A Survey on Long-Video Storytelling Generation: Architectures, Consistency, and Cinematic Quality

Mohamed Elmoghany[1], **Ryan Rossi**[1], **Seunghyun Yoon**[1], **Subhojyoti Mukherjee**[1], **Eslam Bakr**[2], **Puneet Mathur**[1], **Gang Wu**[1], **Viet Dac Lai**[1], **Nedim Lipka**[1], **Ruiyi Zhang**[1], **Varun Manjunatha**[1], **Chien Nguyen**[1,3], **Daksh Dangi**[4], **Abel Salinas**[5], **Hongjie Chen**[6], **Xiaolei Huang**[7], **Joe Barrow**[11], **Nesreen Ahmed**[8], **Hoda Eldardiry**[9], **Namyong Park**[12], **Yu Wang**[3], **Zhengzhong Tu**[10], **Thien Nguyen**[1], **Dinesh Manocha**[13], **Mohamed Elhoseiny**[2], **Franck Dernoncourt**[1]

[1]Adobe Research [2]KAUST [3]University of Oregon [4]Independent Researcher [5]University of Southern California [6]Dolby Labs [7]University of Memphis [8]Cisco [9]Virginia Tech [10]Texas A&M University [11]Pattern Data [12]Meta AI [13]University of Maryland, College Park

Abstract

Despite the recent progress in video generative models, existing state-of-the-art methods can only produce videos lasting 5-16 seconds, often labeled "long-form videos". Furthermore, videos exceeding 16 seconds struggle to maintain consistent character appearances and scene layouts throughout the narrative. In particular, multi-subject long videos still fail to preserve character consistency and motion coherence. While some methods can generate videos up to 150 seconds long, they often suffer from frame redundancy and low temporal diversity. Recent work has attempted to produce long-form videos featuring multiple characters, narrative coherence, and high-fidelity detail. We studied 32 papers on long-video generation to identify key architectural components and training strategies that consistently yield these qualities. We also construct a comprehensive novel taxonomy of existing methods and present comparative tables that categorize papers by their architectural designs and performance characteristics.

1. Introduction

The advent of diffusion-based models DDPM [27] and DDIM [65], together with recent advances in large language models [5], has laid the foundation for cinematic long-form video generation and AI content creation. High-quality, realistic video generation now underpins applications in education [73], marketing [101], autonomous driving [85], gaming [51], entertainment [41], robotic learning [36], medicine [90], and virtual reality [94]. These modeling developments produce large volumes of synthetic video data that benefit all of the aforementioned domains. Such data can support educational content, drive task-specific applications, or facilitate training of machine learning models.

Video generation poses greater challenges than text or image generation. It incorporates temporal complexity alongside spatial complexity, ensuring frames' consistency, which is a key distinction from image generation. Consequently, beyond 16 seconds, autoregressive methods accumulate quality degradation, leading to inconsistencies, visual artifacts, or other perceptual errors [89]. Another limitation is the high memory requirements and computational resources required for video generation, as in the spatiotemporal attention [3, 16, 79].

Video datasets that can be used commercially are very limited, which further hampers progress. Most public datasets require commercial licenses e.g., MovieBench [86], Koala-36M [74], CelebV-HQ [106], Panda-70M [8], HD-VG-130M [77] and MiraData [39], impeding industry-driven innovation. Long-form video generation demands realistic clips with detailed annotations to capture full narrative scenes. However, existing open-source datasets typically contain only a few seconds of footage. Moreover, crucial metadata such as shot type, camera motion, character emotions, background context, and action labels are rarely provided, necessitating custom dataset creation and curation. Richly annotated datasets, as in MovieBench [86], include details about the background, characters, camera shot type, and camera motion. All of these annotation details

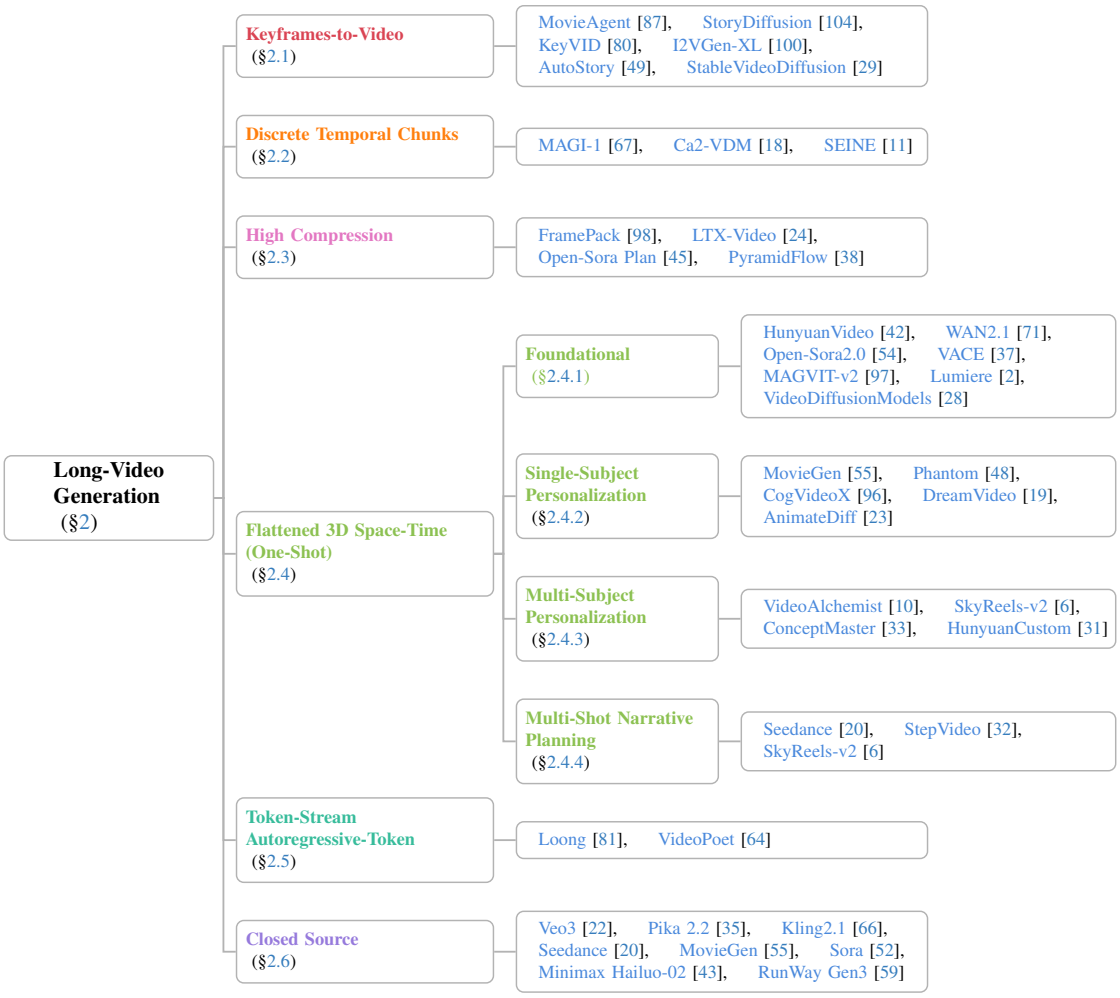

Figure 1. Architectural taxonomy of long-video generation methods. These trees were selected because together they span the key axes of temporal decomposition, compression, personalization scope, narrative structure, and openness that govern modern long-video synthesis. **(a) Keyframes-to-Video:** two-stage generating images as keyframes followed by motion pipelines that scale to minute-long clips **(b) Discrete Temporal Chunks:** constant-memory, parallel decoding of N-frame blocks **(c) High Compression:** heavy latent down-sampling for real-time, low-power inference **(d) Flattened 3D One-Shot:** end-to-end full-tensor synthesis: 1. Foundational: joint spatiotemporal prior for fixed-length clips; 2. Single-Subject: adapters for identity consistency and faithful individual likeness; 3. Multi-Subject: dedicated per-entity fusion modules for coherent multi-character scenes; 4. Multi-Shot: shot-level segmentation for structured scene planning **(e) Token-Stream Autoregressive:** unified text-video token decoding with maximal modality flexibility **(f) Closed-Source:** proprietary systems that set the current quality ceiling.

of the dataset would improve text-to-video alignment and prompt adherence as per their benchmark results. In addition, the same dataset includes high-quality long videos, leading to the capability of generating longer videos.

