Video Generation Models: A Survey of Post-Training and Alignment

Chaoyu Li[1†✉] Xiaoyi Gu[2†] Yogesh Kulkarni[1] Eun Woo Im[1] Mohammadmahdi Honarmand[3] Zeyu Wang[4] Juntong Song[5] Fei Du[6] Xilin Jiang[7] Kexin Zheng[8] Tianzhi Li[9] Fei Tao[5] Pooyan Fazli[1✉]

[1]Arizona State University [2]Twitch [3]Stanford University [4]eBay [5]NewsBreak [6]Microsoft [7]Columbia University [8]University of Southern California [9]Carnegie Mellon University

† *Equal contribution* ✉ *Corresponding Author*

Abstract | Video generation has rapidly progressed from short, low-quality clips to high-resolution, long-duration sequences with complex spatiotemporal dynamics. Despite strong generative priors learned through large-scale pretraining, pretrained video models often fail to reliably follow human intent, maintain temporal coherence, or satisfy physical and safety constraints. Compared with image and text generation, alignment in video generation presents unique challenges, including error accumulation over time, motion-appearance coupling, multi-objective trade-offs, and limited supervision for temporal properties. These challenges motivate systematic post-training strategies that adapt pretrained models without retraining them from scratch. In this survey, we present the first comprehensive review of post-training and alignment in video generation models. We frame post-training as a unifying framework and distinguish between **implicit alignment** and **explicit alignment** based on how alignment signals are enforced. From this perspective, we organize existing approaches into four broad categories: (1) **supervised fine-tuning methods**, (2) **self-training and distillation methods**, (3) **preference- and reward-based methods**, and (4) **inference-time methods**. This taxonomy provides a coherent view of how alignment signals shape model behavior across both training and deployment. Beyond methodological advances, we review commonly used datasets, benchmarks, and evaluation practices, and discuss open challenges such as scalable reward design, long-horizon temporal consistency, stability-expressiveness trade-offs, and safety-aware generation. This survey aims to provide a structured conceptual foundation and practical guidance for advancing controllable and reliable video generation models.

Main Contact: chaoyuli@asu.edu, pooyan@asu.edu
Github: https://github.com/CyL97/Awesome-Video-Generation-Post-Training

1. Introduction

Video generation has advanced rapidly, evolving from low-quality, short clips to high-definition, minute-long sequences with increasingly complex dynamics [1, 2]. Despite this progress, generating realistic videos remains highly challenging. Models must preserve spatiotemporal coherence, ensure physical plausibility, and maintain fine visual details simultaneously. As a result, video generation stands among the most demanding problems in generative AI, requiring both strong generative priors and precise control mechanisms. The evolution of video generation models has followed several major trends. Early research mainly relies on generative adversarial network (GAN)-based methods and probabilistic generative models, focusing on unconditional generation or class-specific synthesis [3–5]. However, these approaches often suffer from limited diversity and training instability, making it difficult to model complex video distributions [6, 7]. With the growth of large-scale datasets and

