Video is worth a thousand images: exploring the latest trends in long video generation

Article

Published Version

Creative Commons: Attribution 4.0 (CC-BY)

Open Access

Waseem, F. and Shahzad, M. ORCID: https://orcid.org/0009-0002-9394-343X (2025) Video is worth a thousand images: exploring the latest trends in long video generation. ACM Computing Surveys, 58 (6). 154. ISSN 1557-7341 doi: 10.1145/3771724 Available at https://centaur.reading.ac.uk/124876/

SURVEY

Video is Worth a Thousand Images: Exploring the Latest Trends in Long Video Generation

FARAZ WASEEM, University of Reading, Reading, Berkshire, U.K.

MUHAMMAD SHAHZAD, University of Reading, Reading, Berkshire, U.K.

PDF Download
3771724.pdf
05 January 2026
Total Citations: 0
Total Downloads: 584

Published: 06 December 2025
Online AM: 10 October 2025
Accepted: 29 September 2025
Revised: 18 September 2025
Received: 22 December 2024

Citation in BibTeX format

ACM Computing Surveys, Volume 58, Issue 6 (April 2026)
https://doi.org/10.1145/3771724
EISSN: 1557-7341

Video is Worth a Thousand Images: Exploring the Latest Trends in Long Video Generation

FARAZ WASEEM, Department of Computer Science, University of Reading, Reading, RG6 6DH, United Kingdom of Great Britain and Northern Ireland
MUHAMMAD SHAHZAD, Department of Computer Science, University of Reading, Reading, RG6 6DH, United Kingdom of Great Britain and Northern Ireland

An image may convey a thousand words, but a video, composed of hundreds or thousands of image frames, tells a more intricate story. Despite significant progress in multimodal large language models (MLLMs), generating extended videos remains a formidable challenge. As of this writing, OpenAI's Sora [1], the current state-of-the-art system, is still limited to producing videos of up to one minute in length. This limitation stems from the complexity of long video generation, which requires more than generative AI techniques for approximating density functions. Critical elements, such as planning, narrative construction, and spatiotemporal continuity, pose significant challenges. Integrating generative AI with a divide-and-conquer approach could improve scalability for longer videos while offering greater control. In this survey, we examine the current landscape of long video generation, covering foundational techniques such as GANs and diffusion models, video generation strategies, large-scale training datasets, quality metrics for evaluating long videos, and future research areas to address the limitations of existing video generation capabilities. We believe it would serve as a comprehensive foundation, offering extensive information to guide future advancements and research in the field of long video generation.

CCS Concepts: • **General and reference → Surveys and overviews**; • **Computing methodologies → Image processing**; **Computer vision**; **Video summarization**; **Artificial intelligence**; **Generative and developmental approaches**; **Temporal reasoning**; **Machine learning**; **Modeling and simulation**; • **Information systems → Multimedia information systems**; **Search engine architectures and scalability;**

Additional Key Words and Phrases: Survey, Text-to-video generation, text-to-image generation, generative AI, video editing, temporal dynamics, scalability in AI, artificial general intelligence, AI models generalization

ACM Reference Format:
Faraz Waseem and Muhammad Shahzad. 2025. Video is Worth a Thousand Images: Exploring the Latest Trends in Long Video Generation. *ACM Comput. Surv.* 58, 6, Article 154 (December 2025), 35 pages. https://doi.org/10.1145/3771724

1 Introduction

The year 2022 marked a significant milestone in the field of the generative AI era with the release of ChatGPT [2]. ChatGPT is an advanced language model that produces human-like text from user input, supporting tasks like answering questions, creative writing, and conversation. This

Authors' Contact Information: Faraz Waseem, Department of Computer Science, University of Reading, Whiteknights, Reading, RG6 6DH, United Kingdom of Great Britain and Northern Ireland; e-mail: faraz.waseem@pgr.reading.ac.uk; Muhammad Shahzad (corresponding author), Department of Computer Science, University of Reading, Whiteknights, Reading, RG6 6DH, United Kingdom of Great Britain and Northern Ireland; e-mail: m.shahzad2@reading.ac.uk.

ACM 0360-0300/2025/12-ART154
https://doi.org/10.1145/3771724

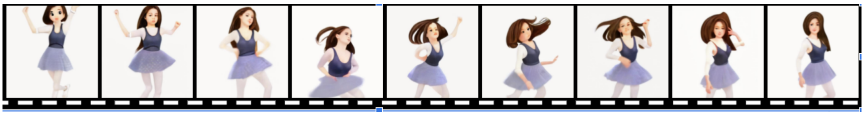


Fig. 1. Example of semantic content not changing with the progress of frames [10].

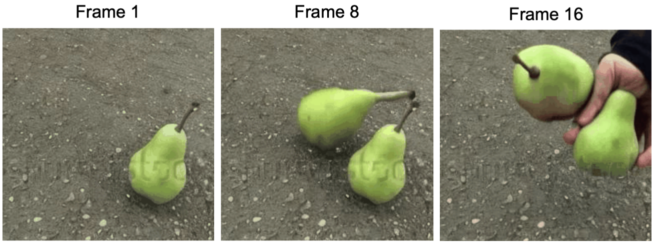


Fig. 2. Example of semantic content changing with the progress of frames [11].

technology uses complex deep neural networks based on large language models trained on extensive text data to capture intricate linguistic patterns and contextual nuances for precise text generation and understanding. Since then, major tech companies have introduced their **large language models** (**LLMs**), such as Facebook's LLama series [3], Google's Gemini [4], and a few other notable models, including Claude [5] and Mistral [6].

The success of LLMs brought a transformative breakthrough in image generation. DALL-E 2 [7] surpassed traditional GANs and VAEs by interpreting natural language and generating diverse concepts and styles, excelling in photorealistic outputs. Other systems, such as Stable Diffusion 3 [8] and MidJourney [9], also demonstrate strong capabilities in creating realistic visuals.

Video generation is far more complex than text or images due to dynamic elements like motion, occlusion, and evolving semantic content–the conceptual mapping of objects, actions, and interactions. Single-scene videos (e.g., a girl dancing against a static background, Figure 1) maintain consistent semantics, while multi-scene videos (Figure 2) introduce new objects or actors, altering semantic content over time.

Due to the complexities of dynamic scenes, early video generation models were limited to producing short clips lasting only a few seconds, often animating a single static frame without incorporating varying backgrounds or objects. For example, Make-A-Video [12] and RunwayML Gen-2 [13] generate 4-5 second videos using a single animated frame with little change in semantic content. CogVideo [14] is among the first long video generation models to create extended videos using autoregressive transformers. However, it operated based on a single prompt and also exhibited minimal changes in the semantic content of the video. Phenaki [15], which employs autoregressive video transformers, is one of the first models to generate long videos with dynamic semantic content based on multiple prompts. Similarly, Gen-L-Video [16] employs a diffusion model to merge short video clips into a seamless, continuous video. Sora [1] has established a new state-of-the-art in video generation. Sora [1], developed by OpenAI, is a "ChatGPT moment" for video generation, utilizing diffusion transformers [17] to sample from a compressed spatiotemporal space, producing photorealistic, coherent videos with complex dynamics. Operating in a similar tier, RunwayML's Gen-4 Alpha [18] is a commercial diffusion transformer model generating 10-second videos, marking progress toward practical use. Despite these advances, models remain nascent compared to human-made videos, struggling with extended coherence, consistent characters, and narrative organization over long sequences (Figure 4).

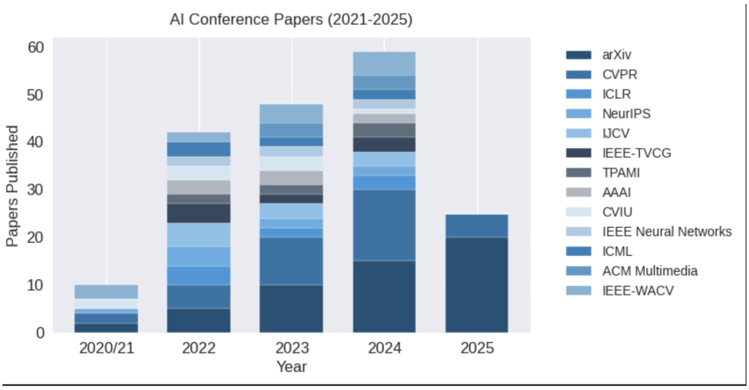


Fig. 3. Most articles focusing on long video generation were published in 2023–2025.

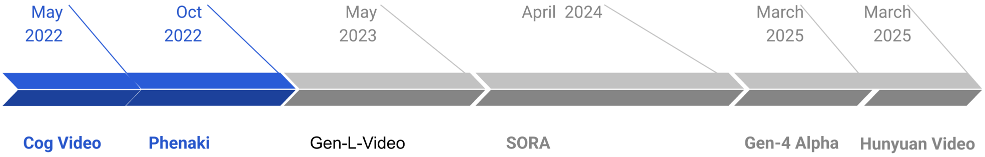


Fig. 4. Evolution of long video generation models. Later models, such as SORA [1] and Gen-4 Alpha [18], focus more on video quality than duration.

1.1 Survey Contributions—Need for Summarizing Long Video Generation Methods

Long video generation presents numerous challenges beyond crafting and maintaining consistent storylines across scenes. These include the lack of large-scale video datasets with detailed captions and the requirement of significant computational resources. Despite these challenges, long video generation has emerged as a transformative frontier in generative AI, unlocking novel possibilities in various fields, including entertainment, education, healthcare, marketing, and gaming. This potential has sparked substantial research attention, and publication rates have accelerated dramatically in recent years. As illustrated in Figure 3, the field has experienced explosive growth, with more than two-thirds of all academic work produced within the past 24 months—a testament to its dynamic transformation.

Given these interests and opportunities, it is time to summarize the state-of-the-art in long video generation and discuss the associated challenges, progress, and future directions that will support the advancement of the field. To our knowledge, there are only two related long video-generation surveys [19] and [20]. The former [19] explores the latest trends in long video generation, highlighting the divide-and-conquer and autoregressive approaches as two primary themes. It also examines photo-realism trends and generative paradigms, such as VAE, GAN, and diffusion-based models. However, while it highlights the divide-and-conquer approach, which simplifies the complexity of long videos by breaking them into smaller, manageable chunks, a detailed exploration of this methodology, such as how short videos can be seamlessly integrated into longer narratives, is lacking. Our work aims to bridge this gap by thoroughly analyzing the various dimensions of the divide-and-conquer strategy and its role in addressing the challenges of long video generation. In contrast, the latter [20] provides a broader overview of the video generation field, encompassing topics such as long videos, video editing, super-resolution, datasets, and metrics. However, long-video generation is only one of many topics discussed. It highlights the need for a focused and in-depth study on the generation of long videos. Our contribution addresses this need by providing a comprehensive analysis of this emerging field and highlighting key approaches, challenges, and prospects.

This work addresses literature gaps by incorporating recent studies and providing a comprehensive analysis of the divide-and-conquer approach. We focus on underexplored aspects, such as agent-based networks and methods for transitioning from short to long videos, which are currently absent from reviews. Our examination extends to methodologies outside the divide-and-conquer and autoregressive categories, as well as the latest research, offering a more holistic perspective on the field.

1.2 Survey Focus—Techniques, Challenges and Key Questions in Long Video Generation

Video generation utilizes various techniques, such as sampling from latent space [21], creating small video segments or images, generating intermediate frames ("divide and conquer")[3.2], employing autoregressive methods [3.1] to predict future frames based on initial ones, and improving latent state representations for longer videos. Training video generation models presents challenges due to the higher computational requirements and more extensive memory needs for video datasets. Many video generation models are based on pretrained image models [[22–24]], which enhance attention mechanisms to ensure consistency between adjacent frames, as video is essentially a sequence of frames. Some long video models are developed by extending short video generation models[[16, 25, 26]] and improving control mechanisms for longer content. Another significant aspect of video generation is input guidance[[5.1, 5.3, 5.2]]. Long video generation requires stronger guidance than images or short clips, typically anchored in text embeddings, such as CLIP [27]. Here, LLMs (Section 3.2.1) take center stage: they decode physical world dynamics, forecast object interactions, and choreograph multistep actions, leveraging their pre-trained knowledge to steer generation toward coherent, long-form outputs. When evaluating the quality of the generated videos[[7.1, 7.2, 7.3, 7.4]], it is important to assess the quality of the individual frames, the fluidity of motion, and the overall aesthetic appeal. Ensuring that generated videos remain faithful to the input text while preserving entity consistency (e.g., cars and actors) across frames is a critical challenge. Long Video generation has motivated researchers to explore novel directions in the field, raising key questions that warrant further investigation.

(1) How can we generate long videos with multiple semantic segments with different actors, actions, and objects?
(2) How can we ensure semantic consistency across long video segments, such as maintaining consistent models of objects like cars?
(3) Discussion of Long-video generation strategies covering segmented stitching, auto-regressive frames, and full latent-space synthesis.

Our survey article centers around these critical questions, providing insights to guide researchers and practitioners in addressing these challenges.

1.3 Survey Methodology

For this survey, we conducted searches across several conferences, including but not limited to CVPR, ICLR, NeurIPS, IJCV, IEEE-TVCG, TPAMI, AAAI, CVIU, IEEE Neural Networks, ICML, ACM Multimedia, IEEE-WACV. We used keywords such as "video generation," "long video generation," and "LLM-guided video generation." Additionally, we searched academic databases, including arXiv, Google Scholar, IEEE Xplore, ACM Transactions, and Scopus, focusing on the term "long video generation" for our survey. Our survey covers articles published between 2021 and 2025 (as of May 2025), with a focus on "video generation" and "generative AI". We gathered over 190+ articles through snowball sampling, using keywords such as text-to-video, generative AI, visual interpretation, and extended video generation.

1.4 Survey Organization

We will begin by discussing the foundational frameworks for video generation, including embedding and LLMs, to set the stage for more advanced topics. The goal is to familiarize readers with these fundamental components, enabling them to explore these building blocks according to their level of expertise and interest. Next, we will explorebackbone mechanisms for video generation, including divide-and-conquer autoregressive models and the use of implicit latent spaces. We will then explore Tokenization Strategies. We will explore input guidance mechanisms, including strategies such as LLM guidance, and categorize them into different levels based on the depth of control the LLM exerts over them. We will also address the necessary modifications to the image and video diffusion models to facilitate such control. We will also discuss the post-processing pipelines required to achieve high temporal and spatial quality in videos generated by diffusion models. We will then discuss the datasets used to train video generation models, as outlined in Section datasets. We will then discuss the metrics used to measure generated video quality, as outlined in Section metrics. Finally, we will talk about future trends and open challenges.

2 Long Video Generation: Backbone Architectures and Methods

Progress in long-video generation builds on advancements in many foundational building blocks. These include GANs-based architecture Section 2.1, Autoencoders Section 2.2, Transformers-based models Section 2.3, LLMs and language understanding Section 2.4, and Image and Video Diffusion models Section 2.5.

2.1 GAN-based Video Generation

GANs [28] dominated generative tasks from 2014 until the early 2020s, although diffusion models and transformer-based approaches in terms of performance and versatility have since surpassed them. The fundamental GAN framework [28] consists of two competing components: a Generator that creates samples from random noise and a Discriminator that evaluates their authenticity. While this adversarial architecture established the foundation for image and video generation, newer methods have advanced beyond its capabilities, as discussed in subsequent sections.

GAN Based Image Generation GANs initially revolutionized image generation, dominating the field. Here, we examine key ideas and milestones in GAN literature, organized by timeline.

Early GANS: GANs [28] was the first to generate images using adversarial networks, but employed simple feedforward neural networks for both the Generator and the Discriminator. DCGANs [29] extend the GAN architecture by incorporating convolutional layers, making them more suitable for image data. They generated images with a resolution of 64×64. LAPGAN [30] increased the resolution of images by developing them at multiple scales. It consists of various GANs, each generating images at different resolutions. GANs, DCGANs, and LAPGAN are primarily designed to create images based on random noise vectors. These models lack text guidance, but text-based GANs incorporate text-based control, which we will discuss in the next section.