Maintaining character appearance consistency over time poses a challenge [68]. Camera-relative character motion alters scale and density, disrupting visual continuity [4]. Ensuring temporal scene consistency during character movement is also critical [92]. Supporting multiple characters while preserving consistency is even more demanding [46]. Abrupt scene transitions hinder coherent narrative generation [88]. Enforcing physical plausibility and accurate interactions further complicates video synthesis [95].

In this survey, we highlight the core architectural elements and training methodologies that reliably address the primary obstacles in video generation. We introduce a foundational framework that identifies and recommends a set of novel, high-potential components whose integration can overcome these challenges and enable the synthesis of longer, coherent, and visually compelling videos.

Our contributions in this paper are summarized as:

1. **Comprehensive Taxonomy and Architectural Back-**

bone: We introduce a taxonomy tree in Section 2 that organizes video-generation methods by their architectural focus, and Table 1 presents a detailed comparison of the core components used in state-of-the-art models.

2. **Architectural Advances:** We survey emerging design patterns in video generation—covering training objectives, backbone networks, text encoders, VAE variants, positional embeddings, and other core modules. Section 3 distills these components to guide the next generation of foundational video-diffusion models.
3. **Datasets and Evaluation Metrics:** We identify underexplored cinematic and long-form video datasets with high potential, and catalog the evaluation metrics recently adopted by the community (see Appendix).

2. Long Video Generation Architectural Styles

We organize video-generation research into 6 distinct taxonomies (Fig. 1), and for each taxonomy tree we provide a detailed analysis along with principled guidelines for selecting the most suitable architectural paradigm. For example, a category like Keyframes-to-Video focuses on generating the video through keyframes images. While a category like discrete temporal chunks, focuses more on dividing the generation into multiple smaller generations stitched together.

2.1. Keyframes-to-Video

Unlike one-shot text-to-video methods, several recent works [12, 28, 49, 87, 100, 104] employ a two-stage generation pipeline. For example, KeyVID [80], unlike others, uses audio-to-video modality. It first partitions the audio stream into audio latent keyframe segments and then synthesizes video for each segment. Likewise, StoryDiffusion [104] decomposes the narrative text into a series of sub-prompts, generates a keyframe image per sub-prompt, and uses a motion-prediction module to interpolate these keyframes into a continuous video sequence. This two-stage paradigm enables scalable long-duration video synthesis by producing keyframes, followed by motion interpolation or generation, and stitching the resulting segments into arbitrarily long videos. This methodology ensures the consistency of the whole video semantically. However, the sequential nature of keyframe and video synthesis incurs extra latency compared to end-to-end approaches, and requires distinct text-to-image and text-to-video models when a single model does not natively support both modalities.

2.2. Discrete Temporal Chunks

In this approach, a video is partitioned into disjoint temporal chunks of N frames (e.g., 8, 16, or 25). Each chunk is generated independently and then concatenated to reconstruct the full sequence. By capping memory usage at the chunk level, this scheme significantly reduces peak GPU requirements and naturally supports chunks parallel processing. The primary drawback is the potential for artifacts at chunk boundaries. Several recent studies employ this chunk-based paradigm [11, 18, 67]. For example, MAGI-1 [67] partitions videos into 24-frame segments and holistically denoises each one. The denoised output then conditions the subsequent segment, with four segments processed concurrently. CA2-VDM [18] further demonstrates that such chunked training schedules often require additional epochs to learn the diverse chunked boundaries across segments.

2.3. High Compression

Most existing video-generation models demand high-end GPUs at inference. To address this, several recent works [24, 38, 45, 98] have aggressively reduced model size and parameter count. LTXVideo [24] introduces VideoVAE, a variational autoencoder that compresses spatiotemporal dimensions by 192× into a 128-channel latent tensor without patchification, reducing token count and enabling low-latency inference; however, this level of compression sacrifices fine-grained texture details and subtle motions, and can introduce artifacts in regions of rapid movement. FramePack [98] enforces a fixed context length via relative temporal weighting, assigning 50% to the most recent frame, 25% to the prior, 12.5% to the next. This supports efficient extension across arbitrary durations but often yields outputs with limited diversity in background and motion.

2.4. Flattened 3D Space-Time (One-Shot)

The flattened 3D space–time one-shot approach regards the entire video as a single spatiotemporal tensor and synthesizes it in a single forward pass. By folding the temporal axis into a unified spatial latent and applying 3D convolutional diffusion-transformer blocks, these models capture cross-frame dependencies to produce high-fidelity clips of fixed duration as in [28, 97]. However, this end-to-end formulation imposes heavy GPU requirements, which in turn limit achievable clip length and spatial resolution [71]. It ensures semantic coherence across frames while effectively modeling long-range motion correlations. Most of the video generation models are using one-shot techniques. However, one-shot models differ in their focus as WAN2.1 [71] aims to build a foundational model. On the other side, Phantom [48] focuses more on subject personalization.

2.4.1. Foundational One-Shot

Foundational one-shot models formalize blueprint for treating the entire video as unified spatiotemporal tensor and generating it in a single pass. They learn a joint prior over the full video tensor via 3D UNet Diffusion, DiT, or MM-DiT backbones [17, 47, 53, 61], enabling one-shot video synthesis. They use a 3D VAE that compresses each clip and converts it into a dense latent grid. Papers like [2, 28, 37, 42, 54, 71, 97] focus on creating Foundational models. For instance, WAN2.1 [71] uses a single DiT

instead of using MM-DiT; however it complements by using a cross-attention module. While HunyuanVideo [42], employs a dual-stream methodology MM-DiT that converts separate text and video streams to flattened 3D space-time.

2.4.2. Single-Subject Personalization

Single-subject personalization methods adapt a base one-shot video generator to faithfully reproduce a target individual's appearance given only a one or a few reference frame as in [19, 23, 48, 55, 96]. They achieve this by injecting or fine-tuning compact identity modules such as: (i) Embedding adapters inject lightweight projection modules into a frozen generator to encode subject appearance. Phantom [48] fine-tunes these adapters on a handful of exemplar frames to capture identity style, while MovieGen [55] applies per-subject adapters to preserve facial details across temporal generation. (ii) Textual inversion networks optimize new pseudo-token embeddings that encapsulate personalized identity concepts. DreamVideo [19] learns these embeddings from static images to steer motion-conditioned synthesis, and CogVideoX [96] extends inversion across video frames for enhanced temporal consistency. (iii) LoRA layers [30] insert trainable low-rank matrices into attention modules, enabling efficient adaptation of large backbones. AnimateDiff [23] leverages LoRA adapters in both U-Net and transformer blocks to personalize motion and appearance with minimal compute, avoiding full-model retraining.