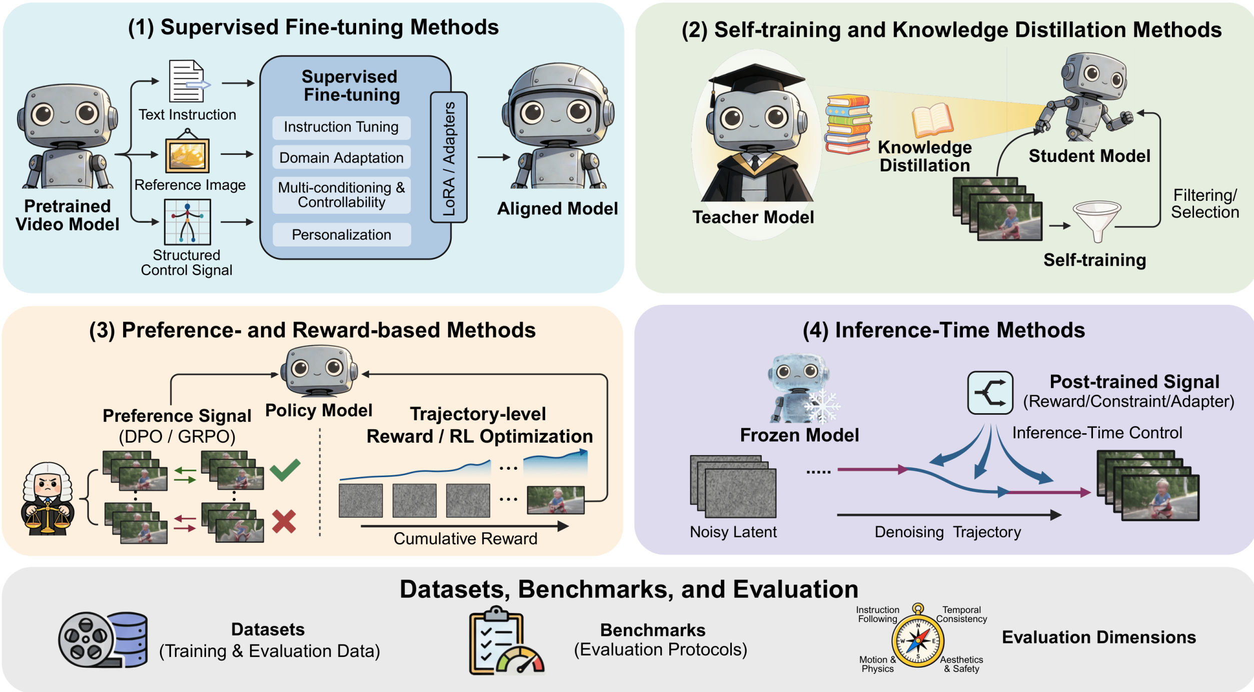

Figure 1 | An overview of the post-training and alignment in video generation models.

computational resources, research shifts toward foundation-style pre-training on massive video-text corpora. This paradigm enables models to learn general visual representations and align visual content with semantic descriptions, substantially improving generalization and text-conditioned generation performance [8, 9]. Among recent architectures, Diffusion Transformers (DiTs) [10] have emerged as the dominant paradigm [11–16]. By combining scalable transformer backbones with diffusion-based denoising, operating in compressed latent spaces, and incorporating multimodal conditioning, DiT-based models show strong scaling behavior and impressive generalization across diverse video generation tasks. Despite the powerful generative priors obtained through large-scale pre-training, such models do not inherently guarantee that generated videos adhere to user intent, physical constraints, or fine-grained control signals.

Aligning video generation models with desired behaviors poses fundamentally different challenges from those in image or text generation. In videos, small frame-level errors can accumulate over time and interact in complex ways, leading to artifacts that may not be evident in short clips or single-frame evaluations. Moreover, alignment objectives, such as motion realism, temporal coherence, identity consistency, and physical plausibility, span multiple dimensions and can conflict with one another, creating trade-offs between stability and expressiveness. Reliable supervision for temporal properties is also scarce and expensive, leading to reliance on proxy metrics or learned evaluators that may introduce bias. Together, these challenges call for alignment strategies specifically designed to handle temporal dynamics, multi-objective trade-offs, and limited supervision.

Motivated by these challenges, recent research has increasingly focused on post-training and alignment techniques that adapt pretrained models through additional optimization stages applied after large-scale pretraining. Figure 1 provides an overview of this landscape, highlighting major post-training paradigms and representative methods discussed in this survey. Rather than modifying core architectures or relying solely on scaling, these approaches refine model behavior through targeted post-training optimization [17, 18]. Across the literature, post-training techniques have expanded along multiple dimensions. As shown in Figure 2, research on post-training alignment for

