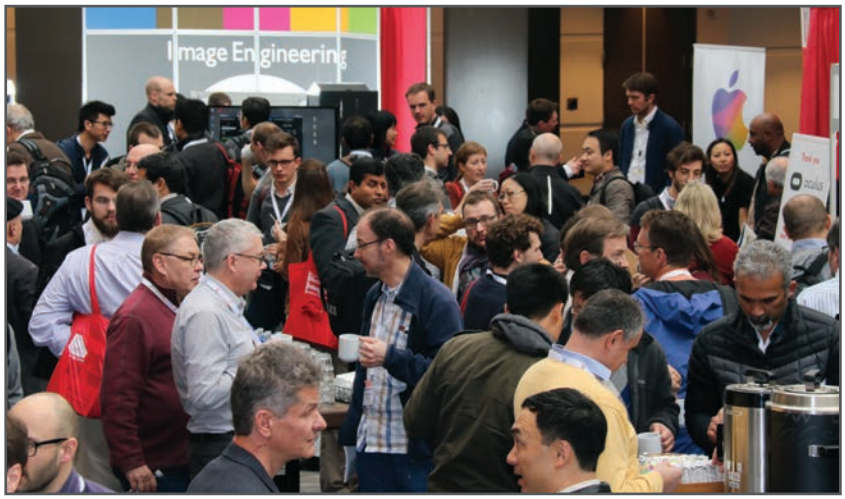
Author Biography

Weijuan Xi earned her PhD in Electrical and Computer Engineering from Purdue University in 2018. She is currently a researcher at Google Labs in Mountain View, CA, where her work focuses on Multimodal Large Language Models (MLLMs) and the deployment of production-scale generative AI.

JOIN US AT THE NEXT EI!

electronic IMAGING

Imaging across applications . . . Where industry and academia meet!



• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

ADVANCING IMAGING SCIENCE 75 1947 IS&T 2022 imaging.org