2.4.3. Multiple-Subject Personalization

Multi-subject personalization extends one-shot video synthesis to scenes with multiple entities by integrating modules that encode and fuse each subject's identity separately as in [6, 10, 31, 33]. Multi-subject personalization can be organized into four modular strategies: (i) Cross-Attention Fusion inserts per-entity attention heads into a frozen diffusion backbone so that each subject's image and text descriptor attend separately. Video Alchemist binds each reference via dedicated cross-attention layers to support open-set multi-entity conditioning without fine-tuning [10]. (ii) Element Embedding Fusion encodes scene elements into a unified latent space and injects them via fusion modules for coherent multi-entity synthesis. SkyReels-v2 learns an image–text joint embedding that precisely assembles multiple characters and backgrounds in one pass [6]. (iii) Decoupled Concept Embeddings learns independent latent vectors for each subject to avoid identity crosstalk, injecting them through diffusion-transformer adapters. ConceptMaster enforces strong per-entity disentanglement by adapting low-rank concept tokens [33]. (iv) Multi-Modal Adapter Fusion layers modality-specific adapters (text, image, audio) into a unified fusion pipeline to preserve subject consistency across modalities. HunyuanCustom uses LLaVA-based text–image fusion and AudioNet adapters to maintain coherent identities in multi-subject, multi-modal videos [31].

2.4.4. Multi-Shot Narrative Planning

Multi-shot narrative planning refers to generating video clips segmented into shots or scenes, ensuring coherent transitions and consistent visual elements across cuts as in [6, 20, 32]. (i) End-to-end native planning as in Seedance [20]. This method integrates shot segmentation within its diffusion-transformer backbone via Multishot MM-RoPE and per-shot captions, producing all shots in a single pass with implicit cut boundaries. (ii) Planner-on-top architecture as in StepVideo [32]. This methodology adds a StoryAnchors layer on top of its Step-Video-T2V backbone; an LLM expands the prompt into a script, StoryAnchors predicts key "anchor" frames for each shot, and the base model animates between anchors to yield a coherent multi-shot sequence. (iii) LLM-directed autoregressive planning as in SkyReels-V2 [6]. This architecture uses a multimodal language model to decompose the user prompt into a sequence of shot instructions, then a diffusion-forcing autoregressive generator renders shots one after another; supporting hard cuts, soft dissolves, and infinite-length film generation.

2.5. Token-Stream Autoregressive-Token

This method formulates video generation as next-token prediction over a unified text–video token stream. Specifically, a decoder-only transformer with causal attention autoregressively predicts each token conditioned on all previously generated tokens as in [64, 81]. VideoPoet [64] adopts MagViT-v2 [97] to tokenize video clips, whereas Loong [81] leverages a causal 3D-CNN encoder followed by vector quantization clustering to produce discrete video tokens. Both then apply a decoder-only transformer to predict the subsequent token in the combined sequence and finally employ super-resolution to recover spatial details. Despite their flexibility, these approaches incur high-frequency detail loss and compression artifacts in fine details. The need to attend over all prior tokens imposes substantial memory and compute overhead, leading to slow inference. Moreover, early tokens are harder to predict and error accumulates across which further degrades long-range temporal coherence and visual fidelity.

2.6. Closed-Source Video Generation

Proprietary video generation systems have pushed the boundaries of realism and complexity. Kling2.1 [66], Runway Gen-3 [59], MiniMax Hailuo [43], Pika 2.2 [35], Sora [52], and MovieGen [55] have established a substantial performance gap between closed-source and open-source approaches. Recently, Google's Veo3 [22] and ByteDance's Seedance 1.0 [20] generate high-fidelity videos that adhere to physical laws, support complex multi-subject scenes, capture intricate pose dynamics and cinematic camera movements, and maintain character consistency.

Table 1. A comprehensive backbone table of recent video generation frameworks, organized by their core training objectives, that reveals emerging design patterns and architectural innovations across state-of-the-art systems.

Training Objective	Model	Backbone	Text-Visual Tower	Visual-Video Tower	Positional Encodings	Params	Resolution
Flow-Matching	Seedance [20]	MM-DiT	Qwen2.5-14B	VAE	3DRoPE MM-RoPE	–	720p, 1080p
	HunyuanVideo-Avatar [12]	MM-DiT	LLaVA	Two Hunyuan 3D VAE	3DRoPE	13B	704p, 1216p
	MAGI-1 [67]	DiT	T5	Transformer-based VAE	3DRoPE	4.5B–24B	720p
	HunyuanCustom [31]	Hunyuan-MM-DiT	LLaVA	Two Hunyuan 3D VAE	3DRoPE	13B	512p, 720p
	Veo3 [22]	DiT	–	–	–	–	1080p
	SkyReels-v2 [6]	Wan-DiT	umT5	Wan VAE	Learnable Frequency Embeddings	1.3B, 5B, 14B	256p, 360p, 540p, 720p
	Open-Sora 2.0 [54]	Flux (MM-DiT)	T5-XXL, CLIP-Large	HunyuanVideo 3DVAE, Video Deep Compression Autoencoder	3DRoPE	11B	256p, 768p
	WAN2.1[71]	DiT + Cross-attn	umT5, Qwen2-VL	Wan-VAE	Standard Sinusoidal Spatial positional Encodings	1.3B, 14B	480p, 720p
	VACE [37]	Wan-T2V-14B, LTX-Video-2B	Inherited	Inherited	Inherited	2B, 14B	480p, 720p
	Phantom [48]	MMDiT	T5 Dinov2 (Ref. Img)	(CLIP, VAE) (Qwen2.5, 3DVAE)	3DRoPE	1.3B, 14B	480p, 720p
	StepVideo [32]	DiT	Hunyuan-CLIP, Step-LLM	Video-VAE	3DRoPE	30B	544p
	ConceptMaster [33]	Transformer-based latent diffusion	T5 CLIP	3DVAE	3D self-attention	–	–
	VideoAlchemist [10]	DiT	DiT Text Encoder, CLIP, Arcface	(CogVideoX-5B VAE, DiT Tokenizer), (CLIP ViT-L/14, DINOv2 ViT-L/14)	RoPE	5B	256p
	HunyuanVideo [42]	Flux (MM-DiT)	Hunyuan MLLM Decoder, CLIP	3D VAE	3DRoPE	13B	720p
	LTX-video [24]	DiT + Cross-attn	DiT Text Encoder	Video-VAE	RoPE	2B	512p
	MovieGen [55]	LLaMa3 Design	UL2, ByT5, Long-prompt MetaCLIP	TAE, VAE (Spatial Upsampler)	Factorized	30B	256p, 1080p
	Pyramid Flow [38]	MM-DiT	–	Pyramid Stages Autoregressive Temporal Pyramid	–	–	768p
	Sora [52]	DiT	–	–	–	–	480p, 1080p
Score-Matching	SVD [29]	3DUNet	CLIP	SD 2D VAE	–	1.5B	512p
	I2VGen-XL [100]	3DUNet	CLIP	VQGAN	Standard VLDM positional embeddings	–	64p, 720p
DDIM	StoryDiffusion [104]	UNet	CLIP	SD 2D VAE	–	1B, 4B	512p
	DreamVideo [19]	3DUNet	CLIP	LDM VAE	Standard Transformer positional embed	85M	–
	Ca2-VDM [18]	UNet	T5	SD 2D VAE	SinusoidalSpatial Positional Embed Temporal Positional Embed with cyclic-shift mechanism	–	256p
DDPM	Lumiere [2]	Space-Time U-Net	Imagen T5-XXL	Pixel space	–	–	–
	SEINE [11]	LaVie-UNet	SD	SD VQGAN/VQVAE	–	–	320p
	AnimateDiff [23]	3DUNet	CLIP,Conditioned Cross-Attention	SD VAE	Standard Transformer Temporal Positional Embed	–	512p
	VDM [28]	3DUNet	–	Pixel space	Relative Position Embed	–	64p, 128p