GAN Text Input: StackGAN [31] is a multistage text-to-image GAN that generates high-quality images. It has two GAN stacks stacked on top of each other. The first one takes the text and generates a low-resolution image. The second one takes both text and input images and creates high-quality images. AttnGAN [32] uses attention mechanisms to create images from text, allowing it to focus on specific words or phrases in the input description.

Style/Image Transfer: StyleGAN [33] was the first generative model to generate high-quality artistic images, and one of the key innovations was transferring an artistic style like Vincent Van Gogh's to real-world pictures and image translation. CycleGAN [34] does image-to-image translation and consists of two generators and two discriminators. StyleGAN2 [35] primarily focuses on generating high-quality, diverse images, particularly faces. It introduced a disentangled latent space. Latent space is where a vector of N dimensions represents each image. Projecting different high-level attributes, such as skin color and hairstyle, onto distinct dimensions in a latent space provides excellent editing capabilities for realistic image generation, semantic manipulation, and local editing. StyleGAN2[35] opened the doors for high-level image manipulation. StyleGAN2 [35] is an improvement over StyleGAN, producing higher-quality images. pix2pix [36] specifically designed for image-to-image translation tasks. It learns a conditional generative model and generates an output image conditioned on the input image. GAN also revolutionized video generation, which we will explore in Section 2.1.1.

2.1.1 *Video/Multi Frame Generation.*

Early Attempts: [37], which generated future frames from observed sequences. [38] advanced this by using separate 2D and 3D convolutional networks for static backgrounds and moving foregrounds, producing 32-frame unconditional videos of various scenes. Further development came with [39]'s two-stage model, which first generated 128×128 resolution time-lapse videos from a single frame and then enhanced them with dynamic motion information. These early works laid the important foundations for unconditional video generation before the advent of prompt-based approaches.

Prompt-based Guidance: Numerous studies have explored the use of conditional inputs in GAN to guide and refine the generation process. These conditions can take various forms, including audio signals, text prompts, semantic maps, images, or other videos. TGANs-C [40] incorporate text guidance using LSTM-based latent vectors. TGANs-C was designed to input a single sentence.

Long Video Generation Using GAN: DIGAN [41] can create a 128-frame video. It introduced an INR (Implicit Neural Representations) based video generator that improves motion dynamics by manipulating space and time coordinates differently and a motion discriminator that efficiently identifies unnatural motions without requiring long frame sequences. StyleGan-V [42] improved the state-of-the-art and was built on StyleGAN2 [35]. It can generate high-resolution 1024-long videos by designing a holistic discriminator that aggregates temporal information by simply concatenating frame features, thereby decreasing the training cost.

2.2 Autoencoder-based Video Generation

Autoencoders [43], variational autoencoders [44], and masked autoencoders [45] belong to the family of models that compress information into a compact latent space and serve as building blocks for image and video generation pipelines. Masked autoencoders can generate videos from this learned latent space. We will discuss the foundations of autoencoders and build up the video generation process via masked autoencoders.

2.2.1 *Autoencoder Formulation.* An Autoencoder is an unsupervised neural network that compresses its input into a compact latent layer and then learns to reproduce its input through back-propagation. The autoencoder is trained to minimize the reconstruction loss between the input $\mathbf{x}$ and the reconstructed output $\hat{\mathbf{x}}$. The most common application of an autoencoder for long videos and video generation is the construction of a compressed latent space. For example, in [46], the authors use an encoder and decoder to project images from pixel space to latent space, thereby decreasing the computational complexity of learning the image distribution.

Variational Autoencoders (VAEs) [44] address the limitations of traditional autoencoders by learning latent distributions instead of fixed representations, allowing new data generation. The VQ-VAE variant [47] has become foundational for video/image generation pipelines like VideoGen [48], VQGAN [49], and DALL-E [50]. Hybrid approaches like Hierarchical Patch VAE-GAN [51] combine VAEs with GANs, while applications extend to anomaly detection through architectures like LSTM-Convolutional VAEs [52].

2.2.2 Masked Autoencoders. Masked autoencoders [45] serve as scalable self-supervised backbones for video generation by reconstructing randomly masked image patches. The approach extends to video through models like VideoMAC [53], which uses convolutional networks to reconstruct symmetrically masked frame pairs at high masking ratios (0.75). The framework has evolved into advanced variants, such as MAGVIT [54], which achieves significantly faster inference than diffusion models, and MAGVLT [55], which unifies vision-language generation under this paradigm.

2.3 Transformer-based Video Generation

GANs have limitations, such as mode collapse [56], training instability, which requires fine-tuning of parameters, and a significant amount of training time and resources. Transformers, introduced in 2017 [57], made inroads into image and generation via autoregressive and masked encoding. Some of the key concepts to understand are Vision Transformers and Video Transformers.

2.3.1 Transformer-based Image Generation. **Vision Transformers (ViTs)** [58] revolutionized computer vision by processing images as patch-based tokens, similar to NLP transformers. DALL-E [59] pioneered this approach for image generation, using a Discrete VAE [60] to compress images into 32×32 tokens and training a GPT-style transformer on 250M image-text pairs. CogView [61] later surpassed DALL-E in FID scores but maintained weaker complex prompt rendering. Both autoregressive models suffered from slow generation due to token-by-token processing, a limitation addressed by CogView2 [62] through masked cross-modal training.

2.3.2 Autoregressive-based Video Generation. **Video transformers (VViTs)** [63] extend ViTs [58] by tokenizing video patches. Phenaki [64] generates long videos from text prompts using T5X embeddings [65] and C-ViViT, a variant of VViT [63] that compresses tokens and employs masked and autoregressive prediction for long sequences. CogVideo [14] builds on CogView2 [62], using hierarchical training for better text-video alignment and a two-stage process involving keyframe generation and interpolation.

2.4 Language Understanding in Video Generation

2.4.1 Text to Image Feature Representation. The core principle behind text-based visual generation tasks is effectively pairing text with the visual content. Many visual generation pipelines leverage pre-existing image-text pair models, such as CLIP (Contrastive Language-Image Pretraining) [27]. CLIP has been pre-trained using a contrastive learning approach that optimizes the cosine similarity between image and text embedding. Given CLIP's robust performance, many visual generation models, such as DALL·E 2 [7], incorporate CLIP's text embedding to leverage its superior semantic understanding. It allows these models to enhance their ability to generate visually relevant and contextually appropriate content, effectively bridging the gap between text and visual representation.

2.4.2 LLMs-based Video Guidance. Many visual generation models, such as LLM Director [66], leverage standalone LLMs [67] to enhance their performance. By integrating LLM, visual generation models can benefit from advanced natural language processing capabilities, enabling them to interpret and generate more nuanced and contextually relevant descriptions of captions in single or multiple prompts, along with detailed scenes. One example of this design is LLM-grounded VDM [68]. When paired with visual inputs, LLMs can transform simple image descriptions into more elaborate storytelling, adding layers of meaning and context that enhance the viewer's experience. LLM can also act as the director of the entire video generation process and create a coherent script, as shown by Vlogger [69]. The details on how the recent long video generation methods leverage LLMs are explained in Section 3.2.1.

2.5 Diffusion Models

Diffusion models have emerged as the state-of-the-art approach for image and video generation, combining components such as VAEs, transformers, and language models. The foundational work [70] established key principles by applying non-equilibrium thermodynamics to unsupervised learning, while [71] advanced the field through parameterized Markov chains trained by variational inference. These breakthroughs created the basis for modern diffusion architectures in visual generation tasks.

2.5.1 Image Diffusion. Image diffusion models generate images through iterative denoising, with [72] contributing gradient-based estimation methods and [71] establishing the foundational DDPM framework. For a text-conditioned generation, models typically employ a U-Net with cross-attention layers using embeddings from CLIP [27], BERT [73], or T5 [74]. Robin Rombach et al. implemented this in [46] through modified attention layers that combine multimodal embeddings [57].

2.5.2 Video Generation from Diffusion Models. Video generation models primarily use two architectures: 3D U-Nets and Transformers. The U-Net approach extends 2D diffusion models to handle 4D tensors (frames × height × width × channels) through factorized spatial-temporal attention, where spatial attention focuses on intra-frame regions and temporal attention captures inter-frame dependencies. Alternatively, Sora [75] implements a Transformer-based diffusion model. A **Diffusion Transformer** (**DiT**) [17] replaces the traditional U-Net backbone in a diffusion model with a Transformer architecture, which is better at processing images as sequences of patches. Sora [75] generates videos efficiently by compressing them into latent spacetime patches that capture both appearance and motion, serving as visual tokens for video construction [17] that processes videos as spacetime patch tokens, as detailed in [75]. These architectural approaches for long-video generation are further explored in subsequent sections.

3 Long Video: Generation Paradigm

We summarize various video generation approaches into three core paradigms.

- Auto-Regressive Paradigm: Videos are generated sequentially, with each frame conditioned on the frames generated before.
- Divide-and-conquer approach: Videos are produced by creating keyframes or short video segments guided by storyline prompts, often with the aid of an LLM.
- Implicit Video Generation: Videos are generated implicitly from the model without needing explicit extrapolation (autoregressive approach) or explicit interpolation (divide-and-conquer) by designing a latent space to represent variable-size videos.

These approaches will be explained in the following sections.

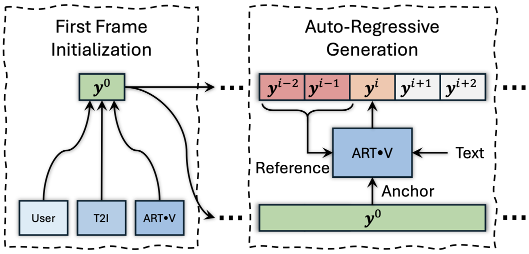


Fig. 5. The basic theme of the auto-regressive approach is that it generates new frames, given the initial anchor frame, previous frames, and optional progressive prompts [81].

3.1 Auto Regressive Approaches

The autoregressive generation paradigm creates videos by sequentially predicting future frames from previous ones, ensuring temporal coherence through this frame-by-frame approach [[15, 76, 77]]. While effective for maintaining consistency (Figure 5), this method faces computational limitations for long videos due to its sequential nature. Key implementations include CogVideo [14], which extends CogView2 [78] but is constrained by sequence length and resolution (160×160, upscalable to 480×480), and NUWA-Infinity [79], which improves resolution through hierarchical generation. Phenaki [15] advances the paradigm by handling multiple prompts through C-ViViT [63] compression and T5X embeddings [65], though its bidirectional training requires significant memory.

A paradigm shift emerged with VideoPoet [80], which demonstrated that multimodal training—combining text, images, and audio in a decoder-only transformer—could overcome these quality limitations. By pretraining on diverse objectives and fine-tuning for specific tasks, VideoPoet achieved state-of-the-art zero-shot generation, proving that autoregressive models can produce high-fidelity videos when augmented with rich multimodal signals.

All approaches discussed [[15, 76, 79]] are based on transformers. Grid Diffusion [82] is based on diffusion. Grid Diffusion first used compression and represented video using an image created from keyframes, which covers the primary motions or events of the video. It is called a ‘grid image,’ which consists of 4 subframes representing video keyframes. During the training phase, they masked these frames and learned to produce masked frames conditioned on previous grid images and non-masked images. This design paradigm is illustrated in Figure 6. As they replaced the challenge of video generation with image generation, they can create long videos up to 128 frames autoregressively with high image quality (low FVD scores [83]. They utilized transfer learning from a pre-trained stable diffusion model [46] and trained on only two Nvidia A100 GPUs. Figure 6 explains this architecture. Building on this insight, ARLON [84] combines the strengths of autoregressive transformers and diffusion models through its **Asymmetric Diffusion Transformer** (**AsymmDiT**) and latent VQ-VAE, achieving 128× compression (8× spatial + 6× temporal downsampling) in a 12-channel latent space. This hybrid approach maintains motion fidelity while enabling scalable, high-quality synthesis—effectively bridging diffusion models (Section 2.5) and token-based methods.

Long video generation faces critical challenges in memory management and temporal coherence across extended sequences. ARLON [84] addresses this via a training-free autoregressive inference method using pre-trained diffusion models. Its sliding-window queue mechanism processes frames with progressively increasing noise levels: fully denoised frames are removed from the head of the queue, while new noisy frames are added to the tail. This approach enables the unbounded generation of infinitely long videos without retraining, maintaining computational efficiency through a fixed-size queue that prevents memory overload.

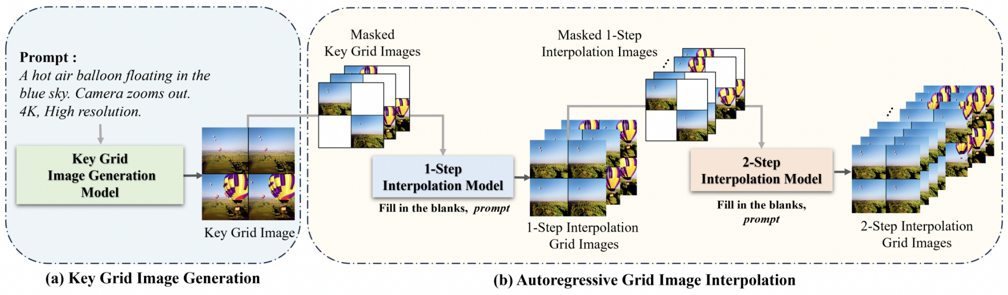

Prompt :
A hot air balloon floating in the blue sky. Camera zooms out. 4K, High resolution.
Key Grid Image Generation Model
Key Grid Image
Masked Key Grid Images
1-Step Interpolation Model
Fill in the blanks, prompt
Masked 1-Step Interpolation Images
1-Step Interpolation Grid Images
2-Step Interpolation Model
Fill in the blanks, prompt
2-Step Interpolation Grid Images
(a) Key Grid Image Generation
(b) Autoregressive Grid Image Interpolation

Fig. 6. Grid Diffusion Model. It first generates a grid image and then learns a spatial auto-regressive model by learning to predict masked subframe conditions on previous images and unmasked frames. [82].

Previous methods for video generation, including NUWA-Infinity [79], CogView [76], Phenaki [15], and GRID [82], face limitations in rendering complex compositional prompts that describe dynamic spatiotemporal interactions—such as "a man walking with a black dog on his right while a blue car drives from the opposite direction." VideoTetris [77] addresses this challenge by introducing spatiotemporal Compositional Diffusion, which manipulates cross-attention maps in denoising networks to synthesize videos that adhere to intricate or evolving instructions. This approach enables the generation of long videos with progressive compositional prompts, where "progressive" refers to continuous changes in object positions, quantities, and attributes, ensuring precise alignment of interacting entities across space and time.

Autoregressive video generation has progressed from CogVideo [76]'s single-prompt, low-resolution outputs to modern systems like VideoTetris, which can model complex scientific dynamics with multiple prompts while preserving quality. While autoregressive methods excel at smooth motion transitions, their sequential nature results in slow generation and limited control over complex scene elements (actors, bounding boxes, spatial relationships). Divide-and-conquer approaches (Section 3.2) address these limitations by enabling parallel frame generation and leveraging LLMs for structured video blueprints, improving the handling of dynamic scenes. Key articles and insights are cataloged in Table 1.

3.2 Divide and Conquer Paradigms

The divide-and-conquer approach generates keyframes or short clips from prompts and interpolates between them, often using an anchor image as a reference. Each keyframe is generated independently, enabling parallel processing. Challenges include maintaining semantic consistency, ensuring smooth motion, and achieving high quality. A key theme is the separation of planning and video generation stages, differentiating it from autoregressive methods (Figure 8). Its paradigms are: LLM as Director, Intermediate Transition Model, and Agent-Based Framework. Some milestone articles with timelines are illustrated in Figure 7.

3.2.1 LLM as Director. The LLM-as-Director paradigm [11, 94, 97, 98] revolutionizes video generation by employing a two-stage process: (1) an LLM Planner creates detailed narrative blueprints (keyframes, layouts, actions) and (2) a Video Generator Backbone produces intermediate frames (Figure 9). This framework supports both zero-shot (training-free) and training-based approaches.