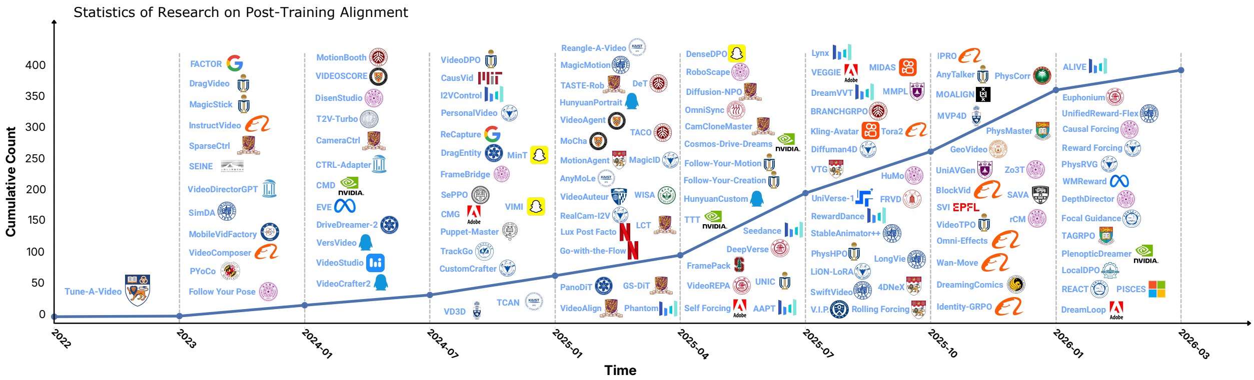

Figure 2 | Research trends in post-training and alignment for video generation models (2022–Feb. 2026).

video generation has grown rapidly since 2022, with expanding diversity in supervision paradigms and deployment strategies. One line of work focuses on supervised adaptation and parameter-efficient tuning, such as LoRA [19], to improve controllability, personalization, and domain transfer [20–22]. Another direction incorporates evaluative signals derived from human preferences or verifiable proxies to better align generation with semantic intent, physical plausibility, or safety constraints [23–25]. Meanwhile, some approaches leverage model-generated data and teacher supervision to enable iterative refinement and improve inference efficiency [26–28]. In parallel, a growing body of work explores how post-trained signals can be reused at deployment time to steer generation without further parameter updates [29–31]. Together, these directions mark a shift from purely scaling-driven improvements toward modular, signal-driven alignment strategies, forming a flexible toolbox for addressing the unique temporal and multi-objective challenges of video generation.

In this survey, **post-training** refers broadly to any optimization, adaptation, or control procedure applied after large-scale pretraining that modifies the behavior of a video generation model without retraining it from scratch. These methods operate on pretrained foundation models and aim to shape model behavior in downstream use. We use the term **alignment** to describe the extent to which a video generation model's behavior conforms to desired objectives at deployment. These objectives include accurately following human intent, maintaining temporal and identity consistency, respecting physical and causal constraints, and avoiding unsafe or undesirable outcomes. In this sense, alignment concerns **behavioral correctness** and reliability, rather than visual quality or data fit alone.

Within this post-training framework, we distinguish between **implicit alignment** and **explicit alignment** based on how alignment signals are applied. **Implicit alignment methods** shape model behavior indirectly. They rely on mechanisms such as supervised adaptation, model-generated or teacher-provided signals, or structured controllability mechanisms, without explicitly evaluating whether generated outputs satisfy alignment objectives. In contrast, **explicit alignment methods** directly optimize model behavior using evaluative signals that assess correctness. These signals may include preference feedback, reward functions, or verifiable criteria related to human intent, physical plausibility, or safety. Importantly, alignment in video generation exists on a **spectrum**: post-training methods differ in how directly, strongly, and reliably they influence aligned behavior, rather than forming a strict binary between aligned and non-aligned approaches. This distinction is orthogonal to the specific training or inference mechanisms employed and reflects *how* alignment is enforced rather than *when* or *where* optimization occurs.

Based on these definitions, we organize post-training and alignment methods for video generation into four broad categories according to the primary source and role of the signals used to shape model behavior. **(1) Supervised Fine-tuning Methods** primarily achieve implicit alignment by adapting pretrained models using labeled or structured supervision. **(2) Self-Training and Distillation Methods** also promote implicit alignment, leveraging model-generated data or teacher supervision to improve robustness, stability, or efficiency without explicitly evaluating correctness. **(3) Preference- and Reward-Based Methods** enable explicit alignment by optimizing model behavior with evaluative signals that assess correctness with respect to human intent, physical plausibility, or safety. **(4) Inference-Time Methods** influence alignment at deployment, either by enforcing explicit alignment through evaluative guidance or by supporting implicit alignment via iterative refinement and structured control. Together, these categories provide a unified and interpretable view of how post-training techniques shape alignment in video generation models across both training and inference. Figure 3 presents the overall taxonomy of post-training and alignment methods. In short, the key contributions of this survey are as follows:

Contributions

- **Post-Training Methods for Video Generation**. We provide a comprehensive review of post-training and alignment methodologies for video generation models, including supervised fine-tuning, preference- and reward-based optimization, self-training and distillation, and inference-time alignment and control techniques.
- **Taxonomy of Alignment Techniques**. We introduce a structured taxonomy that organizes post-training approaches according to their optimization mechanisms and alignment roles, highlighting adaptations to video-specific challenges such as temporal coherence, motion realism, and controllability.
- **Datasets and Benchmarks for Video Alignment**. We systematically summarize commonly used datasets, benchmarks, and evaluation protocols for post-training and alignment in video generation, categorizing them by alignment objectives and temporal characteristics.