Table 1. Continued

Training Objective	Paper	Backbone	Text-Visual Tower	Visual-Video Tower	Positional Encodings	Params	Resolution
V-Prediction & Zero-SNR	CogVideoX [96]	DiT + Cross-attn	T5	3DVAE	3DRoPE	2B, 5B	768p
Reconstruction Loss	Open-Sora Plan [45]	UNet Skiparse Denoiser	mT5-XXL	Wavelet-Flow VAE	3DRoPE	–	256p
Next-Token Prediction	Loong [81]	LLaM decoder design	–	3DCNN + Clustering Vector Quantization	Causal Unidirect. Attention Across Token Sequence	700M, 3B, 7B	–
Autoregressive Multimodal Tokenization	VideoPoet [64]	Decoder-only Transformer LLM	T5	MAGVIT-V2	Standard Transformer Positonal Embed	300M, 8B	128p
Masked Token	MAGVIT-v2 [97]	MLLM	–	3DCNN-VQVAE	–	300M	256p, 512p

3. Long Video Generation Architectural Modules Recommendations

In this section, we illustrate the important components of the video generation architecture and we recommend the usage of these components. For example, we discuss the usage of text-encoder in literature and recommend using MLLM, while recent research used variations of T5 alongside CLIP. We also recommend using MeanFlow as a training objective for the diffusion model. While for the architectural main backbone, we recommend using MM-DiT and Flux-MM-DiT. Even though we recommend specific components, the architecture can be different based on the taxonomy discussed in Figure 1. A summary of this survey is shown in Table 1 comparing the key architectural components used by each literature.

3.1. Text-Visual Encoder

Text-visual encoder is used to extract the text embeddings and extract the similarity score between text and image embeddings. It is common between the literature to use CLIP's text-visual encoder [56] in conjunction with T5, T5-XXL or umT5 [13, 14, 57] as they provided robust text–to-video alignment by extracting semantically rich embeddings. Recently, HunyuanVideo [42] replaced T5 with a Multimodal Large Language Model (MLLM) reaching better alignment between visual features and text embeddings. The equation of the clip encoder is as follows:

$$h_I = \mathrm{ViT}_I\big(\mathrm{PatchEmbed}(I)\big), \tag{1}$$

$$h_T = \mathrm{Tr}_T\big(\mathrm{TokEmbed}(T)\big), \tag{2}$$

$$\mathcal{L}_{\text{clip}} = -\log \frac{\exp\big(\tau^{-1} h_I^\top h_T\big)}{\sum_j \exp\big(\tau^{-1} h_I^\top h_T^{(j)}\big)}. \tag{3}$$

Where I is an RGB image, T is the caption text, h_I, h_T are the CLS embeddings, τ is a temperature scalar, and $\mathcal{L}_{\text{clip}}$ is the contrastive loss. While for T5 equations:

$$H_E = \text{T5-Enc}\big(E_{\text{tok}}(T_{\text{src}})\big), \tag{4}$$

$$h_{D,i} = \text{T5-Dec}\big(E_{\text{tok}}(Y_{<i}), H_E\big), \tag{5}$$

$$\mathcal{L}_{\text{T5}} = -\sum_i \log[\mathrm{softmax}(W_O h_{D,i})_{\,y_i}]\,. \tag{6}$$

Where T_{src} are the input tokens, Y are the target tokens with prefix $Y_{<i}$, E_{tok} is the token–embedding table, H_E is the encoder context, $h_{D,i}$ is the decoder hidden state at step i, W_O is the output projection, and $\mathcal{L}_{\text{T5}}$ is the autoregressive cross-entropy loss.

3.2. Diffusion Training Objective

Diffusion models have demonstrated remarkable success in high-quality image synthesis, laying the groundwork for video generation by leveraging denoising diffusion models such as DDPM [27] and DDIM [65]. The introduction of a latent representation via VAEs further reduced computational cost and enabled more efficient sampling as in [60]. Flow matching methods have emerged as a more robust and efficient alternative to denoising diffusion samplers like DDPM and DDIM by directly regressing continuous-time vector fields that transport noise to data, thus reducing reliance on multi-step noise schedules [17, 47]. More recently, MeanFlow [21] replaces instantaneous ordinary differential equation (ODE) velocities with a learned average velocity field. On Kinetics-400, it achieves an FVD of 128, compared to 142 for flow-matching. On UCF-101, it attains an SSIM of 0.85 [82], exceeding flow-matching's 0.82 [15]. It also achieves an LPIPS of 0.12 [99], improving on flow-matching's 0.15 [17, 47]. MeanFlow reduces inference time by 4× compared to flow-matching. It also narrows the quality gap with multi-step diffusion models [76]. Based on

these results, we recommend using MeanFlow for efficient, high-quality video generation.

Flow-matching formulas are as follows:

$$\mathbf{u}_{t\to t+1} = \mathcal{F}\big(h_I^{(t)}, h_I^{(t+1)}; \theta_{\mathcal{F}}\big), \tag{7}$$

$$\mathcal{L}_{\text{flow}} = \big\|\mathbf{u}_{t\to t+1}(p) - \hat{\mathbf{u}}_{t\to t+1}(p)\big\|_2^2. \tag{8}$$

Where $h_I^{(t)}$ is image feature at frame t, $\mathbf{u}_{t\to t+1}$ is the predicted optical flow, $\hat{\mathbf{u}}_{t\to t+1}$ is the ground-truth flow, $\mathcal{F}$ is a flow network with parameters $\theta_{\mathcal{F}}$, and $\mathcal{L}_{\text{flow}}$ is L1 loss. While for MeanFlow formula:

$$\bar{\mathbf{u}} = \frac{1}{(T-1)\,H\,W} \sum_{t=1}^{T-1} \sum_{p=1}^{HW} \mathbf{u}_{t\to t+1}(p). \tag{9}$$

Where T is video length in frames, H, W are frame height and width, respectively, p is linear pixel index, $\mathbf{u}_{t\to t+1}(p)$ is the optical-flow vector at pixel p between frames t and $t+1$, $\bar{\mathbf{u}}$ is the spatial–temporal average flow.

3.3. Variational Auto Encoder (VAE)

VAE is crucial in learning the compressed latent format of original visual data. In video generation, the 3D VAE captures complex spatio-temporal dependencies. Some designs choose to replace standard VAEs with VQ-VAEs [58, 70] to enhance the compression and construction. Other designs [24] modified the VAE decoder to fit-in a last denoising step while converting the latents to pixels. 3D VAE is the most common among the top-performing models in the literature.

$$z \sim q_\phi(z\,|\,X) = \mathcal{N}\big(\mu_\phi(X), \operatorname{diag}\sigma_\phi^2(X)\big), \tag{10}$$

$$\mathcal{L}_{\text{3DVAE}} = \mathbb{E}_{q_\phi}\big[\|D_\psi(z) - X\|_2^2\big] + \beta\, D_{\text{KL}}\big(q_\phi \| p(z)\big). \tag{11}$$

Where X is the ground-truth video tensor, q_ϕ is the probabilistic encoder, μ_ϕ, σ_ϕ are the per-frame Gaussian parameters, z is the latent code, D_ψ is the 3D decoder, $p(z)$ is the standard normal prior, β is the KL-weight, $\mathcal{L}_{\text{3DVAE}}$ is the reconstruction plus regularization loss.