Free-Bloom [22] demonstrates zero-shot capability through innovative techniques like joint noise sampling and DDIM-based dual-path interpolation [99], though its LLM scripting potential is

Table 1. Auto Regressive Approaches

Model	Theme	Month/Year
StyleGAN-V [85]	Time-continuous signals [86] [87] extended from StyleGAN2 [35].	Dec 2021
DIGAN [41]	Implicit neural representation-based [86] video generation model.	Feb 2022
CogVideo [14]	Long videos using autoregressive and interpolation stages.	May 2022
NUWA-Infinity [79]	Long videos with hierarchical autoregressive modeling.	July 2022
Phenaki [15]	Compresses videos into discrete tokens for efficient frame generation.	Oct 2022
PVDM [88]	PVDM uses diffusion in latent space for video generation.	Feb 2023
MeBT [89]	Memory-efficient transformer for long-range dependency videos.	March 2023
ART·V [81]	Auto-regressive using keyframes and image diffusion.	Nov 2023
StreamingT2V [90]	Long videos with consistent transitions and scene preservation.	March 2024
Grid Diffusion Models [82]	Video generation by merging four keyframes into images.	March 2024
ViD-GPT [91]	GPT-style autoregressive generation into video diffusion models.	June 2024
FlexiFilm [92]	Long videos with temporal conditioning and resampling strategy.	June 2024
VideoPoet [80]	An autoregressive LLM for high-quality synthesis from multi modal inputs.	June 2024
VideoTetris [77]	Text-to-video generation with spatio-temporal compositional diffusion.	Oct 2024
Arlon [84]	AR for long-range temporal guidance and DiT for high-fidelity synthesis.	Jan 2025

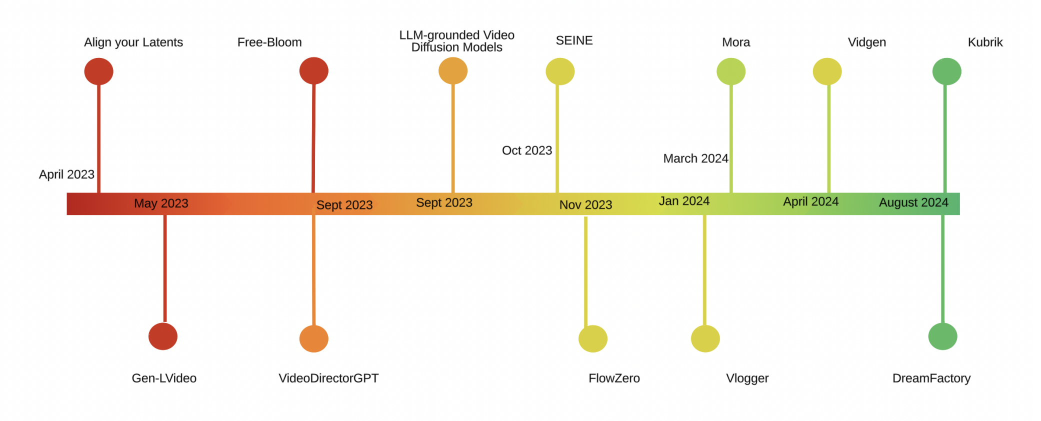


Fig. 7. Divide-and-conquer timeline: We used the dates these articles were published in online resources, such as arXiv or https://openreview.net/. Articles catalog here are Align your Latents [24], Gen-L-Video [16], Free-Bloom [22], VideoDirectorGPT [11], LLM-grounded Video Diffusion Models [68], SEINE [26], FlowZero [23], Mora [93], Vlogger [69], Vidgen [94], DreamFactory [95], and Kubrik [96].

underutilized. VideoDirectorGPT [11] enhances this with GPT-4's comprehensive planning (layouts, bounding boxes) executed via Layout2Vid [97]. The training-based LLM-grounded model [100] further improves realism by learning spatiotemporal dynamics from the text.

Like VideoDirectorGPT, FlowZero [23] adopted a zero-shot (training-free) approach; however, the LLM plays a more detailed role than VideoDirectorGPT. It generates a detailed **dynamic scene syntax** (**DSS**), including scene descriptions, object arrangements, and background motion patterns. The DSS components direct an image diffusion model to generate videos with smooth object movements and consistent frame transitions. The theme of using dynamic scene layout is illustrated in Figure 8. The LLM as a director approach has limitations, as it is a two-stage architecture, and adding speech or integrating short clips will require modifications in the pipeline. We can extend the LLM-based divide-and-conquer approach by incorporating more specialized components, such

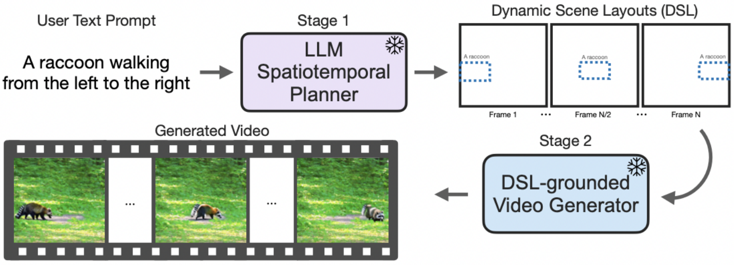

User Text Prompt
A raccoon walking from the left to the right
Stage 1
LLM Spatiotemporal Planner
Dynamic Scene Layouts (DSL)
A raccoon
Frame 1
Frame N/2
Frame N
Stage 2
DSL-grounded Video Generator
Generated Video
...

Fig. 8. The LLM as director approach utilizes LLM as the spatiotemporal director of the script, along with a separate video generation module that can understand the DSL (metadata) generated by LLM [68].

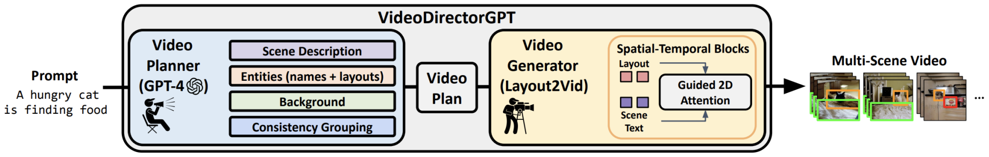


Fig. 9. VideoDirectorGPT: GPT-4 generates a blueprint for video generation, including scene and entity description. Separate module Layout2Vid generates video from this video plan [11].

as a model for generating reference images, a module for video creation, and a plug-and-play module for adding transitional clips and speech. That is achieved using the multi-agent framework, as described in Section 3.2.2.

3.2.2 *Multi-stage/Agent-based Divide and Conquer Approach.*

An Agent-based LLM Framework [101] is a system where LLM serves as the "brain," responsible for overseeing complex operations, while simpler models function as tools, executing more specific, supportive tasks. A multi-agent framework for video generation represents a multi-layered approach to text-to-video generation. It produces high-quality long videos like those generated by Sora [102] by dividing the video creation challenge into multiple systems, each specializing in some aspect of the video generation pipeline, as illustrated in Figure 10.

3.2.3 *Divide and Conquer Compositional/Transition Approach.*

Early long-video generation methods stitched short clips from standard models (diffusion/autoregressive) but struggled with transitions. SEINE [26] innovated by framing transitions as masked diffusion, jointly denoising overlapping segments conditioned on boundary frames—though still requiring independent segment generation. Subsequent work improved continuity: MEVG [103] anchored new clips to the final frames of predecessors, while MAVIN [25] formalized transition learning as "video infilling" by training on corrupted intermediates. Encoder-Empowered GAN [104] enforced temporal coherence via recall mechanisms but sacrificed dynamic content flexibility.

These incremental advances culminated in VideoMerge [105], which reimagined the paradigm entirely. Instead of post hoc stitching, it preemptively ensures coherence through (1) adaptive noise blending to unify short and long temporal scales, (2) latent fusion for boundary-free transitions, and (3) prompt refinement for persistent identity. By addressing consistency at noise, latent, and semantic levels, VideoMerge [105] achieves what prior segment-and-merge methods could not:

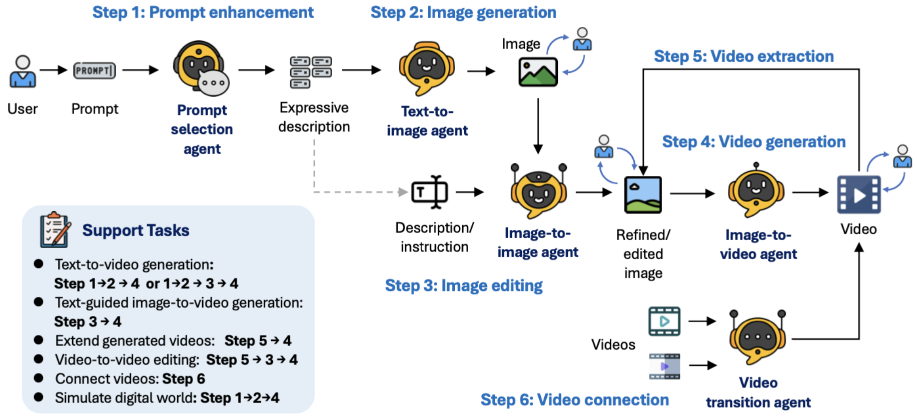


Fig. 10. Mora utilizes a multi-agent framework. The prompt selection agent enhances prompts with detailed instructions, the text-to-image agent generates images from input prompts, the image-to-image agent improves the quality of photos, the text-to-video agent generates video segments, and the video transition agent integrates these videos into a longer video [93].

holistic long-video synthesis without retraining, marking a shift from compositional fixes to native long-form generation.

The autoregressive approach ensures smooth transitions between frames by generating each frame conditioned on the previous ones, but its sequential nature makes it inherently slow for long video generation. The divide-and-conquer approach (see Section 3.2.1) can generate keyframes in parallel but faces challenges involving interpolation between frames with smooth transitions and achieving higher video quality while maintaining semantic consistency. Implicit generation Section 3.3 approaches combine the best of both worlds by generating complete videos directly from a model conditioned on user input without the need for interpolation (divide and conquer) or extrapolation (autoregressive) between frames. A summary of key articles exploring themes related to the Divide and Conquer approach is provided in Table 2.

3.3 Implicit Video Generation Using Compressed Latent Space

Implicit video generation synthesizes complete videos simultaneously through compact latent representations, employing spacetime compression, enhanced attention mechanisms, and hierarchical denoising [21]. Unlike sequential approaches, models like Sora [21] process entire videos via: (1) Spacetime compression to latent patches (visualized in Figure 12), (2) ViT-based denoising, and (3) LLM-augmented CLIP-like conditioning. Open-source advances include FreeNoise [108] for tuning-free semantic preservation, and GLOBER [109] for efficient latent reconstruction. HunyuanVideo [107] Figure 11 advances the field through a diffusion-VAE hybrid architecture in compressed latent space, enabling both quality and long-form coherence–demonstrating implicit generation's unique temporal synthesis capabilities from Sora to multimodal implementations.

Latent-space transformer-based video generation has progressed through key architectural innovations, beginning with Goku [110]'s flow-based transformers for efficient joint image-video learning. Subsequent advances include REDUCIO! [111]'s 3D VAE compression (64× more efficient than 2D VAEs) and Mochi 1 [112]'s 10B-parameter Asymmetric Diffusion Transformer for improved coherence. Meta's Movie Gen[113] is a unified foundation model that generates high-quality images

Table 2. Divide and Conquer Paradigms: Catalog of Key Articles

Model	**Theme**	**Month/Year**	**category**
Align your Latents [24]	Diffusion models for interpolation and upsampling.	April 2023	training-based
Gen-L-Video [16]	Integrate short videos into long, consistent video.	June 2023	training-based
Free-Bloom [22]	LLMs and LDMs [46] for consistent video generation.	Sept 2023	training-free
VideoDirectorGPT [11]	Consistent multi-scene videos using GPT-4 guidance.	Sept 2023	training-free
LVD [68]	Dynamic video scenes using LLM-guided diffusion.	Sept 2023	training-free
SEINE [26]	Long video with smooth transitions from short videos.	Oct 2023	integration
Encoder GAN [104]	Connects short video by temporal relationships.	Oct 2023	integration
FlowZero [23]	Multi-frame story and aligns spatiotemporal layouts.	Nov 2023	training-free
MEVG [106]	Multiple prompts, preserving visual coherence.	Dec 2023	training-based
Vlogger [69]	Specialized models to generate long videos in stages.	Jan 2024	multi-stage
Mora [93]	Collaborative models for script, image, and video.	March 2024	multi-stage
Vidgen [94]	LLM for story pre-processing and textual Inversion Memory Module.	April 2024	training-based
MAVIN [25]	Transition videos creating a cohesive sequence.	May 2024	integration
DreamFactory [95]	LLM collaboration for script and movie creation.	August 2024	multi-stage
Kubrick [96]	Agent collaborations to generate Blender scripts.	August 2024	multi-stage
VideoMerge [105]	training free, merges short clips generated by pretrained text-to-video models.	March 2025	multi-stage

In the Category Column, Training-free or Training-based Represents Section 3.2.1 Pattern, 'multi-stage' Represents the Section 3.2.2 Pattern and 'integration' Represents the Section 3.2.3 Pattern.

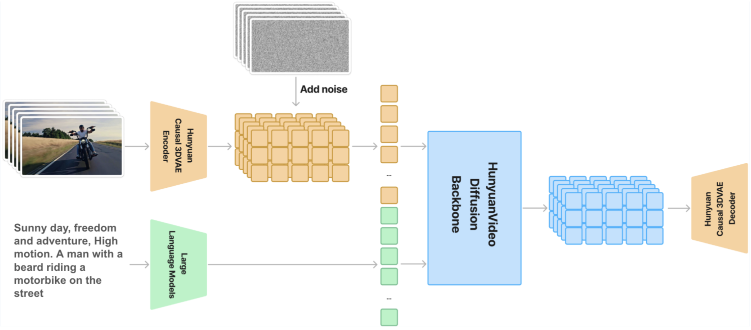

Add noise
Hunyuan Causal 3DVAE Encoder
Sunny day, freedom and adventure, High motion. A man with a beard riding a motorbike on the street
Large Language Models
HunyuanVideo Diffusion Backbone
Hunyuan Causal 3DVAE Decoder

Fig. 11. Using a Causal 3D VAE, Hunyuan Video compresses data into latent space. LLM-encoded text conditions Gaussian noise inputs, generating latents decoded into images/videos via the VAE decoder [107].

and videos from text prompts using efficient joint training in compressed latent space. The field's current pinnacle is Cosmos [84] World Foundation Model with 128× latent compression and two-phase training. However, persistent challenges in motion consistency and semantic alignment remain, evidenced by SORA's [1] one-minute generation limit and documented artifacts [114]. Table 3 compares models using compressed latent spaces for video generation.

Beyond generation strategies, long video modeling requires effective tokenization, which represents videos as compact units for efficient processing, as discussed in the following section.

4 Long Video: Tokenization Strategies

Long video tokenization strategies employ frame-level Section 4.1, Temporal-spatial (3D Conv/VQ-VAE) Section 4.2, and Hierarchical Section 4.3 approaches to efficiently encode spatiotemporal information while balancing computational demands and representation fidelity, which is critical for both generative and discriminative tasks, as detailed below.

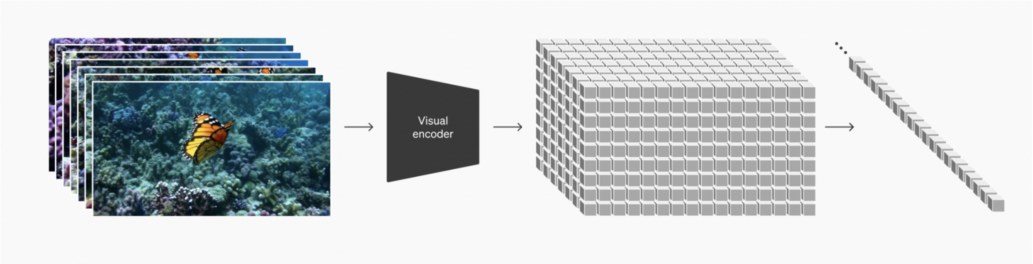


Fig. 12. Transformer-based Diffusion Model Sora compressed video of variable length into a fixed space-time latent compressed representation [75].