Survey Structure

- **Section 2: Preliminaries.** Problem formulation of video generation, dominant base models, and multi-dimensional alignment objectives.
- **Section 3: Supervised Fine-tuning Methods.** Implicit alignment through supervised adaptation, including instruction tuning, domain specialization, controllability, personalization, and structured data pipelines.
- **Section 4: Self-training and Knowledge Distillation.** Implicit alignment via self-generated supervision and teacher-student distillation.
- **Section 5: Preference- and Reward-Based Methods.** Explicit alignment using reinforcement learning, preference optimization, and video reward modeling.
- **Section 6: Inference-Time Methods.** Hybrid alignment at deployment through guidance-based control and iterative refinement.
- **Section 7: Datasets, Benchmarks, and Evaluation Protocols.** Post-training datasets, evaluation benchmarks, and assessment protocols for alignment in video generation.
- **Section 8: Challenges and Future Directions.** Key open challenges and future directions for post-training and alignment in video generation models.

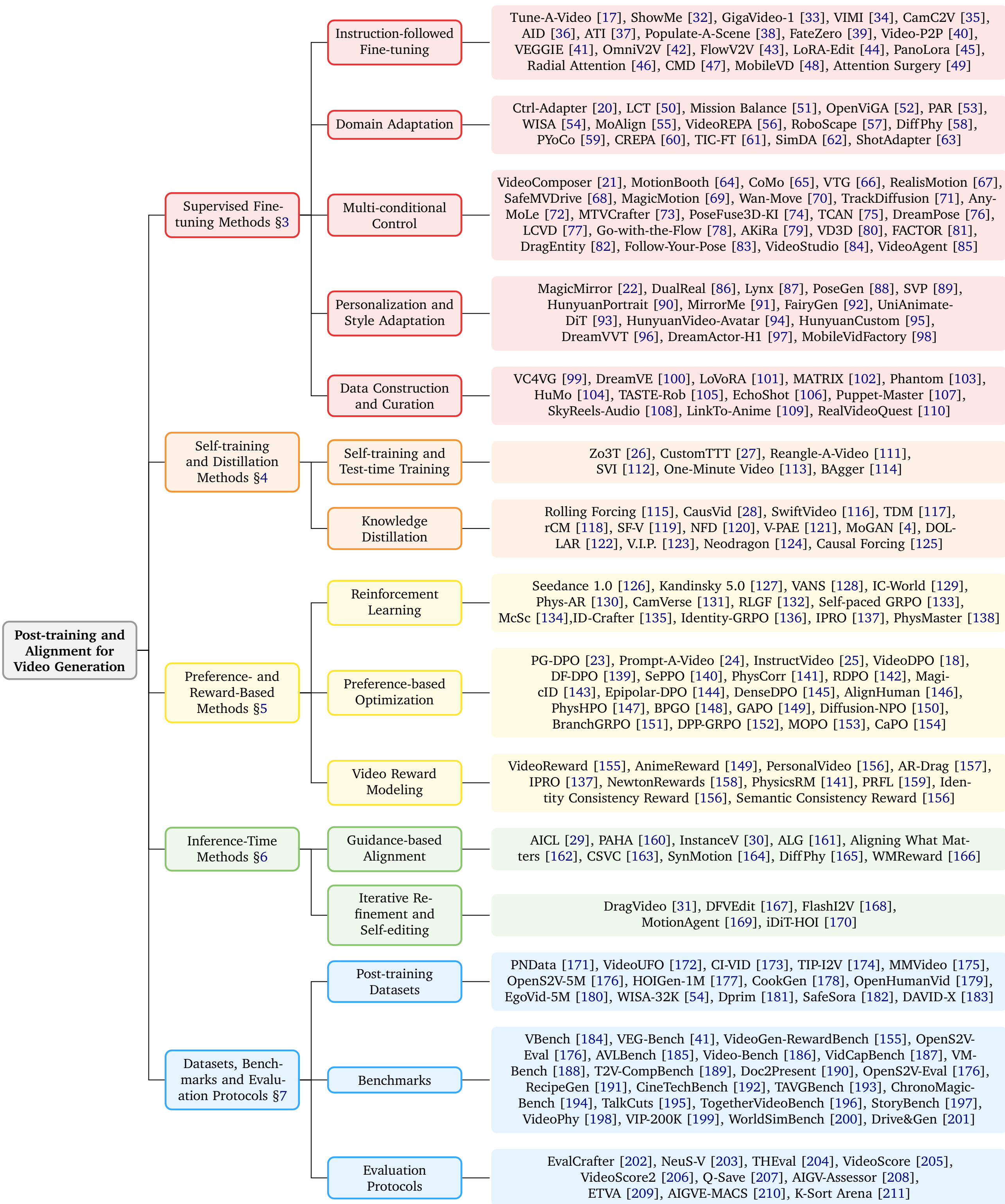

Figure 3 | Taxonomy of post-training and alignment in video generation models. Methods are grouped by alignment type: Supervised fine-tuning and self-training and distillation methods provide *implicit alignment*. Preference- and reward-based methods achieve *explicit alignment*. Inference-time methods function as *hybrid mechanisms*, supporting both implicit and explicit alignment. The bottom node lists representative post-training datasets, benchmarks, and evaluation protocols.

2. Preliminaries: Video Generation Models and Alignment Dimensions

Takeaways

- **Video Generation Paradigms**: Video generation is formalized as learning a conditional distribution under three primary settings: Text-to-Video (T2V), Image-to-Video (I2V), and Video-to-Video (V2V).
- **Dominant Base Models**: Latent Diffusion Transformers (DiTs) are identified as the foundation of modern video generation and post-training, characterized by spatio-temporal latent representations, attention-based conditioning, and diffusion- or flow-based objectives.
- **Alignment Objectives**: Alignment is characterized as a multi-objective problem beyond likelihood-based training, requiring satisfaction of instruction adherence, temporal coherence, motion realism, perceptual quality, and safety constraints.