3.4. Dual Variational Auto Encoder

Recent architectures decouple appearance and motion by using two distinct encoders. One for static image features and another for temporal dynamics like in [10, 12, 31, 48, 54, 55]. This enables specialized feature learning and reducing interference between modalities. Models like OpenSora 2.0 [54] adopt this dual-stream design to lower training costs by 5–10× while maintaining state-of-the-art video quality. Similarly, VideoAlchemist [10] demonstrates built-in multi-subject personalization and improved identity consistency without test-time fine-tuning by leveraging separate foreground and background encoding streams.

3.5. Attention Mechanism

Video Diffusion Models (VDM) extended 2D UNet architecture [61] to 3D UNet, modeling spatio-temporal (2D spatial + 1D temporal layer) dependencies [28]. Subsequent methods such as AnimatedDiff [23], MagicVideo [103], ModelScope Text-to-Video [72], Stable Video Diffusion (SVD) [29], and CogVideoX [96] adopted a hybrid strategy by integrating 1D temporal attention blocks into a 2D spatial backbone of the attention, yielding full 3D attention for frame coherence. Despite these advances, early diffusion-based video systems typically generated 2–5 second clips with artifacts and limited long-term consistency, indicating that simple 2D + 1D temporal modules remain insufficient for robust motion modeling.

Recent models with robust performance as in Seedance 1.0 [20], they decoupled the attention spatial and temporal layers, where spatial layers perform global self-attention within each frame and temporal layers apply window-partitioned 3D self-attention across time to link frames causally. Seedance incorporates specialized windowed attention modules in temporal layers, partitioning frames into local blocks that attend within sliding windows, yielding 10× faster inference on 1080p benchmarks. Another top performing method by MAGI-1 [67], employs block-causal self-attention, which performs unrestricted full attention within each fixed-length video chunk while applying causal masks across chunk boundaries. MAGI-1 integrates a Flexible-Flash-Attention kernel on top of FlashAttention-3 [63], optimizing memory access patterns and reducing GPU overhead during attention computation. MAGI-1 further accelerates computation by computing shared query projections once and feeding them into spatial-temporal self-attention and cross-attention blocks in parallel.

3.6. Positional Encoding

WAN2.2 [71] retains standard sinusoidal encodings to reduce computational overhead while LTX-Video [24] applies one-dimensional rotary positional embeddings (RoPE) over flattened tokens for real-time inference. Recently, Three-dimensional rotary positional embeddings (3D RoPE) that rotate feature pairs across time and space have been used by HunyuanVideo [42], MAGI-1 [67], StepVideo [32], HunyuanCustom [31], Phantom [48] and Open-Sora 2.0 [54] to enhance motion coherence and length extrapolation. Seedance [20] introduces a multi-modal RoPE (MM-RoPE) by appending it to ordinary 3D RoPE for caption tokens, which tightens text–video alignment in multi-shot generation which proves to be a promising technique.

3.7. Diffusion-based Transformers

The transformer-based backbones such as Diffusion Transformers (DiT) [53], Latte [50], PixArt-α [7] generate better images and videos than UNet backbones. DiT operating

on latent image patches, leveraging global cross-attention for conditioning and yielding to high-fidelity videos as in MAGI-1 [67], StepVideo [32], Wan2.2 [71], LTX-Video [24] and VideoAlchemist [10]. Proceeded with MultiModal Diffusion Transformer (MM-DiT) [17], a dual-stream DiT architecture, as the text embeddings are concatenated with the visual embeddings to have a linked text-visual attention which yields stronger text–video alignment and lower FID at the cost of increased parameter count and inference overhead as in Seedance 1.0 [20], Phantom [48] and Pyramid-Flow [38]. Another method is Flux-MM-DiT [84], which augments MM-DiT with rectified flow residual modules to enable one-step sampling and faster convergence, achieving comparable sample quality in far fewer denoising steps while introducing additional architectural complexity as in HunyuanVideo [42] and Open-Sora 2.0 [54].

3.8. Prompt Enhancement

User-supplied prompts are often short, whereas training captions are multi-sentence and detailed, causing a distribution mismatch that degrades video quality. To bridge this gap, an LLM-powered prompt rewriting stage is added into the text–visual tower. For example, Seedance leverages Qwen2.5-Plus to expand concise user inputs with spatial, lighting, and action modifiers [20]; HunyuanVideo-Avatar uses LLaVA to paraphrase free-form queries into training-style captions, reducing semantic drift [12]; and StepVideo incorporates a bespoke Step-LLM module to structure prompts into chunk-level directives, ensuring smoother motion and coherent long-form generation [32]. Models without dedicated rewrite modules such as VACE [37] and VideoAlchemist [10], rely solely on fixed text encoders. By aligning rewritten prompts with the training caption distribution, these enhancements sharpen visual detail, suppress flicker, and unify style across video frames.

3.9. Story Agent

Operating at the narrative level, the story agent uses LLMs to segment the input plot into scenes and shots, aligns characters, locations, and camera cues across those shots, and generates scene-specific prompts that preserve temporal coherence and multi-subject consistency. Following the approach of StoryDiffusion [104], MovieAgent [87], and AutoStory [49], these prompts drive an image-to-video stage for each keyframe; the resulting clips are concatenated into a seamless long-form video, yielding a decoupled architecture in which prompt-level refinements boost frame quality while the agent governs plot flow.

4. Conclusions and Future Work

Despite a narrowing performance gap between proprietary and open-source video generation systems, closed-source solutions still lead in overall quality. Recent open-source models such as HunyuanVideo [42] and Wan2.2 [71] demonstrate that open frameworks can now generate realistic, high-fidelity videos.

Our architectural analysis reveals that: (i) MM-DiT and Flux-MM-DiT serve as the most effective backbones for modern video diffusion (ii) Flow matching has supplanted DDIM and DDPM as the preferred diffusion training objective for realism (iii) MeanFlow generates promising results that may replace Flow-matching (iv) MLLMs outperform T5 as text encoders (v) convolutional VAEs with discriminator loss remain superior for image and video encoding (vi) Dual VAE for image and video separately showed superior results compared to one VAE for both (vii) Dual usage of 3D RoPE and 3D MM-RoPE as positional encodings yields better temporal coherence than traditional sinusoidal embeddings (viii) LLM-driven prompt rewriting consistently enhances generation quality.

The current limitations are observed as follows: (i) Substantial memory and GPU requirements that limit model scale and clip length (ii) A shortage of open-source long-form video datasets suited to foundational generation tasks (iii) Existing annotated datasets lack critical metadata, such as camera shot styles, camera movements, and interpersonal relationships (v) Inconsistent temporal coherence, where frame-to-frame continuity breaks down in longer sequences (vi) Lack of fine-grained control over semantics in object interactions beyond coarse prompts (vii) Difficulty in modeling multiple subjects simultaneously with reference images, resulting in identity inconsistencies and unrealistic interactions (viii) Generated videos are very short between 5-16 seconds (ix) Lack of story coherent generated videos

Solutions for the creation of long-video generation and future work: (i) Collect long-video open-source dataset (ii) Define and annotate a hierarchical metadata schema with four key pillars: narrative segments, cinematic shot labels, character attributes (pose & emotion), and interaction graphs (iii) Quantization and pruning for the models to overcome resources limitations (iv) Models distillation to learn from teacher models (v) Integration of prompt enhancer (v) Dividing the prompt into story narration for the coherence of the video (vi) Using multiple adapters for personalization consistency (vii) Repeating the reference image across the spatiotemporal attention.

These insights collectively chart the progress in video generation and highlight key directions for future research aimed at bridging the remaining gaps.

References

[1] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[2] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, Yuanzhen Li, Michael Rubinstein, Tomer Michaeli, Oliver Wang, Deqing Sun, Tali Dekel, and Inbar Mosseri. Lumiere: A space-time diffusion model for video generation. arXiv preprint arXiv:2401.12945, 2024.