Table 3. Implicit Video Generation: Milestone Models Catalog

Model	Theme	Year
GLOBER [109]	Global features to synthesize coherent video frames.	Sept 2023
FreeNoise [108]	Extended videos using pre-trained video diffusion models.	Oct 2023
SORA [21]	Compact latent and patch-based representations.	Feb 2024
Mochi 1 [112]	Asymmetric Diffusion Transformer (AsymmDiT) design.	October 2024
Hunyuan-Video [107]	Open source, MLLM Text Encoder, and multiple video resolution.	Dec 2024
Goku [110]	Flow-based transformer and Vector-quantized VAE.	Feb 2025
REDUCIO! [111]	Radical latent space compression, 3D VAE that compresses videos into ultra-compact motion representations.	Nov 2024
Cosmos [115]	Physical AI based on world foundation models.	March 2025

4.1 Frame-level Tokenization

Early approaches to video generation framed the problem as modeling a sequence of images. The introduction of VQ-VAE [47] was pivotal, as it compressed images into discrete token representations that autoregressive models could efficiently process. Building on this, early video generation systems, such as CogVideo [14], utilized a VQ-VAE [47] architecture to achieve frame-based tokenization for video. NÜWA [116] and HARP [117] are based on VQ-GAN [49] and are also built on frame-level tokenization as shown in Figure 13 [117]. Some other methods [118] represent images not as 2D grids but as compact and highly efficient 1D token sequences, achieving semantically rich representations that are compact for generation. The limitation of frame-level tokenization is the repetition of information in adjacent frames, which will be discussed in Temporal-Spatial Tokenization Section 4.2.

4.2 Temporal-spatial Tokenization

Temporal tokenization compresses video into motion-aware latent representations, a concept pioneered by Phenaki [15] using its C-ViViT architecture [15]. This spatio-temporal approach was also adopted by VideoGPT [48] and Hunyuan-Video [107], which utilize 3D convolutional VAEs. A common implementation involves a VQ-VAE encoder trained on video data that uses 3D convolutions for spatiotemporal downsampling before attention residual blocks. Other frameworks like MAGVIT [54] model dynamics with a 3D-VQGAN architecture, extending the standard VQGAN encoder-decoder—comprising cascaded residual blocks with down- and upsampling layers—into the temporal domain. In contrast, OmniTokenizer [119] first divides data into patches and then uses a decoupled spatial-temporal transformer architecture, leveraging both VAE and VQ-VAE encoders. Although VQ methods often face training instability [120], techniques like Index Backpropagation Quantization [121] can resolve this with a differentiable codebook. While these methods improve

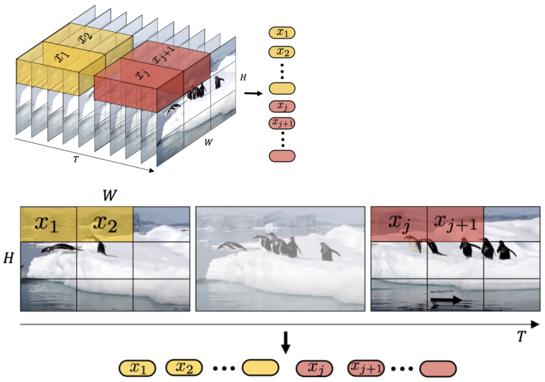
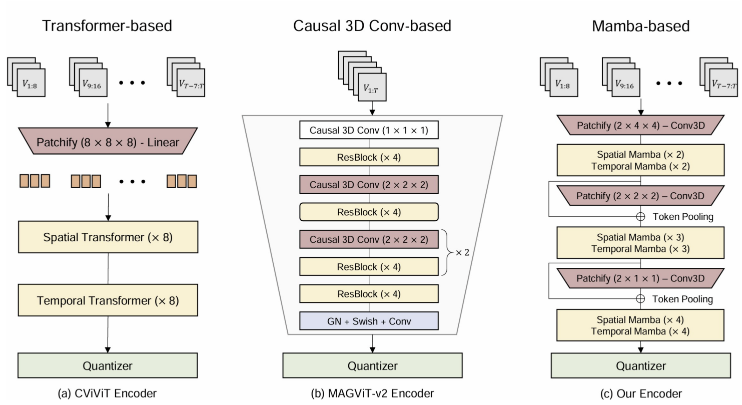


Fig. 13. Comparison of tokenization methods: ViViT [63] (top) uses a spatiotemporal strategy, while the bottom image [63] illustrates a frame-level approach.



(a) CViViT Encoder (b) MAGViT-v2 Encoder (c) Our Encoder

Fig. 14. This figure from [124] compares tokenization strategies. The encoder network of CViViT [15] is a Transformer-based tokenizer, MAGVIT [125]employs a causal 3D convolution-based tokenizer, and the Mamba-based tokenizer [124] introduces a new encoder architecture, with all models designed for spatio-temporal compression [123].

motion capture, modeling the hierarchical storyline of long videos remains a challenge, a topic explored further in Section 4.3.

4.3 Hierarchical Tokenization

Hierarchical video tokenization captures multi-scale spatiotemporal dependencies. Early models, such as HERO [122], employ a two-level transformer to capture both local and global context. HiTVideo [123] utilizes a multilayer codebook to strike a balance between semantics and detail. MambaVideo [124] represents an evolution from prior works like C-ViViT [15] and MAGVIT [125] as shown in Figure 14. It advances spatiotemporal feature encoding through a hierarchical encoder-decoder with 3D convolutions.

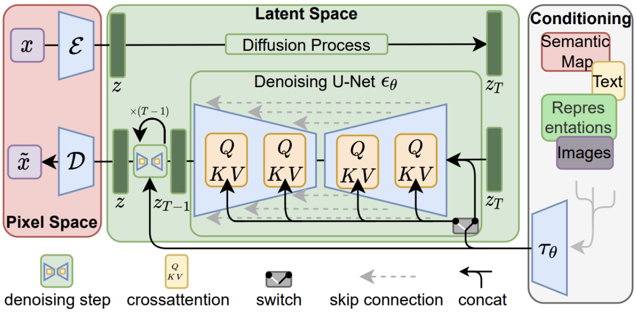


Fig. 15. A latent diffusion model with input conditioning generates data by applying a reverse diffusion process on latent representations, conditioning the generation of additional input information (e.g., text or images) to guide output [46].

In addition to the tokenization strategy, another theme in the long videos is the use of input control mechanisms, such as text, bounding boxes, and images, for video guidance. We will discuss that in the next section.

5 Long Video: Input Control

Input conditioning involves diffusion models, GANs, or autoencoders using signals from user text prompts, entity layouts, bounding boxes, and images to condition video generation. Although long video generation utilizes many of the same techniques for input control as image and video generation models, it also faces the additional challenge of preserving long-term dependencies. Video generation models utilize innovative strategies like the use of LLM to create progressive prompts from a single input prompt [11, 22, 68, 126, 127] and enhancement in generation mechanism to create semantic consistency between frames [22, 23, 126].

Popular mechanisms for input conditioning of long videos include User Textual Prompt, User Textual Prompt with Scene Layout, and Image Input with Textual Prompt and Scene Layout, which we will discuss next.

5.1 User Textual Prompt

Text prompts serve as the primary conditioning method for video generation models, with implementations ranging from single-prompt to multi-prompt approaches. Early transformer-based models, such as DALL-E [50] and CogVideo [10], utilized autoregressive transformers on joint text-image token distributions, albeit with limitations to single prompts. Phenaki [64] advanced this by incorporating T5X embeddings [65] for sequential prompt conditioning, though facing coherence challenges in transitions. Contemporary solutions address these limitations through various approaches: Free-Bloom [22] employs LLM-generated coherent prompts with spatial-temporal attention; LLM-Grounded Video Diffusion [68] alternates between language guidance and denoising steps; VideoStudio [128] modifies cross-attention mechanisms; and DirecT2V [126] utilizes GPT-4 for step-by-step prompt generation. Figure 16 illustrates the DirecT2V architecture [126], which modifies the attention block of U-Net and incorporates modulated self-attention. Text-only prompts can guide long video generation, but they lack the semantics necessary for fine-grained control over this process. In addition to frame descriptions, adding metadata, such as bounding boxes for entities like persons and cars, as well as background information, can help generate a more accurate depiction of videos and facilitate fine alignment between text and video.

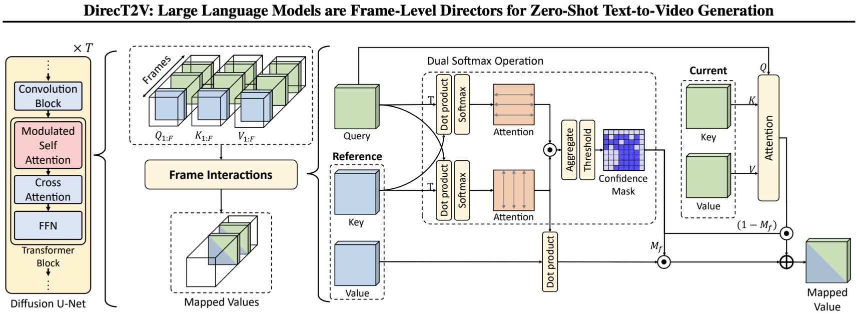


Fig. 16. DirectT2V Modulated self-attention for capturing interactions between frames [126].

5.2 User Textual Prompt with Scenes Layout

Multi-modal input control mechanisms extend beyond text prompts to include images, edge maps, bounding boxes, and music, thereby enhancing generation with metadata such as entity trajectories and layouts (Figure 8). Stable Diffusion [46] pioneered this via cross-attention in UNet backbones (Figure 15), while ControlNet [129] introduced a trainable copy linked via zero convolutions to preserve pre-trained features. This architecture enables diverse applications [130–133]. Further advances include Layout2Vid in [11] (extending ModelScopeT2V [97] with layout/entity metadata), LLM-Grounded VDM's training-free dynamic scene layouts [68], and FlowZero's spatiotemporal scene indices [23]. Dynamic scene layout increases control and detail alignment between text and video, but does not provide aesthetic infusion. Images can guide aesthetics if generated by a high-quality text-to-image diffusion model or provided as a reference. We will discuss a few articles incorporating images as a guidance mechanism.

5.3 Image Input with Textual Prompt and Scenes Layout

Images enhance video generation by providing high-quality spatial and aesthetic details (e.g., object textures and entity positions) beyond text descriptions. NUWA-Infinity [79] supports both text and image inputs, while VideoStudio [134] leverages reference images (e.g., actors and objects) alongside text prompts to guide generation. Microcinema [127] employs a multi-stage pipeline, first generating images via SDXL [135] or DALL-E [50], then using them for video synthesis. Similarly, Video-Booth [136] projects reference images into CLIP [27] text space, and VideoDrafter [128] uses a two-stage control pipeline (text-to-image, then image+text-to-video). Key works are compared in Table 4.

User textual prompts, additional metadata for scene layouts, entity descriptions, and reference images provide a rich context for the video generation model, facilitating fine alignment between user intention and generated videos. We also require extensive video datasets with labels or captions to train the long video generation and input control mechanisms, which will be discussed in the next section.

6 Existing Datasets

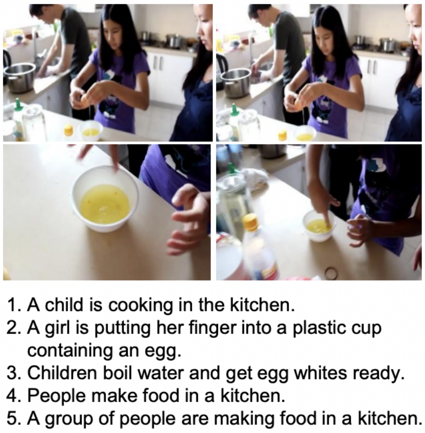
The existing datasets for long video generation can be categorized as classification datasets Section 6.1 and captions datasets Section 6.2.

6.1 Classification Datasets

Video classification datasets have evolved significantly in scale and annotation granularity. Early datasets like UCF-101 [138] (13,320 clips, 101 action classes) were limited to single-label

Table 4. Input Conditioning: Milestone Articles. 'Text Prompt' Represents Section 5.1, 'Text, Scene Layout' Represents Section 5.2, and 'Image, Text, Layout' Represents Section 5.3

Model	Theme	input control	Year
Free-Bloom [22]	LLM prompts with cross-attention and step-blocks.	Text Prompt	2023
MEVG [103]	Initial latent code, cross-attention.	Text Prompt	2023
SEINE [26]	Transition frames using CLIP and Lavie [137].	Text Prompt	2023
GLOBER [109]	CLIP encoder, cross-modal instructions with attention.	Text Prompt	2023
FlowZero [23]	Frame sequences using LLM with cross-frame attention.	Text, Scene Layout	2023
VideoDirectorGPT [11]	LLM generates video plan, interpreted by U-Net.	Text, Scene Layout	2024
LLM grounded VDM [68]	LLM story with attention maps and bounding layouts.	Text, Scene Layout	2024
VideoTetris [77]	ControlNet for autoregressive video generation.	Text, Scene Layout	2024
VideoDrafter [128]	LLM scripts scenes with CLIP embeddings and prompts.	Image, Text, Layout	2024
MAVIN [25]	3D UNET with cross-attention and CLIP embeddings.	Image, Text, Layout	2024
VideoBooth [136]	Video from images and prompts using latent space.	Image, Text, Layout	2024
MicroCinema [127]	LLM scripts multi-stage 3D Unet with cross-attention.	Image, Text, Layout	2024
Sora [21]	LLM prompts with optional visual input encoding.	Image, Text, Layout	2024



1. A child is cooking in the kitchen.
2. A girl is putting her finger into a plastic cup containing an egg.
3. Children boil water and get egg whites ready.
4. People make food in a kitchen.
5. A group of people are making food in a kitchen.

Fig. 17. Here are examples from the MSR-VTT dataset showcasing video clips paired with labeled sentences. Each example includes four frames representing the video clip and five human-generated sentences that describe the content [142].

categorization. The Kinetics series [139] expanded this to 306,245 videos across 700 classes while maintaining single-label classification. YouTube-8M [140] introduced multi-label annotation at scale (8M videos, 350k+ hours). HowTo100M [141] further advanced this with 136M clips featuring 23k tasks and narrative descriptions. The shift toward richer annotations is exemplified by captioning datasets like MSR-VTT [142], which pairs video frames with descriptive sentences. This progression reflects a broader trend from constrained single-label datasets to large-scale, multi-modal video-text collections.

6.2 Captions Datasets

MSR-VTT pioneered natural language video descriptions with 41.2 hours of videos and 200K clip-sentence pairs (Figure 17). Later datasets dramatically scaled up volume through automated methods: WebVid-2M [143] compiled 2M videos with algorithmic captions (similar to Conceptual Captions [144]). In contrast, InternVid [145] expanded to 234M clips with 4.1B words. However,

Table 5. Datasets for Long Videos

Dataset	**Size**	**Month/Year**	**Avg duration**	**Category**
UCF-101 [138]	13,320	Dec 2012	7.21 sec	classification
Kinetics-400 [139]	306,245	May 2017	10 sec	classification
Kinetics-600 [149]	480,000	Aug 2018	10 sec	classification
HowTo100M [141]	136 million	June 2019	6.5 min	classification
YouTube 8M [140]	6.1 million	Sept 2016	230 sec	classification
YouTube 8M Segments [140]	237k	June 2019	25 sec	classification
WebVid-2M [143]	2.5 million	April 2021	18 sec	captions
Pandas 70m [150]	70.8 million	Feb 2024	8.5 sec	captions
HD-VG-130M [151]	130 million	May 2023	10 sec	captions
InternVid [145]	234 million	July 2023	39 sec	captions
VidProM [152]	6.69 million	Sept 2024	2.5 sec	captions
Ego4D [153]	3,670 (hours)	Oct 2021	180–300 sec	captions
E-SyncVidStory [154]	6k	May 2024	39(s)	captions
LGVQ [155]	2,808	July 2024	8–96 sec	captions
VideoInstruct-100K [146]	100k	June 2024	2–3 min	captions
MiraData [148]	788k	July 2024	72.1(s)	captions
Vimeo25M [137]	25M	Sept 2023	19.6(s)	captions

these lack detailed spatiotemporal context due to limitations in automated captioning. Recent datasets address this quality gap: VideoInstruct-100K [146] enriched ActivityNet [147] subsets with human-annotated spatial/temporal details, and MiraData [148] employed GPT4-V to generate structured "dense captions" covering subjects, motion, and scene attributes. Examples from the MSR-VTT dataset show video clips paired with descriptive sentences, each with four frames and five human-labeled captions [142]. Using such datasets presents the challenge of measuring generated video quality, including frame fidelity, transition smoothness, and text-video alignment, as discussed in Section 7.3. Table 5 catalogs milestone datasets for long video generation.