This section introduces the foundational formulation of video generation models and the alignment objectives that guide post-training. We first formalize common video generation settings, including text-to-video, image-to-video, and video-to-video, and outline the dominant architectural paradigms behind modern systems. We then characterize alignment in video generation as a multi-objective problem spanning instruction adherence, temporal coherence, motion realism, and safety. These preliminaries provide the conceptual and technical basis for understanding how subsequent post-training methods shape model behavior.

2.1. Video Generation Problem Setting

Formally, video generation is modeled as learning a conditional probability distribution $p(\mathbf{v}|\mathbf{c})$, where $\mathbf{v}$ represents a video sequence, and $\mathbf{c}$ denotes conditioning signals such as text, images, or edit instructions [11, 212]. As illustrated in Figure 4, video generation tasks can be categorized into three paradigms based on input modalities and generation mechanisms.

(1) Text-to-Video (T2V). The goal of T2V is to synthesize a video $\mathbf{v}$ from a textual prompt $\mathbf{c}_{\text{text}}$ by sampling from a learned conditional distribution:

$$\mathbf{v} \sim p_\theta(\mathbf{v} \mid \mathbf{c}_{\text{text}}), \tag{1}$$

where θ denotes the parameters of the video generation model. This process corresponds to generation from scratch, requiring the model to produce both spatial content and temporal dynamics solely from the learned prior and textual conditioning [213, 214]. In practice, sampling typically begins with random noise $\mathbf{z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ in latent space, which is iteratively denoised using a spatio-temporal backbone, such as a 3D U-Net [215] or a Diffusion Transformer (DiT) [10]. The primary alignment challenges in T2V include semantic adherence to complex instructions and maintaining physical plausibility in open-domain, long-horizon video generation [130].

(2) Image-to-Video (I2V). I2V, conditions generation on both a text prompt $\mathbf{c}_{\text{text}}$ and a reference image $\mathbf{I}_{\text{ref}}$ (typically the first frame) [9, 216]. The objective is to generate temporal dynamics that extend the context of $\mathbf{I}_{\text{ref}}$ while preserving its identity and visual details.

$$\mathbf{v} \sim p_\theta(\mathbf{v}|\mathbf{c}_{\text{text}}, \mathbf{I}_{\text{ref}}) \quad \text{s.t.} \quad \mathbf{v}_0 \approx \mathbf{I}_{\text{ref}}. \tag{2}$$

This is achieved by conditioning spatio-temporal diffusion backbones on image representations (e.g., via cross-attention or feature injection), expanding the static image into a coherent temporal sequence. The alignment focus is on motion fidelity and preventing identity degradation over time [69, 90].