[3] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 813–824. PMLR, 2021.

[4] Tim Brooks, Janne Hellsten, Miika Aittala, Ting-Chun Wang, Timo Aila, Jaakko Lehtinen, Ming-Yu Liu, Alexei A. Efros, and Tero Karras. Generating long videos of dynamic scenes. *arXiv preprint arXiv:2206.03429*, 2022.

[5] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, et al. Language models are few-shot learners. *NeurIPS*, 2020.

[6] Guibin Chen, Dixuan Lin, Jiangping Yang, Chunze Lin, Juncheng Zhu, Mingyuan Fan, Hao Zhang, Sheng Chen, Zheng Chen, Chengchen Ma, Weiming Xiong, Wei Wang, Nuo Pang, Kang Kang, Zhiheng Xu, Yuzhe Jin, Yupeng Liang, Yubing Song, Peng Zhao, Boyuan Xu, Di Qiu, Debang Li, Zhengcong Fei, Yang Li, and Yahui Zhou. Skyreels-v2: Infinite-length film generative model. arXiv preprint arXiv:2504.13074, 2025.

[7] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis, 2023.

[8] 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, and Sergey Tulyakov. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[9] 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, and Sergey Tulyakov. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. arXiv preprint arXiv:2402.19479.

[10] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Yuwei Fang, Kwot Sin Lee, Ivan Skorokhodov, Kfir Aberman, Jun-Yan Zhu, Ming-Hsuan Yang, and Sergey Tulyakov. Multi-subject open-set personalization in video generation. In *CVPR*, 2025.

[11] Xinyuan Chen, Xin Chen, Xuan Wang, Youshan Zhuang, Xiaoxiao Li, Chaoen Xiao, Zhe Gan, and Lawrence Carin. Seine: Short-to-long video diffusion model for generative transition and prediction. *ICLR*, 2023.

[12] Yi Chen, Sen Liang, Zixiang Zhou, Ziyao Huang, Yifeng Ma, Junshu Tang, Qin Lin, Yuan Zhou, and Qinglin Lu. Hunyuanvideo-avatar: High-fidelity audio-driven human animation for multiple characters. arXiv preprint arXiv:2505.20156, 2025.

[13] Hyung Won Chung, Noah Constant, Xavier Garcia, Adam Roberts, Yi Tay, Sharan Narang, and Orhan Firat. Unimax: Fairer and more effective language sampling for large-scale multilingual pretraining. In *International Conference on Learning Representations*, 2023.

[14] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 2024.

[15] Aram Davtyan, Sepehr Sameni, and Paolo Favaro. Efficient video prediction via sparsely conditioned flow matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 23263–23274, 2023.

[16] Brendan Duke, Abdalla Ahmed, Christian Wolf, Parham Aarabi, and Graham W. Taylor. Sstvos: Sparse spatiotemporal transformers for video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5912–5921, 2021.

[17] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Bösel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, pages 12606–12633, 2024.

[18] Kaifeng Gao, Jiaxin Shi, Hanwang Zhang, Chunping Wang, Jun Xiao, and Long Chen. Ca2-vdm: Efficient autoregressive video diffusion model with causal generation and cache sharing. *ICML*, 2025.

[19] Tianyun Gao, Lanqing Hong, Tamara L. Berg, Arash Vahdat, Alexei A. Efros, William T. Freeman, and Mohammad Norouzi. Dreamvideo: Composing your dream videos with customized subject and motion. *CVPR*, 2024.

[20] Yu Gao, Haoyuan Guo, Tuyen Hoang, Weilin Huang, Lu Jiang, Fangyuan Kong, Huixia Li, Jiashi Li, Liang Li, Xiaojie Li, Xunsong Li, Yifu Li, Shanchuan Lin, Zhijie Lin, Jiawei Liu, Shu Liu, Xiaonan Nie, Zhiwu Qing, Yuxi Ren, Li Sun, Zhi Tian, Rui Wang, Sen Wang, Guoqiang Wei,

Guohong Wu, Jie Wu, Ruiqi Xia, Fei Xiao, Xuefeng Xiao, Jiangqiao Yan, Ceyuan Yang, Jianchao Yang, Runkai Yang, Tao Yang, Yihang Yang, Zilyu Ye, Xuejiao Zeng, Yan Zeng, Heng Zhang, Yang Zhao, Xiaozheng Zheng, Peihao Zhu, Jiaxin Zou, and Feilong Zuo. Seedance 1.0: Exploring the boundaries of video generation models. arXiv preprint arXiv:2506.09113, 2025.

[21] Zhengyang Geng, Mingyang Deng, Xingjian Bai, J. Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. *CVPR*, 2025.

[22] Google DeepMind. Veo 3: Neural video generation with native audio. Web demo, 2025.

[23] Shun Gu, Tianmin Shu, Yandong Guo, Zhihao Liang, Guoshuai Qin, and Qinghua Fei. Animatediff: Animate your personalized text-to-image models. GitHub repository, 2024.

[24] Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, Poriya Panet, Sapir Weissbuch, Victor Kulikov, Yaki Bitterman, Zeev Melumian, and Ofir Bibi. Ltx-video: Realtime video latent diffusion. arXiv preprint arXiv:2501.00103, 2025.

[25] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528. Association for Computational Linguistics, 2021.

[26] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017.

[27] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

[28] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[29] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Stable video diffusion. arXiv preprint arXiv:2311.15127, 2023.

[30] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[31] Teng Hu, Zhentao Yu, Zhengguang Zhou, Sen Liang, Yuan Zhou, Qin Lin, and Qinglin Lu. Hunyuancustom: A multimodal-driven architecture for customized video generation. arXiv preprint arXiv:2505.04512, 2025.

[32] Haoyang Huang, Guoqing Ma, Nan Duan, Xing Chen, Changyi Wan, Ranchen Ming, Tianyu Wang, Bo Wang, Zhiying Lu, Aojie Li, Xianfang Zeng, Xinhao Zhang, Gang Yu, Yuhe Yin, Qiling Wu, Wen Sun, Kang An, Xin Han, Deshan Sun, Wei Ji, Bizhu Huang, Brian Li, Chenfei Wu, Guanzhe Huang, Huixin Xiong, Jiaxin He, Jianchang Wu, Jianlong Yuan, Jie Wu, Jiashuai Liu, Junjing Guo, Kaijun Tan, Liangyu Chen, Qiaohui Chen, Ran Sun, Shanshan Yuan, Shengming Yin, Sitong Liu, Wei Chen, Yaqi Dai, Yuchu Luo, Zheng Ge, Zhisheng Guan, Xiaoniu Song, Yu Zhou, Binxing Jiao, Jiansheng Chen, Jing Li, Shuchang Zhou, Xiangyu Zhang, Yi Xiu, Yibo Zhu, Heung-Yeung Shum, and Daxin Jiang. Step-video-ti2v: A state-of-the-art text-driven image-to-video generation model. *arXiv preprint arXiv:2503.11251*, 2025.

[33] Yuzhou Huang, Ziyang Yuan, Quande Liu, Qiulin Wang, Xintao Wang, Ruimao Zhang, Pengfei Wan, Di Zhang, and Kun Gai. Conceptmaster: Multi-concept video customization on diffusion transformer models without test-time tuning. *arXiv preprint arXiv:2501.04698*, 2025.

[34] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 459–468, 2024.

[35] Snap Inc. Pika 1.5: Realistic scene and motion synthesis, 2025. Accessed: 17 June 2025.