Long Video generation needs extensive metrics to measure aesthetic, motion, and semantic alignment between text prompts and video. We will discuss these metrics in the next section.

7 Performance Measures

Video generation metrics can be mainly categorized into four categories, including *Image Quality Metrics* Section 7.1, *Video Quality Metrics* Section 7.2, *Semantics Quality Metrics* Section 7.3, and *Composite Metrics* Section 7.4.

7.1 Image Quality Metrics

Image quality metrics are critical for evaluating generative models, with the **Inception Score** (**IS**) [156] being a widely adopted measure. IS uses a pre-trained Inception model [157] to assess quality (classification accuracy) and diversity (variety of classes), but fails to capture perceptual quality or generalize beyond ImageNet domains. To address these limitations, **Fréchet Inception Distance** (**FID**) [158] was introduced, which compares feature distributions between generated and authentic images for a more robust evaluation of perceptual and statistical fidelity. While these metrics excel for individual frames, they cannot assess temporal dynamics, such as motion transitions, a gap addressed by video-specific metrics.

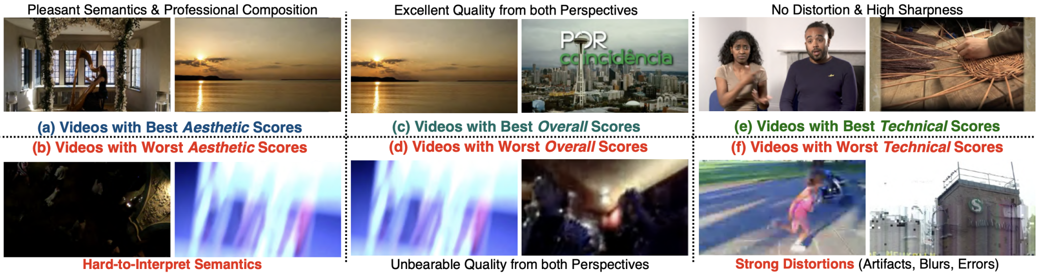

Pleasant Semantics & Professional Composition
Excellent Quality from both Perspectives
No Distortion & High Sharpness
POR coincidência
(a) Videos with Best *Aesthetic* Scores
(c) Videos with Best *Overall* Scores
(e) Videos with Best *Technical* Scores
(b) Videos with Worst *Aesthetic* Scores
(d) Videos with Worst *Overall* Scores
(f) Videos with Worst *Technical* Scores
Hard-to-Interpret Semantics
Unbearable Quality from both Perspectives
Strong Distortions (Artifacts, Blurs, Errors)

Fig. 18. Dover score. Samples from a dataset with human labeling of aesthetics and technical aspects of images. Dover score could be computed by aggregating or averaging the human-assigned aesthetic and technical scores [159].

7.2 Video Quality Metrics

Video quality assessment requires metrics that evaluate both spatial and temporal coherence, extending beyond traditional image metrics such as FID. The **Fréchet Video Distance** (**FVD**) [83] extends FID to videos by using 3D convolutions to capture temporal dynamics (motion, transitions), though it remains computationally intensive and doesn't assess aesthetic/technical flaws. To address this, Dover [159] evaluates technical quality (blur, noise, flicker) and aesthetics, leveraging its DIVIDE-3k dataset with 450K+ subjective annotations (Figure 18). For motion analysis, RAFT [160] measures optical flow to quantify object movement and temporal alignment between frames. Together, these metrics provide complementary insights into visual fidelity, temporal coherence, and motion quality.

While video quality metrics like FVD and Dover work well for assessing the technical quality of generated videos, they have limitations in measuring how well a generated video aligns with the user's intentions or the semantic content outlined in the prompts. To address this gap, we must explore semantic alignment metrics and composite metrics, which combine technical quality and alignment with user-defined content. The metrics presented in Table 6 evaluate the accuracy with which the generated videos align with the intentions described in their input prompts. Additionally, Table 7 highlights the evolution of video quality and semantic alignment metrics over time, showcasing advancements in generative models.

7.3 Semantics Alignment Metrics

Semantic quality metrics assess how well-generated videos align with user intent, particularly in response to text prompts. CLIP [27] is a foundational tool for image-text alignment, using contrastive learning to embed both modalities into a shared space and measure their semantic similarity. Its extension, CLIPScore [161], leverages CLIP's embeddings to provide a reference-free metric for assessing image-text correspondence without ground-truth labels, making it efficient for evaluating generation models.

CLIPSIM [162] extends this paradigm for videos. CLIPSIM computes the similarity between the text and each video frame and then averages these scores to measure semantic matching. Table 7 shows how models have been improved on these frame-level metrics and their derivatives over time.

While CLIP and CLIPScore effectively measure basic image-text alignment through embeddings, they struggle with complex object interactions or nuanced descriptions (e.g., attributes and actions). GRiT [163] addresses this by leveraging region-to-text understanding, enabling finer-grained interpretation of scenes (e.g., "a brown dog running") and better alignment with user intent.

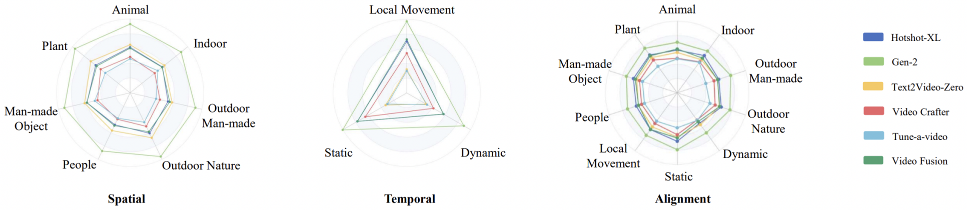


Fig. 19. The prompt dataset is designed to evaluate the model by focusing on three key quality aspects: (1) spatial quality (frame appearance), (2) temporal quality (frame coherence), and (3) text-to-video alignment (content-text correspondence) [155].

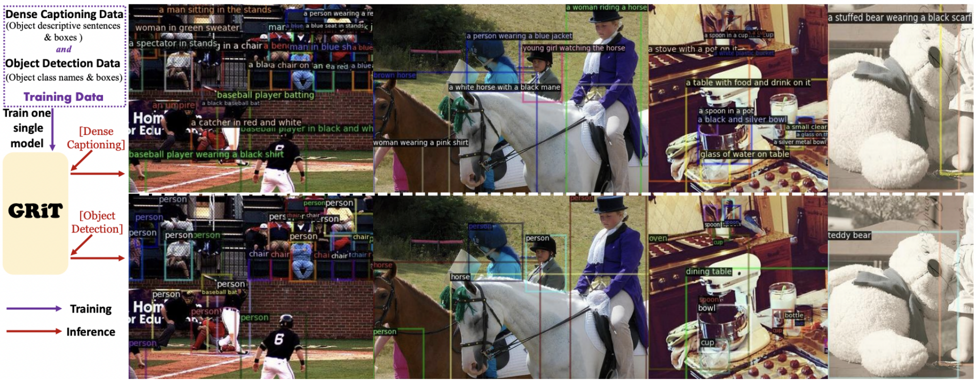


Fig. 20. GRiT locates different entities in scenes with their relations and matches with dense captions [163].

Table 6. Metrics for Video Quality Evaluation. In the "Direction" Column, ↑ Represents Higher Score Is Better, while ↓ Represents That a Lower Score Is Better.

Metrics	Type	Year	Direction
IS [156]	frame	June 2016	↑
FID [158]	frame	Jan 2018	↓
FVD [83]	video	March 2018	↓
CLIPScore [161]	image	March 2021	↑
FETV [164]	video	Nov 2023	↑
VBench [165]	video	Nov 2023	↑
T2VQA [166]	frame	March 2024	↑
FVD Motion [167]	video	June 2024	↓
UGVQ [155]	video	July 2024	↑
T2V-CompBench [168]	video	July 2024	↑
MiraBench [148]	video	July 2024	↑
Cross-Scene Face/Style Consistency Score [95]	video	August 2024	↑

As illustrated in Figure 20, GRiT employs a transformer-based architecture to learn the relationships between different image regions and their corresponding textual descriptions. It enables the model to break down the image into distinct regions and understand how each part corresponds to specific components of the prompt. In conclusion, while CLIP and CLIPScore provide

Table 7. Comprehensive Benchmark Comparison of Video Generation Models Ordered by Publication Year (2017–2024) [11, 22, 23, 95, 96, 100, 103, 108–111, 136, 137, 169–175]

Model (Citation)	Year	Zero-Shot	Samples	FVD	FID	CLIPSIM	IS
2021 Models							
GODIVA [162]	2021	No	30	–	–	0.2402	–
NUWA [176]	2021	X	0.97M	–	28.46	–	–
VideoGPT [48]	2021	–	–	103.3	24.69	–	24.69
Video Transformer [63]	2021	–	–	94 ± 2	–	–	–
TGAN-F [177]	2021	–	–	–	7817	–	22.91
LVT [178]	2021	–	–	125.8	–	–	–
VGAN [179]	2021	–	–	–	–	–	8.31
2022 Models							
Make-A-Video [170]	2022	Yes	1	367.23	13.17	0.3049	33.00
CogVideo (Chinese) [14]	2022	Yes	1	701.59	23.59	0.2614	23.55
CogVideo (English) [14]	2022	Yes	1	701.59	–	0.2631	25.27
VideoFusion [180]	2022	–	–	639.90	–	–	–
LVDM [171]	2022	–	–	372.00	–	–	–
Video Diffusion [181]	2022	–	–	–	295	–	57
Phenaki [64]	2022	Yes	15M	–	37.74	–	–
MagicVideo [182]	2022	–	–	655.00	–	–	–
2023 Models							
PYoCo [183]	2023	–	–	355.19	–	**0.3204**	**47.76**
VideoPoet (Pretrain) [169]	2023	–	–	355	–	0.3049	38.44
VideoFactory [184]	2023	–	–	410.00	–	0.3005	–
ModelScope [185]	2023	–	–	410.00	12.32	0.2930	–
LaViE [186]	2023	–	–	526.30	–	0.2949	–
Video LDM [187]	2023	–	–	550.61	–	0.2929	33.45
Vlogger [172]	2023	Yes	10M	292.43	37.23	–	–
ModelScopeT2V [185]	2023	–	–	–	12.32	0.2909	–
VideoDirectorGPT [174]	2023	–	–	–	12.22	0.2860	–
InternVid [188]	2023	–	–	617	–	0.2951	21.04
MicroCinema [173]	2023	–	–	342.86	–	–	–
PixelDance [189]	2023	–	–	242.82	–	–	42.10
Emu-Video [190]	2023	–	–	317.10	–	–	42.70
2024 Models							
Lumiere [191]	2024	–	–	332.49	–	–	–
Reducio-DiT [111]	2024	–	–	318.50	–	–	–
Goku-2B (256×256)[110]	2024	Yes	–	246.17	–	–	45.77 ± 1.10

FVD: UCF-101 metrics (from Video LDM comparison table).
BAIR FVD scores shown in FVD column.
† ModelScope replication results.

FID/CLIPSIM: MSR-VTT metrics.
X: Not applicable, Yes: Zero-shot capable.
Bold: Best results in each category.

practical methods for measuring the similarity between images and text, GRiT offers a significant advancement by enabling a deeper semantic understanding of the content within images. By considering not only individual objects but also their relationships, attributes, and actions, GRiT enhances the ability to evaluate generated content on a much more complex and nuanced level. These advances in semantic quality metrics are crucial for enhancing the alignment of videos generated with user intentions, ensuring that the videos are both visually accurate and semantically meaningful.

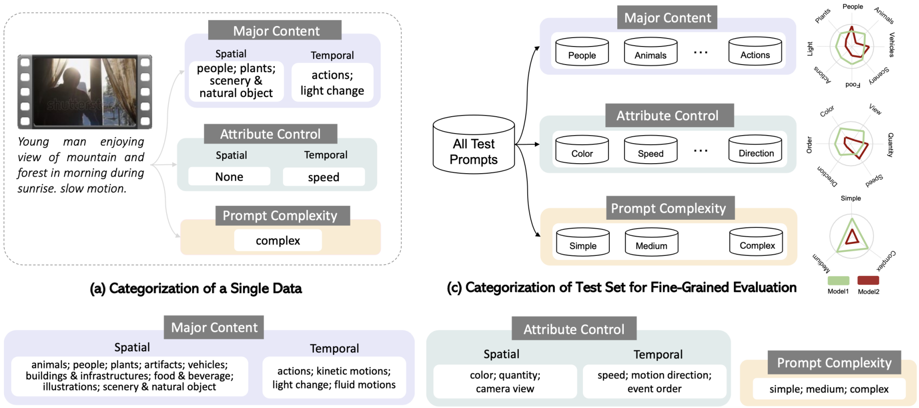

Young man enjoying view of mountain and forest in morning during sunrise. slow motion.
Major Content
Spatial
people; plants; scenery & natural object
Temporal
actions; light change
Attribute Control
Spatial
None
Temporal
speed
Prompt Complexity
complex
(a) Categorization of a Single Data
All Test Prompts
Major Content
People
Animals
...
Actions
Attribute Control
Color
Speed
...
Direction
Prompt Complexity
Simple
Medium
Complex
People
Animals
Vehicles
Scenery
Food
Actions
Light
Plants
Color
View
Quantity
Speed
Direction
Order
Simple
Medium
Complex
Model1
Model2
(c) Categorization of Test Set for Fine-Grained Evaluation
Major Content
Spatial
animals; people; plants; artifacts; vehicles; buildings & infrastructures; food & beverage; illustrations; scenery & natural object
Temporal
actions; kinetic motions; light change; fluid motions
Attribute Control
Spatial
color; quantity; camera view
Temporal
speed; motion direction; event order
Prompt Complexity
simple; medium; complex

Fig. 21. FETV is multi-faceted, classifying prompts into three distinct aspects: the main content, controllable attributes, and prompt complexity [164].

Table 8. Comparative Analysis of Video Generation Models Across 12 VBench Dimensions [164, 165]

Models	Subj. Cons.	Bkg. Cons.	Temp. Flicker	Motion	Aesthetic	Obj. Class
LaVie [193]	91.41	97.47	98.30	96.38	54.94	91.82
ModelScope [194]	89.87	95.29	98.28	95.79	52.06	82.25
CogVideo [14]	92.19	96.20	97.64	96.47	38.18	73.40
VideoCrafter [195]	96.85	98.22	98.41	97.73	63.13	92.55
Gen-2 [196]	97.61	97.61	99.56	99.58	66.96	90.92
AnimateDiff [197]	95.30	97.68	98.75	97.76	67.16	90.90
Latte-1 [198]	88.88	95.40	98.89	94.63	61.59	86.53
Pika-1.0 [199]	96.94	97.36	99.74	99.50	62.04	88.72
Kling [200]	98.33	97.60	99.30	99.40	61.21	87.24
Gen-3 [196]	97.10	96.62	98.61	99.23	63.34	87.81
CogVideoX [201]	96.23	96.52	98.66	96.92	61.98	85.23
Models	**Mult. Obj.**	**Human Act.**	**Color**	**Spatial**	**Temp. Style**	**Overall**
LaVie [193]	33.32	96.80	86.39	34.09	25.93	26.41
ModelScope [194]	38.98	92.40	81.72	33.68	25.37	25.67
CogVideo [14]	18.11	78.20	79.57	18.24	7.80	7.70
VideoCrafter [195]	40.66	95.00	92.92	35.86	25.84	28.23
Gen-2 [193]	55.47	89.20	89.49	66.91	24.12	26.17
AnimateDiff [197]	36.88	92.60	87.47	34.60	26.03	27.04
Latte-1 [198]	34.53	90.00	85.31	41.53	24.76	27.33
Pika-1.0 [199]	43.08	86.20	90.57	61.03	24.22	25.94
Kling [200]	68.05	93.40	89.90	73.03	24.17	26.42
Gen-3 [196]	53.64	96.40	80.90	65.09	24.71	26.69
CogVideoX [201]	62.11	99.40	82.81	66.35	25.38	27.59

These semantic alignment metrics have limitations in the video domain because the video may contain hundreds of frames, and they must match caption boundaries with corresponding frames. Composite metrics, aggregates of many individual algorithms, and manual scoring address this limitation.