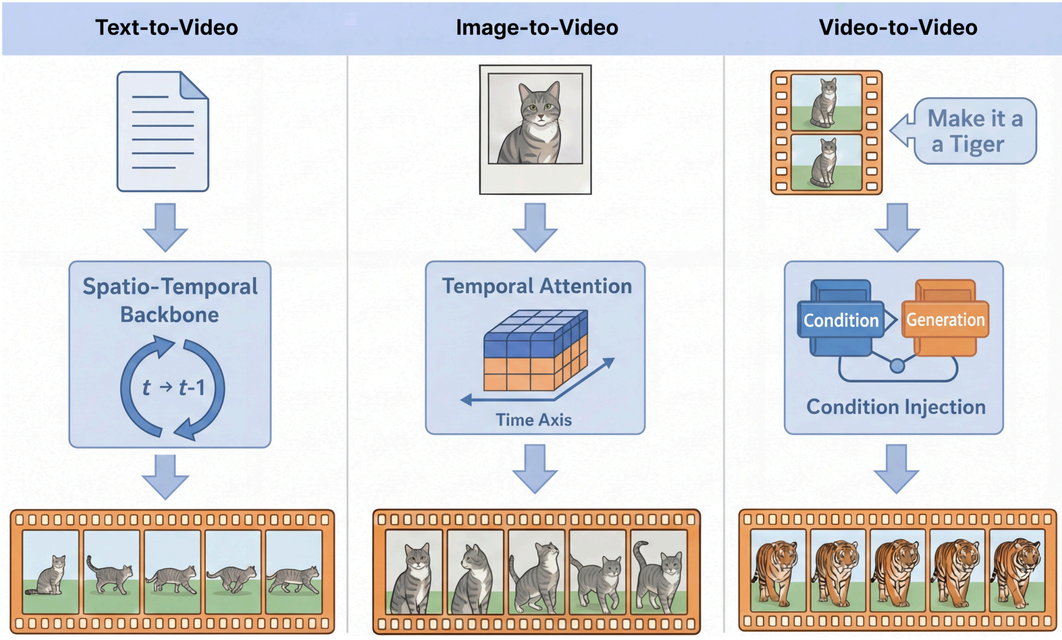

Figure 4 | Overview of video generation tasks. Left: Text-to-Video (T2V) models learn a video prior to map noise to pixels guided by text prompts. **Middle**: Image-to-Video (I2V) injects dynamics into a static image, typically freezing spatial layers and training temporal attention modules. **Right**: Video-to-Video (V2V) focuses on structure-preserving editing, injecting spatial guidance to align the generation with the source video's layout.

(3) Video-to-Video (V2V) and Editing. V2V aims to transform a source video $\mathbf{v}_{\text{src}}$ into a target video $\mathbf{v}_{\text{tgt}}$ according to an editing instruction $\mathbf{c}_{\text{edit}}$, while preserving its spatial-temporal layout (e.g., object motion, depth) [39, 40].

$$\mathbf{v}_{\text{tgt}} \sim p_\theta(\mathbf{v}|\mathbf{c}_{\text{edit}}, \mathbf{v}_{\text{src}}) \quad \text{s.t.} \quad \mathcal{S}(\mathbf{v}_{\text{tgt}}) \approx \mathcal{S}(\mathbf{v}_{\text{src}}), \tag{3}$$

where $\mathcal{S}(\cdot)$ represents structural features. Compared to text-to-video generation, V2V editing introduces an explicit structural consistency constraint that couples semantic modification with temporal coherence. The core challenge lies in balancing edit strength with structural preservation, as aggressive edits may disrupt motion dynamics, whereas conservative edits may fail to realize the intended transformation. To address this, existing methods typically employ **conditioning injection mechanisms** (e.g., ControlNet [217] or lightweight adapters [41]) or inversion-based guidance to anchor spatial layouts while modifying high-level semantics, prioritizing structural consistency and localized editability.