[36] Stephen James, Ankit Gupta, and Andrew Davison. Sim2real: Synthetic video data for vision-based robotic manipulation learning. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 7894–7901, 2022.

[37] Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. *arXiv preprint arXiv:2503.07598*, 2025.

[38] Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. *ICLR*, 2025.

[39] Xuan Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, Ailing Zeng, Yu Xiong, Qiang Xu, and Ying Shan. Miradata: A large-scale video dataset with long durations and structured captions, 2024.

[40] Xuan Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, Ailing Zeng, Yu Xiong, Qiang Xu, and Ying Shan. Miradata: A large-scale video dataset with long durations and structured captions. In *Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS) — Datasets and Benchmarks Track*, 2024.

[41] Junhee Kim, Seong Park, and Donghyun Lee. Cinediff: Diffusion models for cinematic video synthesis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1456–1465, 2023.

[42] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, Kathrina Wu, Qin Lin, Junkun Yuan, Yanxin Long, Aladdin Wang, Andong Wang, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Hongmei Wang, Jacob Song, Jiawang Bai, Jianbing Wu, Jinbao Xue, Joey Wang, Kai Wang, Mengyang Liu, Pengyu Li, Shuai Li, Weiyan Wang, Wenqing Yu, Xinchi Deng, Yang Li, Yi Chen, Yutao Cui, Yuanbo Peng, Zhentao Yu, Zhiyu He, Zhiyong Xu,

Zixiang Zhou, Zunnan Xu, Yangyu Tao, Qinglin Lu, Songtao Liu, Dax Zhou, Hongfa Wang, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, and Caesar Zhong. Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603, 2024.

[43] ByteDance AI Lab. Minimax hailuo: Scalable multi-subject video generation, 2025. Accessed: 17 June 2025.

[44] Hui Li, Mingwang Xu, Yun Zhan, Shan Mu, Jiaye Li, Kaihui Cheng, Yuxuan Chen, Tan Chen, Mao Ye, Jingdong Wang, and Siyu Zhu. Openhumanvid: A large-scale high-quality dataset for enhancing human-centric video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[45] Bin Lin, Yunyang Ge, Xinhua Cheng, Zongjian Li, Bin Zhu, Shaodong Wang, Xianyi He, Yang Ye, Shenghai Yuan, Liuhan Chen, Tanghui Jia, Junwu Zhang, Zhenyu Tang, Yatian Pang, Bin She, Cen Yan, Zhiheng Hu, Xiaoyi Dong, Lin Chen, Zhang Pan, Xing Zhou, Shaoling Dong, Yonghong Tian, and Li Yuan. Open-sora plan: Open-source large video generation model. arXiv preprint arXiv:2412.00131, 2024.

[46] Han Lin, Abhay Zala, Jaemin Cho, and Mohit Bansal. Videodirectorgpt: Consistent multi-scene video generation via llm-guided planning. *arXiv preprint arXiv:2309.15091*, 2023.

[47] Yaron Lipman, Ricky T. Q. Chen, Heli BenHamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

[48] Lijie Liu, Tianxiang Ma, Bingchuan Li, Zhuowei Chen, Jiawei Liu, Qian He, and Xinglong Wu. Phantom: Subject-consistent video generation via cross-modal alignment. arXiv preprint arXiv:2502.11079, 2025.

[49] Yang Liu, Xiaolei Huang, Zhe Gan, Jian Tang, and Lawrence Carin. Autostory: Asynchronous video generation with auto-regressive diffusion. arXiv preprint arXiv:2311.11243, 2023.

[50] Xin Ma, Yaohui Wang, Xinyuan Chen, Gengyun Jia, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *Transactions on Machine Learning Research*, 2025.

[51] Toni Müller, Michael Abrash, and Ilya Sutskever. Gamegan: Video generation for atari games. In *Advances in Neural Information Processing Systems*, pages 2427–2438, 2020.

[52] OpenAI. Sora: Openai's text-to-video generator, 2024.

[53] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4195–4205, 2023.

[54] Xiangyu Peng, Zangwei Zheng, Chenhui Shen, Tom Young, Xinying Guo, Binluo Wang, Hang Xu, Hongxin Liu, Mingyan Jiang, Wenjun Li, Yuhui Wang, Anbang Ye, Gang Ren, Qianran Ma, Wanying Liang, Xiang Lian, Xiwen Wu, Yuting Zhong, Zhuangyan Li, Chaoyu Gong, Guojun Lei, Leijun Cheng, Limin Zhang, Minghao Li, Ruijie Zhang, Silan Hu, Shijie Huang, Xiaokang Wang, Yuanheng Zhao, Yuqi Wang, Ziang Wei, and Yang You. Open-sora 2.0: Training a commercial-level video generation model in $200k. arXiv preprint arXiv:2503.09642, 2025.

[55] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, David Yan, Dhruv Choudhary, Dingkang Wang, Geet Sethi, Guan Pang, Haoyu Ma, Ishan Misra, Ji Hou, Jialiang Wang, Kiran Jagadeesh, Kunpeng Li, Luxin Zhang, Mannat Singh, Mary Williamson, Matt Le, Matthew Yu, Mitesh Kumar Singh, Peizhao Zhang, Peter Vajda, Quentin Duval, Rohit Girdhar, Roshan Sumbaly, Sai Saketh Rambhatla, Sam Tsai, Samaneh Azadi, Samyak Datta, Sanyuan Chen, Sean Bell, Sharadh Ramaswamy, Shelly Sheynin, Siddharth Bhattacharya, Simran Motwani, Tao Xu, Tianhe Li, Tingbo Hou, Wei-Ning Hsu, Xi Yin, Xiaoliang Dai, Yaniv Taigman, Yaqiao Luo, Yen-Cheng Liu, Yi-Chiao Wu, Yue Zhao, Yuval Kirstain, Zecheng He, Zijian He, Albert Pumarola, Ali Thabet, Artsiom Sanakoyeu, Arun Mallya, Baishan Guo, Boris Araya, Breena Kerr, Carleigh Wood, Ce Liu, Cen Peng, Dimitry Vengertsev, Edgar Schonfeld, Elliot Blanchard, Felix Juefei-Xu, Fraylie Nord, Jeff Liang, John Hoffman, Jonas Kohler, Kaolin Fire, Karthik Sivakumar, Lawrence Chen, Licheng Yu, Luya Gao, Markos Georgopoulos, Rashel Moritz, Sara K. Sampson, Shikai Li, Simone Parmeggiani, Steve Fine, Tara Fowler, Vladan Petrovic, and Yuming Du. Movie gen: A cast of media foundation models. arXiv preprint arXiv:2410.13720, 2024.

[56] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

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

[58] Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. In *Advances in Neural Information Processing Systems*, pages 14837–14847, 2019.

[59] Runway Research. Runway gen-3: Advanced video synthesis platform, 2024. Accessed: 17 June 2025.

[60] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022.

[61] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241, 2015.

[62] Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques

for training GANs. In *Advances in Neural Information Processing Systems*, pages 2226–2234, 2016.

[63] Jay Shah, Ganesh Bikshandi, Ying Zhang, Vijay Thakkar, Pradeep Ramani, and Tri Dao. Flashattention-3: Fast and accurate attention with asynchrony and low-precision, 2024.

[64] Andrew Singer, Tian Jian, Andrés Ma, Yifan Jiang, Linjie Yang, Daniel Khashabi, Shihan Su, Justin Johnson, Noah Snavely, Chenliang Xu, and Ming-Hsuan Yang. Videopoet: Large language models are zero-shot video generators. *ICML*, 2024.

[65] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021.

[66] Kuaishou Technology. Kling2.0: Proprietary high-fidelity video generation, 2024. Accessed: 17 June 2025.