7.4 Composite Metrics

Composite metrics combine multiple evaluation approaches to assess text-to-video generation across diverse categories (animals, objects, people) and dimensions (spatial, temporal, motion alignment). FETV Bench [164] pioneered this approach with its three-aspect prompt system (content, attributes, complexity Figure 21) and hybrid evaluation using both manual labeling and automated metrics like CLIPScore [161] and BLIPScore [192]. UGVQ [155] expanded this framework through its LGVQ dataset, evaluating spatial quality, motion coherence, and text-video alignment with partitioned prompts (foreground/background/motion) and manual scoring of six models (Figure 19). Subsequent benchmarks, such as T2V-CompBench, VBench/VBench++, and MiraBench, have further developed comprehensive evaluation protocols, although they reveal persistent gaps in the reliability of automated metrics—particularly for assessing long video temporal consistency and semantic fidelity (Table 8).

8 Conclusion and Future Trends

This survey equips users with a broad overview of the history, recent progress, and ongoing challenges in long video generation, focusing on video generation strategies, datasets, metrics, and open research areas. Long video generation is one of the actual north goals of generative AI, aiming to produce coherent and realistic videos over extended durations. Some of the challenges to be addressed by long video generation are maintaining temporal coherence and visual consistency while ensuring that the generated video aligns with a narrative or specific user intentions. Several strategies have been explored to tackle this challenge, including divide-and-conquer autoregressive models and intrinsic methods. Despite progress in these areas, motion consistency, semantic alignment, and parallel processing remain key obstacles to achieving scalable, high-quality long video generation. Future research in long video generation can focus on enhanced autoregressive models, novel frame and video segment merging techniques, and enhanced training paradigms.

One of the significant open research areas in long video generation is the generation of longer videos that accurately reflect spatial, temporal, and physical dynamics. A key challenge is the need for large-scale video datasets with comprehensive spatial, temporal, and physical context (e.g., trajectories, shadows, and interactions). Existing large-scale datasets, such as HD-VG-130M [151], offer scale but have limitations in terms of caption quality (e.g., captions are restricted to 15–20 words and lack rich spatial and temporal contextual information). On the other hand, datasets such as VideoInstruct-100K [146] provide rich spatial and temporal context but fall short in scale. Developing datasets that balance both scale and rich context is critical for advancing long video generation research. In addition to datasets, measuring the quality of generated videos presents another challenge. Current state-of-the-art metrics, such as FETV [164], MiraBench [148], and VBench [165], rely on manual human feedback to assess video quality, which is time-consuming, subjective, and challenging to scale. Future research should focus on developing fully automated metrics that can objectively evaluate the quality of generated videos in a more scalable manner.

Another open area of research in long video generation is the integration of audio. Currently, most commercial video generation models, such as SORA and Stability AI, do not produce accompanying audio. Developing methods to generate audio that aligns seamlessly with visual content is crucial for creating immersive and comprehensive videos, making this a key focus in the field of long-form video generation.

Long video generation promises to revolutionize multiple fields, including entertainment, education, virtual reality, and game development. However, it also introduces significant challenges, such as the potential for fake video creation, bias, violence, and moral concerns. Additionally, issues like hallucinations can limit the applicability of generative videos, particularly in domains like education and science. In conclusion, this survey provides readers with an in-depth overview of the current state-of-the-art in long video generation, highlighting key research areas and opportunities for future exploration. Lastly, please find below the link containing a collection of video generation projects and demo which the readers may find useful:

Long Video Generation Videos Home

References

[1] Sora Team. 2024. Video generation models as world simulators by Open A.I. (2024). Retrieved 1 July 2024 from https://openai.com/index/video-generation-models-as-world-simulators/

[2] Open A.I Team. 2022. Introducing ChatGPT by Open A.I. (2022). Retrieved 22 October 2024 from https://openai.com/index/chatgpt/

[3] Meta A.I Team. 2023. Meta LLama Models. (2023). Retrieved 21 September 2024 from https://www.llama.com/

[4] Google A.I Team. 2023. Google Gemini series. (2023). Retrieved 20 June 2024 from https://gemini.google.com/

[5] Claude A.I Team. 2023. Anthropic by Claude. (2023). Retrieved 8 June 2024 from https://www.anthropic.com/claude

[6] Mistral A.I Team. 2024. Mistral Large Model by Mistral. (2024). Retrieved 1 September 20 from https://mistral.ai/news/mistral-large/

[7] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with CLIP latents. *ArXiv* abs/2204.06125, (2022). Retrieved from https://arxiv.org/abs/2204.06125

[8] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*. Retrieved from https://openreview.net/forum?id=FPnUhsQJ5B

[9] Midjourney A.I Team. 2024. The Midjourney V5.2 model For Image generation. (2024). Retrieved 10 August 2024 from https://docs.midjourney.com/docs/model-version-5

[10] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. 2023. CogVideo: Large-scale pretraining for text-to-video generation via transformers. In *Proceedings of the 11th International Conference on Learning Representations*. Retrieved from https://openreview.net/forum?id=rB6TpjAuSRy

[11] Han Lin, Abhay Zala, Jaemin Cho, and Mohit Bansal. 2024. VideoDirectorGPT: Consistent Multi-Scene Video Generation via LLM-Guided Planning. (2024). Retrieved from https://openreview.net/forum?id=5PkgaUwiY0

[12] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. 2023. Make-a-video: Text-to-video generation without text-video data. In *Proceedings of the 11th International Conference on Learning Representations*. Retrieved from https://openreview.net/forum?id=nJfylDvgzlq

[13] Midjourney Team. 2023. Gen2 by Runway ML. (2023). Retrieved 20 June 2024 from https://runwayml.com/research/gen-2

[14] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. 2023. CogVideo: Large-scale pretraining for text-to-video generation via transformers. In *Proceedings of the 11th International Conference on Learning Representations*. Retrieved from https://openreview.net/forum?id=rB6TpjAuSRy

[15] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. 2023. Phenaki: Variable length video generation from open domain textual descriptions. In *Proceedings of the International Conference on Learning Representations*. Retrieved from https://openreview.net/forum?id=vOEXS39nOF

[16] Fu-Yun Wang, Wenshuo Chen, Guanglu Song, Han-Jia Ye, Yu Liu, and Hongsheng Li. 2023. Gen-L-Video: Multi-text to long video generation via temporal co-denoising. *CoRR* abs/2305.18264, (2023). Retrieved from https://doi.org/10.48550/arXiv.2305.18264

[17] William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. 4172–4182. DOI: http://dx.doi.org/10.1109/ICCV51070.2023.00387

[18] Midjourney Team. 2024. Gen-4 Alpha by MidJourney. (2024). Retrieved 23 March 2025 from https://runwayml.com/research/introducing-runway-gen-4

[19] Chengxuan Li, Di Huang, Zeyu Lu, Yang Xiao, Qingqi Pei, and Lei Bai. 2024. A survey on long video generation: Challenges, methods, and prospects. *arXiv preprint arXiv:2403.16407* (2024).

[20] Pengyuan Zhou, Lin Wang, Zhi Liu, Yanbin Hao, Pan Hui, Sasu Tarkoma, and Jussi Kangasharju. 2024. A survey on generative AI and LLM for video generation, understanding, and streaming. *arXiv preprint arXiv:2404.16038* (2024).

[21] Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, and others. 2024. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv:2402.17177* (2024).

[22] Hanzhuo Huang, Yufan Feng, Cheng Shi, Lan Xu, Jingyi Yu, and Sibei Yang. 2023. Free-bloom: Zero-shot text-to-video generator with LLM director and LDM animator. In *Proceedings of the 37th Conference on Neural Information Processing Systems*. Retrieved from https://openreview.net/forum?id=paa2OU5jN8

[23] Yu Lu, Linchao Zhu, Hehe Fan, and Yi Yang. 2023. Flowzero: Zero-shot text-to-video synthesis with LLM-driven dynamic scene syntax. *arXiv preprint arXiv:2311.15813* (2023).

[24] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. 2023. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 22563–22575. DOI: http://dx.doi.org/10.1109/CVPR52729.2023.02161

[25] Bowen Zhang, Xiaofei Xie, Haotian Lu, Na Ma, Tianlin Li, and Qing Guo. 2024. Mavin: Multi-action video generation with diffusion models via transition video infilling. *arXiv preprint arXiv:2405.18003* (2024).

[26] Xinyuan Chen, Yaohui Wang, Lingjun Zhang, Shaobin Zhuang, Xin Ma, Jiashuo Yu, Yali Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. 2024. SEINE: Short-to-long video diffusion model for generative transition and prediction. In *Proceedings of the 12th International Conference on Learning Representations.* Retrieved from https://openreview.net/forum?id=FNq3nIvP4F

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and others. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning.* PMLR, 8748–8763.

[28] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Communications of the ACM* 63, 11 (2020), 139–144. DOI: http://dx.doi.org/10.1145/3422622

[29] Alec Radford, Luke Metz, and Soumith Chintala. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434* (2015).

[30] Emily Denton, Soumith Chintala, Arthur Szlam, and Rob Fergus. 2015. Deep generative image models using a Laplacian pyramid of adversarial networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems—Volume 1 (NIPS'15).* MIT Press, Cambridge, MA, USA, 1486–1494.

[31] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris Metaxas. 2017. StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV).* 5908–5916. DOI: http://dx.doi.org/10.1109/ICCV.2017.629

[32] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. 2018. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 1316–1324. DOI: http://dx.doi.org/10.1109/CVPR.2018.00143

[33] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and improving the image quality of StyleGAN. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* IEEE Computer Society, Los Alamitos, CA, USA, 8107–8116. DOI: http://dx.doi.org/10.1109/CVPR42600.2020.00813

[34] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV).* 2242–2251. DOI: http://dx.doi.org/10.1109/ICCV.2017.244

[35] Yuri Viazovetskyi, Vladimir Ivashkin, and Evgeny Kashin. 2020. StyleGAN2 distillation for feed-forward image manipulation. In *Proceedings of the Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Part XXII.* Springer-Verlag, Berlin, 170–186. DOI: http://dx.doi.org/10.1007/978-3-030-58542-6_11

[36] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 5967–5976. DOI: http://dx.doi.org/10.1109/CVPR.2017.632

[37] Michael Mathieu, Camille Couprie, and Yann LeCun. 2015. Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv:1511.05440* (2015).

[38] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. 2016. Generating videos with scene dynamics. *Advances in Neural Information Processing Systems* 29 (2016).

[39] Wei Xiong, Wenhan Luo, Lin Ma, Wei Liu, and Jiebo Luo. 2018. Learning to generate time-lapse videos using multi-stage dynamic generative adversarial networks. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 2364–2373. DOI: http://dx.doi.org/10.1109/CVPR.2018.00251

[40] Yingwei Pan, Zhaofan Qiu, Ting Yao, Houqiang Li, and Tao Mei. 2017. To create what you tell: Generating videos from captions. In *Proceedings of the 25th ACM International conference on Multimedia.* 1789–1798.

[41] Sihyun Yu, Jihoon Tack, Sangwoo Mo, Hyunsu Kim, Junho Kim, Jung-Woo Ha, and Jinwoo Shin. 2022. Generating videos with dynamics-aware implicit generative adversarial networks. In *Proceedings of the International Conference on Learning Representations.* Retrieved from https://openreview.net/forum?id=Czsdv-S4-w9

[42] Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. 2022. StyleGAN-V: A continuous video generator with the price, image quality and perks of StyleGAN2. In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 3616–3626. DOI: http://dx.doi.org/10.1109/CVPR52688.2022.00361

[43] Standford tutorial. Auto-Encoders. Retrieved 10 July 2024 from http://ufldl.stanford.edu/tutorial/unsupervised/Autoencoders/

[44] Diederik P. Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[45] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 15979–15988. DOI: http://dx.doi.org/10.1109/CVPR52688.2022.01553

[46] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* IEEE Computer Society, Los Alamitos, CA, USA, 10674–10685. DOI: http://dx.doi.org/10.1109/CVPR52688.2022.01042

[47] Aaron Van Den Oord, Oriol Vinyals, and others. 2017. Neural discrete representation learning. *Advances in Neural Information Processing Systems* 30 (2017).

[48] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. 2021. VideoGPT: Video generation using VQ-VAE and transformers. *arXiv preprint arXiv:2104.10157* (2021).

[49] Patrick Esser, Robin Rombach, and Bjorn Ommer. 2021. Taming transformers for high-resolution image synthesis . In *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* IEEE Computer Society, Los Alamitos, CA, USA, 12868–12878. DOI: http://dx.doi.org/10.1109/CVPR46437.2021.01268

[50] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. CLIP: Connecting vision and language with contrastive learning. In *Proceedings of the 38th International Conference on Machine Learning.* PMLR, 8821–8831. Retrieved from https://proceedings.mlr.press/v139/ramesh21a.html

[51] Shir Gur, Sagie Benaim, and Lior Wolf. 2020. Hierarchical patch VAE-GAN: Generating diverse videos from a single sample. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.). Curran Associates, Inc., 16761–16772. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2020/file/c2f32522a84d5e6357e6abac087f1b0b-Paper.pdf

[52] Faraz Waseem, Rafael Perez Martinez, and Chris Wu. 2022. Visual anomaly detection in video by variational autoencoder. *arXiv preprint arXiv:2203.03872* (2022).

[53] Gensheng Pei, Tao Chen, Xiruo Jiang, Huafeng Liu, Zeren Sun, and Yazhou Yao. 2024. VideoMAC: Video masked autoencoders meet convnets . In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* IEEE Computer Society, Los Alamitos, CA, USA, 22733–22743. DOI: http://dx.doi.org/10.1109/CVPR52733.2024.02145

[54] Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G. Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, and Lu Jiang. 2023. MAGVIT: Masked generative video transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 10459–10469.

[55] Sungwoong Kim, Daejin Jo, Donghoon Lee, and Jongmin Kim. 2023. MAGVLT: Masked generative vision-and-language transformer . In *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* IEEE Computer Society, Los Alamitos, CA, USA, 23338–23348. DOI: http://dx.doi.org/10.1109/CVPR52729.2023.02235

[56] Quan Hoang, Tu Dinh Nguyen, Trung Le, and Dinh Phung. 2017. Multi-generator generative adversarial nets. *arXiv preprint arXiv:1708.02556* (2017).

[57] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17).* Curran Associates Inc., Red Hook, NY, USA, 6000–6010.

[58] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations.* Retrieved from https://openreview.net/forum?id=YicbFdNTTy

[59] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *Proceedings of the 38th International Conference on Machine Learning.* Marina Meila and Tong Zhang (Eds.), Proceedings of Machine Learning Research, Vol. 139, PMLR, 8821–8831. Retrieved from https://proceedings.mlr.press/v139/ramesh21a.html

[60] Jason Tyler Rolfe. 2016. Discrete variational autoencoders. *arXiv preprint arXiv:1609.02200* (2016).

[61] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and Jie Tang. 2024. CogView: Mastering text-to-image generation via transformers. In *Proceedings of the 35th International Conference on Neural Information Processing Systems (NIPS '21).* Curran Associates Inc., Red Hook, NY, USA, Article 1516, 14 pages.

[62] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. 2024. CogView2: Faster and better text-to-image generation via hierarchical transformers. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS '22).* Curran Associates Inc., Red Hook, NY, USA, Article 1229, 13 pages.