2.2. Base Models

While early work on video generation explores GAN-based architectures [6] and 3D U-Nets [215], these approaches have increasingly been replaced in large-scale settings by Transformer-based backbones [10]. This shift is driven by the superior scalability of Transformers and their ability to model complex, long-horizon spatio-temporal dependencies. As a result, contemporary post-training methods primarily focus on two modern paradigms: latent diffusion models with Transformer backbones and autoregressive video generation models.

Latent Diffusion Transformers (DiT) with Flow Matching. Figure 5 illustrates the dominant latent diffusion pipeline adopted by modern video generation models in 2024–2025 (e.g., Wan [218],

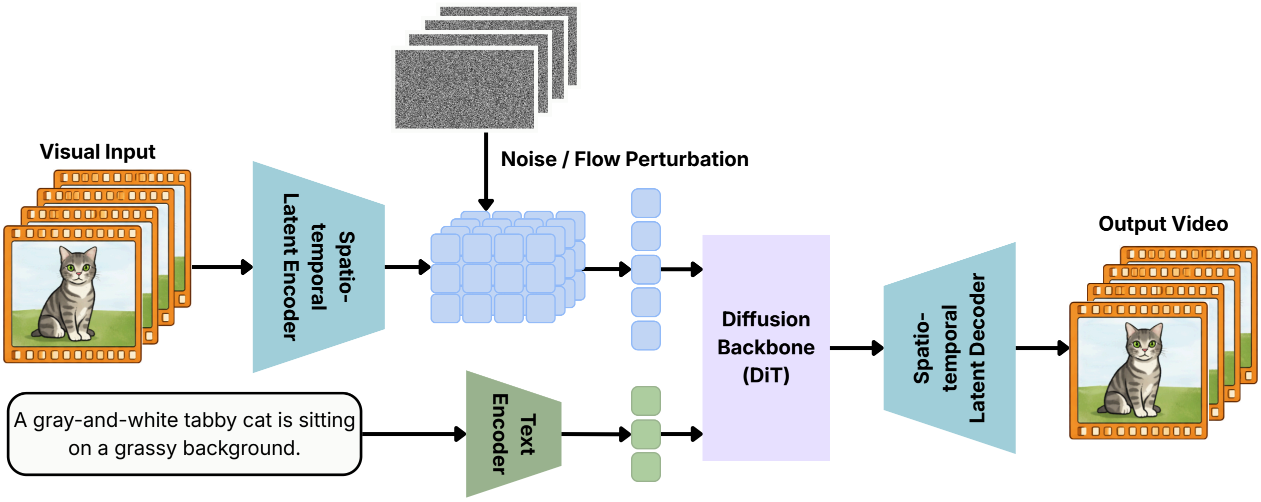

Figure 5 | Architecture of a modern Latent Diffusion Transformer (DiT) for video generation. The pipeline has three stages: (1) **Compression:** A spatio-temporal latent encoder compresses the input video into latent representations. (2) **Diffusion Modeling:** Stochastic perturbations are applied in latent space, and the resulting spatio-temporal tokens (blue tokens) are processed by a DiT backbone together with text embeddings (green tokens) to learn a denoising objective. (3) **Decoding:** The predicted clean latents are mapped back to pixel space by a spatio-temporal latent decoder.

HunyuanVideo [12], OpenSora [13]), which combines a spatio-temporal latent encoder, typically implemented as a 3D VAE, with a Diffusion Transformer backbone. Understanding its internal components is vital for effective alignment:

- **Spatio-temporal Latent Compression:** Videos are compressed into a latent space $\mathbf{z} = \mathcal{E}(\mathbf{v})$. Unlike frame-wise image VAEs, modern video encoders perform both spatial and temporal compression [219], significantly reducing sequence length but increasing the difficulty of fine-grained temporal control during post-training.
- **Spatio-temporal Patchification and Positional Encoding:** Latent tensors are flattened into spatio-temporal tokens and augmented with factorized or 3D Rotary Positional Embeddings (RoPE) [220], enabling variable temporal lengths and robust spatio-temporal generalization. Long-video post-training methods often require explicit handling of token positions and positional encodings.
- **Diffusion Transformer Backbone and Conditioning:** The denoising network is implemented as a ViT-style backbone, where conditioning signals (e.g., text prompts) are injected via cross-attention or Adaptive Layer Normalization (AdaLN). This module serves as the primary interface for post-training and alignment, with adapter-based methods (e.g., LoRA, ControlNet) attaching to attention blocks.
- **Pre-training Objective (Flow Matching):** Many modern models adopt Flow Matching, often instantiated as Rectified Flow [221], which learns a velocity field in latent space to map noise to data [222, 223]. This objective provides the foundation for subsequent post-training and preference-based alignment methods.