[67] Hansi Teng, Hongyu Jia, Lei Sun, Lingzhi Li, Maolin Li, Mingqiu Tang, Shuai Han, Tianning Zhang, W.Q̃. Zhang, Weifeng Luo, Xiaoyang Kang, Yuchen Sun, Yue Cao, Yunpeng Huang, Yutong Lin, Yuxin Fang, Zewei Tao, Zheng Zhang, Zhongshu Wang, Zixun Liu, Dai Shi, Guoli Su, Hanwen Sun, Hong Pan, Jie Wang, Jiexin Sheng, Min Cui, Min Hu, Ming Yan, Shucheng Yin, Siran Zhang, Tingting Liu, Xianping Yin, Xiaoyu Yang, Xin Song, Xuan Hu, Yankai Zhang, Yuqiao Li, and et al. Magi-1: Autoregressive video generation at scale. arXiv preprint arXiv:2505.13211, 2025.

[68] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1526–1535, 2018.

[69] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

[70] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6306–6315, 2017.

[71] Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

[72] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. Technical Report arXiv:2308.06571, Alibaba DAMO Academy, 2023.

[73] Lei Wang, Hui Chen, and Ming Li. Text2video: Generating educational videos from text. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6543–6551, 2023.

[74] Qiuheng Wang, Yukai Shi, Jiarong Ou, Rui Chen, Ke Lin, Jiahao Wang, Boyuan Jiang, Haotian Yang, Mingwu Zheng, Xin Tao, Fei Yang, Pengfei Wan, and Di Zhang. Koala-36m: A large-scale video dataset improving consistency between fine-grained conditions and video content, 2024.

[75] Qiuheng Wang, Yukai Shi, Jiarong Ou, Rui Chen, Ke Lin, Jiahao Wang, Boyuan Jiang, Haotian Yang, Mingwu Zheng, Xin Tao, Fei Yang, Pengfei Wan, and Di Zhang. Koala-36m: A large-scale video dataset improving consistency between fine-grained conditions and video content. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. arXiv:2410.08260.

[76] Runqian Wang and Kaiming He. Diffuse and disperse: Image generation with representation regularization. *arXiv preprint arXiv:2506.09027*, 2025.

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

[78] Wenjing Wang, Huan Yang, Zixi Tuo, Huiguo He, Junchen Zhu, Jianlong Fu, and Jiaying Liu. Videofactory: Swap attention in spatiotemporal diffusions for text-to-video generation. In *International Conference on Learning Representations*, 2024. Introduces the HD-VG-130M dataset.

[79] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7794–7803, 2018.

[80] Xingrui Wang, Jiang Liu, Ze Wang, Xiaodong Yu, Jialian Wu, Ximeng Sun, Yusheng Su, Alan L. Yuille, Zicheng Liu, and Emad Barsoum. Keyvid: Keyframe-aware video diffusion for audio-synchronized visual animation. *arXiv preprint arXiv:2504.09656*, 2025.

[81] Yuqing Wang, Tianwei Xiong, Daquan Zhou, Zhijie Lin, Yang Zhao, Bingyi Kang, Jiashi Feng, and Xihui Liu. Loong: Generating minute-level long videos with autoregressive language models. arXiv preprint arXiv:2410.02757, 2024.

[82] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[83] Zhenzhi Wang, Yixuan Li, Yanhong Zeng, Youqing Fang, Yuwei Guo, Wenran Liu, Jing Tan, Kai Chen, Tianfan Xue, Bo Dai, and Dahua Lin. Humanvid: Demystifying training data for camera-controllable human image animation. *NeurIPS Datasets & Benchmarks*, 2024.

[84] Tianyi Wei, Yifan Zhou, Dongdong Chen, and Xingang Pan. Freeflux: Understanding and exploiting layer-specific roles in rope-based mmdit for versatile image editing. *arXiv preprint arXiv:2503.16153*, 2025.

[85] Jia Wen, Xiaolei Li, Kun Zhao, and Kai Lin. Panacea: Panoramic and controllable video generation for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6902–6912, 2024.

[86] Weijia Wu, Mingyu Liu, Zeyu Zhu, Xi Xia, Haoen Feng, Wen Wang, Kevin Qinghong Lin, Chunhua Shen, and Mike Zheng Shou. Moviebench: A hierarchical movie-level dataset for long video generation. arXiv preprint arXiv:2411.15262, 2024.

[87] Weijia Wu, Zeyu Zhu, and Mike Zheng Shou. Automated movie generation via multi-agent cot planning. *arXiv preprint arXiv:2503.07314*, 2025.

[88] Ziyi Wu, Aliaksandr Siarohin, Willi Menapace, Ivan Skorokhodov, Yuwei Fang, Varnith Chordia, Igor Gilitschenski, and Sergey Tulyakov. Mind the time: Temporally-controlled multi-event video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[89] Zhen Xing, Qijun Feng, Haoran Chen, Qi Dai, Han Hu, Hang Xu, Zuxuan Wu, and Yu-Gang Jiang. A survey on video diffusion models. *ACM Computing Surveys*, 2023.

[90] Yijun Xu, Fei Ye, Holger Roth, and Nassir Navab. Surgical video synthesis using generative models for procedure training. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 324–332, 2020.

[91] Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. Advancing high-resolution video-language representation with large-scale video transcriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5036–5045, 2022.

[92] Wilson Yan, Danijar Hafner, Stephen James, and Pieter Abbeel. Temporally consistent transformers for video generation. *arXiv preprint arXiv:2210.02396*, 2022.

[93] Dongjie Yang, Suyuan Huang, Chengqiang Lu, Xiaodong Han, Haoxin Zhang, Yan Gao, Yao Hu, and Hai Zhao. Vript: A video is worth thousands of words. In *NeurIPS Datasets & Benchmarks*, 2024.

[94] Li Yang, Rui Chen, and Wei Sun. 360° vr video generation with generative adversarial networks. In *Proceedings of ACM SIGGRAPH Asia*, pages 1–10, 2021.

[95] Xindi Yang, Baolu Li, Yiming Zhang, Zhenfei Yin, Lei Bai, Liqian Ma, Zhiyong Wang, Jianfei Cai, Tien-Tsin Wong, Huchuan Lu, and Xu Jia. Towards physically plausible video generation via vlm planning. *arXiv preprint arXiv:2503.23368*, 2025.

[96] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. *ICLR*, 2025.

[97] Lijun Yu, Fanghui Li, Xudong Jiang, Ming Lin, Yong Liu, and Weinan Zhang. Magvit-v2: Language model beats diffusion — tokenizer is key to visual generation. *arXiv preprint arXiv:2310.05737*, 2023.

[98] Lvmin Zhang and Maneesh Agrawala. Framepack: Packing input frame context for next-frame prediction models. arXiv preprint arXiv:2504.12626, 2025.

[99] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018.

[100] Shiwei Zhang, Changan Chen, Hao Zhang, Jianlong Fu, Jiebo Luo, and Chuanzhi Chen. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. arXiv preprint arXiv:2311.04145, 2023.

[101] Yifan Zhang, Deepa Patel, and Arjun Roy. Diffusion-driven promotional video generation for marketing. In *Proceedings of ACM Multimedia*, pages 1122–1131, 2023.

[102] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. Introduces the HDTF dataset.

[103] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022.

[104] Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. Storydiffusion: Consistent self-attention for long-range image and video generation. *NeurIPS*, 2024.

[105] Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. Celebv-hq: A large-scale video facial attributes dataset. In *European Conference on Computer Vision*, 2022.

[106] Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. CelebV-HQ: A large-scale video facial attributes dataset. In *ECCV*, 2022.