[63] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lucic, and Cordelia Schmid. 2021. ViViT: A video vision transformer. In *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE Computer Society, Los Alamitos, CA, USA, 6816–6826. DOI: http://dx.doi.org/10.1109/ICCV48922.2021.00676

[64] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. 2023. Phenaki: Variable length video generation from open domain textual descriptions. In *Proceedings of the International Conference on Learning Representations*. Retrieved from https://openreview.net/forum?id=vOEXS39nOF

[65] Adam Roberts, Hyung Won Chung, Gaurav Mishra, Anselm Levskaya, James Bradbury, Daniel Andor, Sharan Narang, Brian Lester, Colin Gaffney, Afroz Mohiuddin, and others. 2023. Scaling up models and data with t5x and seqio. *Journal of Machine Learning Research* 24, 377 (2023), 1–8.

[66] Hanxin Zhu, Tianyu He, Anni Tang, Junliang Guo, Zhibo Chen, and Jiang Bian. 2024. Compositional 3D-aware video generation with LLM director. In *Proceedings of the 38th Annual Conference on Neural Information Processing Systems (NeurIPS 2024)*. Retrieved from https://nips.cc/virtual/2024/poster/93599 Poster presentation.

[67] 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. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS '20)*. Curran Associates Inc., Red Hook, NY, USA, Article 159, 25 pages.

[68] Long Lian, Baifeng Shi, Adam Yala, Trevor Darrell, and Boyi Li. 2024. LLM-grounded Video Diffusion Models. In *Proceedings of the 12th International Conference on Learning Representations*. Retrieved from https://openreview.net/forum?id=exKHibougU

[69] Shaobin Zhuang, Kunchang Li, Xinyuan Chen, Yaohui Wang, Ziwei Liu, Yu Qiao, and Yali Wang. 2024. Vlogger: Make your dream a vlog. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8806–8817. DOI: http://dx.doi.org/10.1109/CVPR52733.2024.00841

[70] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*. pmlr, 2256–2265.

[71] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS '20)*. Curran Associates Inc., Red Hook, NY, USA, Article 574, 12 pages.

[72] Yang Song and Stefano Ermon. 2019. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems* 32 (2019).

[73] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 4171–4186.

[74] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research* 21, 1, Article 140 (2020), 67 pages.

[75] Open A.I Team. Video generation models as world simulators. Retrieved from https://openai.com/index/video-generation-models-as-world-simulators/

[76] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and Jie Tang. 2021. CogView: Mastering text-to-image generation via transformers. In *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 19822–19835. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2021/file/a4d92e2cd541fca87e4620aba658316d-Paper.pdf

[77] Ye Tian, Ling Yang, Haotian Yang, Yuan Gao, Yufan Deng, Xintao Wang, Zhaochen Yu, Xin Tao, Pengfei Wan, Di ZHANG, et al. 2024. VideoTetris: Towards compositional text-to-video generation. In *Proceedings of the 38th Annual Conference on Neural Information Processing Systems*. Retrieved from https://openreview.net/forum?id=RPM7STrnVz

[78] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. 2022. CogView2: Faster and better text-to-image generation via hierarchical transformers. In *Proceedings of the Advances in Neural Information Processing Systems*. S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35, Curran Associates, Inc., 16890–16902. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2022/file/6baec7c4ba0a8734ccbd528a8090cb1f-Paper-Conference.pdf

[79] Jian Liang, Chenfei Wu, Xiaowei Hu, Zhe Gan, Jianfeng Wang, Lijuan Wang, Zicheng Liu, Yuejian Fang, and Nan Duan. 2022. NUWA-infinity: Autoregressive over autoregressive generation for infinite visual synthesis. In *Proceedings of the Advances in Neural Information Processing Systems*. Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (Eds.). Retrieved from https://openreview.net/forum?id=0Kv7cLhuhQT

[80] Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming-Chang Chiu, et al. 2024. VideoPoet: A large language model for zero-shot video generation. In *Proceedings of the 41st International Conference on Machine Learning (ICML'24)*. JMLR.org, Article 1005, 20 pages.

[81] Wenming Weng, Ruoyu Feng, Yanhui Wang, Qi Dai, Chunyu Wang, Dacheng Yin, Zhiyuan Zhao, Kai Qiu, Jianmin Bao, Yuhui Yuan, et al. 2024. ART·V: Auto-regressive text-to-video generation with diffusion models. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE Computer Society, Los Alamitos, CA, USA, 7395–7405. DOI: http://dx.doi.org/10.1109/CVPRW63382.2024.00735

[82] Taegyeong Lee, Soyeong Kwon, and Taehwan Kim. 2024. Grid diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8734–8743.

[83] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. 2019. FVD: A new metric for video generation. In *DGS@ICLR*. Retrieved from https://api.semanticscholar.org/CorpusID:198489709

[84] Zongyi Li, Shujie Hu, Shujie Liu, Long Zhou, Jeongsoo Choi, Lingwei Meng, Xun Guo, Jinyu Li, Hefei Ling, and Furu Wei. 2024. ARLON: Boosting diffusion transformers with autoregressive models for long video generation. *arXiv preprint arXiv:2410.20502* (2024).

[85] Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. 2022. StyleGAN-V: A continuous video generator with the price, image quality and perks of StyleGAN2. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 3626–3636.

[86] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. 2020. Implicit neural representations with periodic activation functions. *Advances in Neural Information Processing Systems* 33 (2020), 7462–7473.

[87] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2021. NeRF: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM* 65, 1 (2021), 99–106.

[88] Sihyun Yu, Kihyuk Sohn, Subin Kim, and Jinwoo Shin. 2023. Video probabilistic diffusion models in projected latent space. In *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 18456–18466. DOI: http://dx.doi.org/10.1109/CVPR52729.2023.01770

[89] Jaehoon Yoo, Semin Kim, Doyup Lee, Chiheon Kim, and Seunghoon Hong. 2023. Towards end-to-end generative modeling of long videos with memory-efficient bidirectional transformers. In *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 22888–22897. DOI: http://dx.doi.org/10.1109/CVPR52729.2023.02192

[90] Anonymous. 2024. StreamingT2V: Consistent, dynamic, and extendable long video generation from text. In *Proceedings of the Submitted to The 13th International Conference on Learning Representations*. Retrieved from https://openreview.net/forum?id=26oSbRRpEY under review.

[91] Kaifeng Gao, Jiaxin Shi, Hanwang Zhang, Chunping Wang, and Jun Xiao. 2024. ViD-GPT: Introducing GPT-style autoregressive generation in video diffusion models. *arXiv preprint arXiv:2406.10981* (2024).

[92] Yichen Ouyang, Hao Zhao, Gaoang Wang, and others. 2024. Flexifilm: Long video generation with flexible conditions. *arXiv preprint arXiv:2404.18620* (2024).

[93] Zhengqing Yuan, Yixin Liu, Yihan Cao, Weixiang Sun, Haolong Jia, Ruoxi Chen, Zhaoxu Li, Bin Lin, Li Yuan, Lifang He, and others. 2024. Mora: Enabling generalist video generation via a multi-agent framework. *arXiv preprint arXiv:2403.13248* (2024).

[94] Ram Selvaraj, Ayush Singh, Shafiudeen Kameel, Rahul Samal, and Pooja Agarwal. 2024. Vidgen: Long-form text-to-video generation with temporal, narrative and visual consistency for high quality story-visualisation tasks. In *Proceedings of the 2024 IEEE 9th International Conference for Convergence in Technology (I2CT)*. 1–8. DOI: http://dx.doi.org/10.1109/I2CT61223.2024.10544050

[95] Zhifei Xie, Daniel Tang, Dingwei Tan, Jacques Klein, Tegawend F. Bissyand, and Saad Ezzini. 2024. Dreamfactory: Pioneering multi-scene long video generation with a multi-agent framework. *arXiv preprint arXiv:2408.11788* (2024).

[96] Liu He, Yizhi Song, Hejun Huang, Pinxin Liu, Yunlong Tang, Daniel Aliaga, and Xin Zhou. 2024. Kubrick: Multimodal agent collaborations for synthetic video generation. *arXiv preprint arXiv:2408.10453* (2024).

[97] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. 2023. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571* (2023).

[98] Hanzhuo Huang, Yufan Feng, Cheng Shi, Lan Xu, Jingyi Yu, and Sibei Yang. 2024. Free-bloom: Zero-shot text-to-video generator with LLM director and LDM animator. In *Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS '23)*. Curran Associates Inc., Red Hook, NY, USA, Article 1138, 24 pages.

[99] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2021. Denoising diffusion implicit models. In *Proceedings of the International Conference on Learning Representations*. Retrieved from https://openreview.net/forum?id=St1giarCHLP

[100] Long Lian, Baifeng Shi, Adam Yala, Trevor Darrell, and Boyi Li. 2024. LLM-grounded video diffusion models. In *Proceedings of the 12th International Conference on Learning Representations*. Retrieved from https://openreview.net/forum?id=exKHibougU

[101] Zhengqing Yuan, Yixin Liu, Yihan Cao, Weixiang Sun, Haolong Jia, Ruoxi Chen, Zhaoxu Li, Bin Lin, Li Yuan, Lifang He, and others. 2024. Mora: Enabling generalist video generation via a multi-agent framework. *arXiv preprint arXiv:2403.13248* (2024).

[102] Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, and others. 2024. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv:2402.17177* (2024).

[103] Gyeongrok Oh, Jaehwan Jeong, Sieun Kim, Wonmin Byeon, Jinkyu Kim, Sungwoong Kim, and Sangpil Kim. 2024. MEVG: Multi-event Video Generation with Text-to-Video Models. In *Proceedings of the Computer Vision—ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Part XLIII*. Springer-Verlag, Berlin, 401–418. DOI: http://dx.doi.org/10.1007/978-3-031-72775-7_23

[104] Jingbo Yang and Adrian G. Bors. 2023. Enabling the encoder-empowered GAN-based video generators for long video generation. In *Proceedings of the 2023 IEEE International Conference on Image Processing (ICIP)*. 1425–1429. DOI: http://dx.doi.org/10.1109/ICIP49359.2023.10222725

[105] Siyang Zhang, Harry Yang, and Ser-Nam Lim. 2025. Videomerge: Towards training-free long video generation. *arXiv preprint arXiv:2503.09926* (2025).

[106] Gyeongrok Oh, Jaehwan Jeong, Sieun Kim, Wonmin Byeon, Jinkyu Kim, Sungwoong Kim, and Sangpil Kim. 2024. Mevg: Multi-event video generation with text-to-video models. In *European Conference on Computer Vision*. Springer, 401–418.

[107] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, and others. 2024. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603* (2024).

[108] Haonan Qiu, Menghan Xia, Yong Zhang, Yingqing He, Xintao Wang, Ying Shan, and Ziwei Liu. 2024. FreeNoise: Tuning-free longer video diffusion via noise rescheduling. In *Proceedings of the 12th International Conference on Learning Representations*. Retrieved from https://openreview.net/forum?id=ijoqFqSC7p

[109] Mingzhen Sun, Weining Wang, Zihan Qin, Jiahui Sun, Sihan Chen, and Jing Liu. 2023. GLOBER: Coherent non-autoregressive video generation via GLOBal guided video decodER. In *Proceedings of the 37th Conference on Neural Information Processing Systems*. Retrieved from https://openreview.net/forum?id=TRbklCR2ZW

[110] Shoufa Chen, Chongjian Ge, Yuqi Zhang, Yida Zhang, Fengda Zhu, Hao Yang, Hongxiang Hao, Hui Wu, Zhichao Lai, Yifei Hu, and others. 2025. Goku: Flow based video generative foundation models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 23516–23527.

[111] Rui Tian, Qi Dai, Jianmin Bao, Kai Qiu, Yifan Yang, Chong Luo, Zuxuan Wu, and Yu-Gang Jiang. 2024. REDUCIO! Generating 1K Video within 16 Seconds using Extremely Compressed Motion Latents. *arXiv preprint arXiv:2411.13552* (2024).

[112] Genmo Team. 2024. Mochi 1. Retrieved from https://github.com/genmoai/models. (2024).

[113] Meta Research Team. 2024. MovieGen: A cast of media foundation models. *arXiv preprint* (October 2024). Retrieved from https://ai.meta.com/research/publications/movie-gen-a-cast-of-media-foundation-models/

[114] Rui Sun, Yumin Zhang, Tejal Shah, Jiahao Sun, Shuoying Zhang, Wenqi Li, Haoran Duan, Bo Wei, and Rajiv Ranjan. 2024. From sora what we can see: A survey of text-to-video generation. *arXiv preprint arXiv:2405.10674* (2024).

[115] NVIDIA, Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, Daniel Dworakowski, Jiaojiao Fan, Michele Fenzi, Francesco Ferroni, Sanja Fidler, Dieter Fox, Songwei Ge, Yunhao Ge, Jinwei Gu, Siddharth Gururani, Ethan He, Jiahui Huang, Jacob Huffman, Pooya Jannaty, Jingyi Jin, Seung Wook Kim, Gergely Klár, Grace Lam, Shiyi Lan, Laura Leal-Taixé, Anqi Li, Zhaoshuo Li, Chen-Hsuan Lin, Tsung-Yi Lin, Huan Ling, Ming-Yu Liu, Xian Liu, Alice Luo, Qianli Ma, Hanzi Mao, Kaichun Mo, Arsalan Mousavian, Seungjun Nah, Sriharsha Niverty, David Page, Despoina Paschalidou, Zeeshan Patel, Lindsey Pavao, Morteza Ramezanali, Fitsum Reda, Xiaowei Ren, Vasanth Rao Naik Sabavat, Ed Schmerling, Stella Shi, Bartosz Stefaniak, Shitao Tang, Lyne Tchapmi, Przemek Tredak, Wei-Cheng Tseng, Jibin Varghese, Hao Wang, Haoxiang Wang, Heng Wang, Ting-Chun Wang, Fangyin Wei, Xinyue Wei, Jay Zhangjie Wu, Jiashu Xu, Wei Yang, Lin Yen-Chen, Xiaohui Zeng, Yu Zeng, Jing Zhang, Qinsheng Zhang, Yuxuan Zhang, Qingqing Zhao, and Artur Zolkowski. 2025. Cosmos world foundation model platform for physical AI. *arXiv preprint arXiv:2501.03575* (2025). Retrieved from https://arxiv.org/abs/2501.03575

[116] Chenfei Wu, Jian Liang, Lei Ji, Fan Yang, Yuejian Fang, Daxin Jiang, and Nan Duan. 2022. NÜWA: Visual synthesis pre-training for neural visual world creation. In *European Conference on Computer Vision*. Springer, 720–736.

[117] Younggyo Seo, Kimin Lee, Fangchen Liu, Stephen James, and Pieter Abbeel. 2022. HARP: Autoregressive latent video prediction with high-fidelity image generator. In *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, 3943–3947.

[118] Qihang Yu, Mark Weber, Xueqing Deng, Xiaohui Shen, Daniel Cremers, and Liang-Chieh Chen. 2024. An image is worth 32 tokens for reconstruction and generation. *Advances in Neural Information Processing Systems* 37 (2024), 128940–128966.

[119] Junke Wang, Yi Jiang, Zehuan Yuan, Bingyue Peng, Zuxuan Wu, and Yu-Gang Jiang. 2024. OmniTokenizer: A joint image-video tokenizer for visual generation. *Advances in Neural Information Processing Systems* 37 (2024), 28281–28295.

[120] David Dehaene and Rémy Brossard. 2021. Re-parameterizing VAEs for stability. *arXiv preprint arXiv:2106.13739* (2021).

[121] Fengyuan Shi, Zhuoyan Luo, Yixiao Ge, Yujiu Yang, Ying Shan, and Limin Wang. 2025. Scalable image tokenization with index backpropagation quantization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 16037–16046.

[122] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. 2020. HERO: Hierarchical encoder for video+ language omni-representation pre-training. *arXiv preprint arXiv:2005.00200* (2020).

[123] Ziqin Zhou, Yifan Yang, Yuqing Yang, Tianyu He, Houwen Peng, Kai Qiu, Qi Dai, Lili Qiu, Chong Luo, and Lingqiao Liu. 2025. HiTVideo: Hierarchical tokenizers for enhancing text-to-video generation with autoregressive large language models. *arXiv preprint arXiv:2503.11513* (2025).

[124] Dawit Mureja Argaw, Xian Liu, Joon Son Chung, Ming-Yu Liu, and Fitsum Reda. 2025. MambaVideo for discrete video tokenization with channel-split quantization. *arXiv preprint arXiv:2507.04559* (2025).

[125] Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, and others. 2023. MAGVIT: Masked generative video transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10459–10469.

[126] Susung Hong, Junyoung Seo, Heeseong Shin, Sunghwan Hong, and Seungryong Kim. 2024. Large language models are frame-level directors for zero-shot text-to-video generation. In *Proceedings of the 1st Workshop on Controllable Video Generation @ICML24*. Retrieved from https://openreview.net/forum?id=VmOO0GsG0K

[127] Yanhui Wang, Jianmin Bao, Wenming Weng, Ruoyu Feng, Dacheng Yin, Tao Yang, Jingxu Zhang, Qi Dai, Zhiyuan Zhao, Chunyu Wang, et al. 2024. MicroCinema: A divide-and-conquer approach for text-to-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8414–8424.

[128] Fuchen Long, Zhaofan Qiu, Ting Yao, and Tao Mei. 2024. VideoStudio: Generating consistent-content and multi-scene videos. In *European Conference on Computer Vision*. Springer, 468–485.

[129] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3836–3847.

[130] Zhihao Hu and Dong Xu. 2023. VideoControlNet: A motion-guided video-to-video translation framework by using diffusion model with controlnet. *arXiv preprint arXiv:2307.14073* (2023).

[131] David Junhao Zhang, Dongxu Li, Hung Le, Mike Zheng Shou, Caiming Xiong, and Doyen Sahoo. 2024. Moonshot: Towards controllable video generation and editing with multimodal conditions. *arXiv preprint arXiv:2401.01827* (2024).

[132] Cong Wang, Jiaxi Gu, Panwen Hu, Haoyu Zhao, Yuanfan Guo, Jianhua Han, Hang Xu, and Xiaodan Liang. 2024. EasyControl: Transfer controlnet to video diffusion for controllable generation and interpolation. *arXiv preprint arXiv:2408.13005* (2024).

[133] Tianjun Zhang, Yi Zhang, Vibhav Vineet, Neel Joshi, and Xin Wang. 2023. Controllable text-to-image generation with GPT-4. *arXiv preprint arXiv:2305.18583* (2023).

[134] Fuchen Long, Zhaofan Qiu, Ting Yao, and Tao Mei. 2024. Videostudio: Generating consistent-content and multi-scene videos. In *European Conference on Computer Vision*. Springer, 468–485.

[135] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. SDXL: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952* (2023).

[136] Yuming Jiang, Tianxing Wu, Shuai Yang, Chenyang Si, Dahua Lin, Yu Qiao, Chen Change Loy, and Ziwei Liu. 2024. VideoBooth: Diffusion-based video generation with image prompts. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 6689–6700. DOI: http://dx.doi.org/10.1109/CVPR52733.2024.00639

[137] Y. Wang, X. Chen, X. Ma, et al. 2025. LaVie: High-quality video generation with cascaded latent diffusion models. *Int J Comput Vis* 133 (2025), 3059–3078. https://doi.org/10.1007/s11263-024-02295-1

[138] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402* (2012).

[139] Joao Carreira and Andrew Zisserman. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[140] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. 2016. YouTube-8M: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675* (2016).

[141] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. HowTo100M: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2630–2640. DOI: http://dx.doi.org/10.1109/ICCV.2019.00272

[142] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. MSR-VTT: A large video description dataset for bridging video and language. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5288–5296. DOI: http://dx.doi.org/10.1109/CVPR.2016.571

[143] Max Bain, Arsha Nagrani, Gul Varol, and Andrew Zisserman. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE Computer Society, Los Alamitos, CA, USA, 1708–1718. DOI: http://dx.doi.org/10.1109/ICCV48922.2021.00175

[144] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the ACL*.

[145] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. 2024. InternVid: A large-scale video-text dataset for multimodal understanding and generation. In *Proceedings of the 12th International Conference on Learning Representations*. Retrieved from https://openreview.net/forum?id=MLBdiWu4Fw

[146] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2023. Video-ChatGPT: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424* (2023).

[147] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. ActivityNet: A large-scale video benchmark for human activity understanding. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 961–970. DOI: http://dx.doi.org/10.1109/CVPR.2015.7298698

[148] Xuan Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, Ailing Zeng, Yu Xiong, Qiang Xu, and Ying Shan. 2024. MiraData: A large-scale video dataset with long durations and structured captions. In *Proceedings of the 38th Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. Retrieved from https://openreview.net/forum?id=2myGfVgfva

[149] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. 2018. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340* (2018).

[150] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-Wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. 2024. Panda-70M: Captioning 70M Videos with Multiple Cross-Modality Teachers. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, Los Alamitos, CA, USA, 13320–13331. DOI: http://dx.doi.org/10.1109/CVPR52733.2024.01265

[151] Wenjing Wang, Huan Yang, Zixi Tuo, Huiguo He, Junchen Zhu, Jianlong Fu, and Jiaying Liu. 2023. VideoFactory: Swap attention in spatiotemporal diffusions for text-to-video generation. *arXiv preprint arXiv:2305.10874*.

[152] Wenhao Wang and Yi Yang. 2024. VidProM: A million-scale real prompt-gallery dataset for text-to-video diffusion models. In *Proceedings of the 2024 NeurIPS Conference*.

[153] Kristen Grauman, Andrew Westbury, Eugene Byrne, Vincent Cartillier, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Devansh Kukreja, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina Gonzalez, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jachym Kolar, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbelaez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. 2025. Ego4D: Around the world in 3,600 hours of egocentric video. *IEEE Transactions on Pattern Analysis & Machine Intelligence* 47, 11 (November 2025), 9468–9509. DOI: https://doi.org/10.1109/TPAMI.2024.3381075

[154] Dingyi Yang, Chunru Zhan, Ziheng Wang, Biao Wang, Tiezheng Ge, Bo Zheng, and Qin Jin. 2024. Synchronized video storytelling: Generating video narrations with structured storyline. *arXiv preprint arXiv:2405.14040* (2024).

[155] Zhichao Zhang, Wei Sun, Li Xinyue, Jun Jia, Xiongkuo Min, Zicheng Zhang, Chunyi Li, Zijian Chen, Wang Puyi, Sun Fengyu, and others. 2025. Benchmarking multi-dimensional AIGC video quality assessment: A dataset and unified model. *ACM Transactions on Multimedia Computing, Communications and Applications* 21, 9 (2025), 1–24.

[156] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training GANs. In *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16)*. Curran Associates Inc., Red Hook, NY, USA, 2234–2242.

[157] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2818–2826. DOI: http://dx.doi.org/10.1109/CVPR.2016.308

[158] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 6629–6640.

[159] Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. 2023. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE Computer Society, Los Alamitos, CA, USA, 20087–20097. DOI: http://dx.doi.org/10.1109/ICCV51070.2023.01843

[160] Zachary Teed and Jia Deng. 2020. RAFT: Recurrent all-pairs field transforms for optical flow. In *Proceedings of the 16th European Conference on Computer Vision—ECCV 2020, Glasgow, UK, August 23–28, 2020, Part II*. Springer-Verlag, Berlin, 402–419. DOI: http://dx.doi.org/10.1007/978-3-030-58536-5_24

[161] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718* (2021).

[162] Chenfei Wu, Lun Huang, Qianxi Zhang, Binyang Li, Lei Ji, Fan Yang, Guillermo Sapiro, and Nan Duan. 2021. GODIVA: Generating Open-DomaIn Videos from nAtural Descriptions. arXiv:2104.14806. Retrieved from https://arxiv.org/abs/2104.14806

[163] Jialian Wu, Jianfeng Wang, Zhengyuan Yang, Zhe Gan, Zicheng Liu, Junsong Yuan, and Lijuan Wang. 2024. GRiT: A generative region-to-text transformer for object understanding. In *European Conference on Computer Vision*. Springer, 207–224.

[164] Yuanxin Liu, Lei Li, Shuhuai Ren, Rundong Gao, Shicheng Li, Sishuo Chen, Xu Sun, and Lu Hou. 2024. FETV: A benchmark for fine-grained evaluation of open-domain text-to-video generation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS '23)*. Curran Associates Inc., Red Hook, NY, USA, Article 2723, 36 pages.

[165] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. 2024. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 21807–21818. DOI: http://dx.doi.org/10.1109/CVPR52733.2024.02060

[166] Tengchuan Kou, Xiaohong Liu, Zicheng Zhang, Chunyi Li, Haoning Wu, Xiongkuo Min, Guangtao Zhai, and Ning Liu. 2024. Subjective-aligned dataset and metric for text-to-video quality assessment. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 7793–7802.

[167] Jiahe Liu, Youran Qu, Qi Yan, Xiaohui Zeng, Lele Wang, and Renjie Liao. 2024. Fréchet video motion distance: A metric for evaluating motion consistency in videos. In *Proceedings of the 1st Workshop on Controllable Video Generation @ICML24*. Retrieved from https://openreview.net/forum?id=tTZ2eAhK9D

[168] Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. 2025. T2V-compbench: A comprehensive benchmark for compositional text-to-video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 8406–8416.

[169] Jiaxuan Guo, Yichun Li, Shangzhe Wang, Yinan Zhang, Xihui Liu, Yu Wang, Hanyang Yang, Jing Yang, and Ziwei Liu. 2023. VideoPoet: A large language model for zero-shot video generation. arXiv:2312.14125. Retrieved from https://arxiv.org/abs/2312.14125

[170] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Junjie An, Songyang Zhang, Qiyuan Hu, Oran Yang, Omri Ashual, Oran Gafni, and others. 2022. Make-A-Video: Text-to-video generation without Text-Video Data. *arXiv:2209.14792* (2022).

[171] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. 2022. Latent video diffusion models for high-fidelity long video generation. *arXiv preprint arXiv:2211.13221* (2022).

[172] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2023. VLogger: Generating Long Videos of Dynamic Human Activities. arXiv:2306.04308. Retrieved from https://arxiv.org/abs/2306.04308

[173] Yinan He, Yaohui Wang, Ceyuan Yang, Shangchen Zhou, Xiangyu Zhang, Xiaodong Yang, Yu Qiao, Dahua Lin, and Ying Shan. 2023. MicroCinema: A divide-and-conquer approach for text-to-video generation. arXiv:2312.04889. Retrieved from https://arxiv.org/abs/2312.04889

[174] Han Lin, Abhay Zala, Jaemin Cho, and Mohit Bansal. 2023. VideoDirectorGPT: Consistent multi-scene video generation via LLM-guided planning. *arXiv preprint arXiv:2309.15091* (2023).

[175] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. 2022. Video diffusion models. *Advances in Neural Information Processing Systems* 35 (2022), 8633–8646.

[176] Chenfei Wu, Jian Liang, Lei Ji, Fan Yang, Yuejian Fang, Daxin Jiang, and Nan Duan. 2021. NUWA: Visual synthesis pre-training for neural visual world creation. arXiv:2111.12417. Retrieved from https://arxiv.org/abs/2111.12417

[177] Masaki Saito, Eiichi Matsumoto, and Shunta Saito. 2017. Temporal generative adversarial nets with singular value clipping. In *Proceedings of the IEEE International Conference on Computer Vision*. 2830–2839.

[178] Qingqiu Huang, Wentao Yu, Yuanze Xu, Yitong Wang, and Dacheng Zhang. 2021. LVT: Language-vision transformer for multi-modal video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[179] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. 2016. Generating videos with scene dynamics. In *Proceedings of the NeurIPS*.

[180] Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liangsheng Wang, Yujun Shen, Deli Zhao, Jinren Zhou, and Tien-Ping Tan. 2023. VideoFusion: Decomposed diffusion models for high-quality video generation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'23)*. 10209–10218. Retrieved from https://api.semanticscholar.org/CorpusID:257532642

[181] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. 2022. Video diffusion models. *Advances in Neural Information Processing Systems* 35 (2022), 8633–8646.

[182] Emiel Hoogeboom, Jonathan Heek, and Tim Salimans. 2023. simple diffusion: End-to-end diffusion for high resolution images. In *International Conference on Machine Learning*. PMLR, 13213–13232.

[183] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. 2023. PYoCo: Latent diffusion priors for zero-shot video editing. arXiv:2303.04734. Retrieved from https://arxiv.org/abs/2303.04734

[184] Wenjing Wang, Huan Yang, Zixi Tuo, Huiguo He, Junchen Zhu, Jianlong Fu, and Jiaying Liu. 2025. Swap attention in spatiotemporal diffusions for text-to-video generation. *International Journal of Computer Vision* (2025). 1–19.

[185] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. 2023. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571* (2023).

[186] Julio J. Valdés and Alain B. Tchagang. 2023. Understanding the structure of qm7b and qm9 quantum mechanical datasets using unsupervised learning. *arXiv preprint arXiv:2309.15130* (2023).

[187] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. 2023. VideoLDM: High-resolution video generation with latent diffusion models. arXiv:2304.08818. Retrieved from https://arxiv.org/abs/2304.08818

[188] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziwei Huang, Yu Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Yinan Yang, et al. 2023. InternVid: A large-scale video-text dataset for multimodal understanding and generation. arXiv:2307.06942. Retrieved from https://arxiv.org/abs/2307.06942

[189] Yan Zeng, Guoqiang Wei, Jiani Zheng, Jiaxin Zou, Yang Wei, Yuchen Zhang, and Hang Li. 2023. Make Pixels Dance: High-Dynamic Video Generation. arXiv:2311.10982. Retrieved from https://arxiv.org/abs/2311.10982

[190] Rohit Girdhar, Mannat Singh, Andrew Brown, Quentin Duval, Samaneh Azadi, Sai Saketh Rambhatla, Akbar Shah, Xi Yin, Devi Parikh, and Ishan Misra. 2024. Emu Video: Factorizing Text-to-Video Generation by Explicit Image Conditioning. arXiv:2311.10709. Retrieved from https://arxiv.org/abs/2311.10709

[191] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. 2024. Lumiere: A Space-Time Diffusion Model for Video Generation. arXiv:2401.12945. Retrieved from https://arxiv.org/abs/2401.12945

[192] Yujie Lu, Xianjun Yang, Xiujun Li, Xin Eric Wang, and William Yang Wang. 2023. LLMScore: Unveiling the Power of Large Language Models in Text-to-Image Synthesis Evaluation. arXiv:2305.11116. Retrieved from https://arxiv.org/abs/2305.11116

[193] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. 2023. LAVIE: High-quality video generation with cascaded latent diffusion models. arXiv:2309.15103. Retrieved from https://arxiv.org/abs/2309.15103

[194] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. 2023. Modelscope text-to-video technical report. arXiv:2308.06571. Retrieved from https://arxiv.org/abs/2308.06571

[195] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. 2024. VideoCrafter2: Overcoming Data Limitations for High-Quality Video Diffusion Models. arXiv:2401.09047. Retrieved from https://arxiv.org/abs/2401.09047

[196] 2024. Gen-3. Retrieved June 17, 2024 from https://runwayml.com/research/introducing-gen-3-alpha

[197] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Y. Qiao, Dahua Lin, and Bo Dai. 2023. AnimateDiff: Animate your personalized text-to-image diffusion models without specific tuning. *ArXiv* abs/2307.04725, (2023). Retrieved from https://api.semanticscholar.org/CorpusID:259501509

[198] Xin Ma, Yaohui Wang, Gengyun Jia, Xinyuan Chen, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. 2024. Latte: Latent diffusion transformer for video generation. arXiv:2401.03048. Retrieved from https://arxiv.org/abs/2401.03048

[199] 2023. Pika Labs. Retrieved September 25, 2023 from https://www.pika.art/

[200] 2024. Kling. Retrieved June 6, 2024 from https://klingai.kuaishou.com/

[201] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. 2024. CogVideoX: Text-to-video diffusion models with an expert transformer. arXiv:2408.06072. Retrieved from https://arxiv.org/abs/2408.06072

Received 22 December 2024; revised 18 September 2025; accepted 29 September 2025