Autoregressive Video Generation. As a complementary paradigm to diffusion-based models, autoregressive approaches (e.g., VideoPoet [224], VideoMAR [225]) formulate video generation as a sequence modeling problem over discretized spatio-temporal tokens. Videos are first mapped into a token sequence, and generation proceeds in a causal manner via next-token prediction, analogous to large language models (LLM):

$$p_\theta(\mathbf{v}) = \prod_i p_\theta(z_i \mid z_{<i}, \mathbf{c}). \tag{4}$$

A key advantage of this formulation is that it enables the *direct application* of established LLM alignment algorithms (e.g., standard PPO [226] or DPO [227] on token logits) without the adaptations required for continuous diffusion processes [228]. However, given that the current open-source landscape and recent post-training advancements are predominantly centered on diffusion architectures, most alignment methods discussed in this survey focus on optimizing continuous diffusion trajectories rather than discrete token sequences.

2.3. Alignment Dimensions for Video Generation

Despite powerful architectures, pre-trained models optimize for *data likelihood* rather than *human utility*. They tend to reproduce the "average" web video, often containing motion blur, static scenes, or uncurated compositions. Post-training and alignment aim to bridge this gap. Unlike image generation, alignment in video generation must simultaneously ensure per-frame visual quality and coherent dynamics across time. Based on these challenges, we categorize the primary alignment objectives into four key dimensions.

(1) Instruction Following and Fine-grained Controllability. A central objective of alignment is ensuring that generated videos accurately reflect user intent across multiple modalities. Beyond basic text-semantic matching, this requires models to correctly interpret and execute complex instructions, including multi-step logic, compositional descriptions, and explicit constraints such as edits or exclusions (e.g., "remove the object" or "keep the background unchanged") [25, 83]. In practical scenarios, alignment must also support fine-grained controllability, where generation is conditioned on structured signals such as camera trajectories, depth maps, skeletal poses, or spatial layouts. These controls allow users to explicitly specify how scenes evolve and how actions are performed, rather than only describing what content should appear.

(2) Temporal Consistency and Identity Preservation. Beyond correctly interpreting user intent, a core challenge in video generation is maintaining coherence over time. Alignment methods in this dimension aim to reduce temporal artifacts caused by frame-level inconsistencies, such as flickering textures, unstable backgrounds, or unintended shape changes [229]. Beyond short-term stability, alignment must ensure that the identity of subjects remains consistent throughout a video [230, 231]. In long-form or personalized generation, characters or objects are expected to maintain the same appearance, including clothing, facial features, and overall visual style, even as they move, change viewpoint, or become partially occluded. When these requirements are not met, identity drift gradually accumulates, resulting in videos that appear unrealistic or inconsistent.

(3) Motion Quality and Physical Plausibility. While temporal consistency emphasizes stability over time, overly conservative generation can lead to static or lifeless videos. Pre-trained video generation models often favor static scenes or very small movements, since limited motion reduces the risk of visible errors during generation. Alignment in this dimension aims to encourage more expressive and dynamic motion that better reflects realistic actions and interactions [232, 233]. At the same time, the generated motion must obey basic physical rules [184, 234]. This includes respecting constraints such as gravity, collisions between objects, object permanence, and simple cause-and-effect relationships. Effective alignment helps prevent visible artifacts, for example, objects unrealistically disappearing, intersecting with each other, or behaving in ways that contradict the physical structure of the scene.

(4) Aesthetic Fidelity and Safety. Beyond motion and physical correctness, alignment must also address overall perceptual quality and responsible generation behavior. From an aesthetic perspective, alignment aims to produce videos with high visual clarity, stable composition, and minimal perceptual artifacts, such as motion blur or distorted body parts. It also enables models to match human aesthetic preferences, including consistent lighting, color tone, and recognizable artistic styles [12, 18, 95]. At the same time, alignment must ensure safe and reliable generation behavior. This includes reducing the production of harmful, biased, or NSFW content, as well as ensuring that models appropriately refuse unsafe requests or suppress undesirable concepts during generation [182].

3. Supervised Fine-tuning Methods

Takeaways

- Supervised fine-tuning is the primary mechanism for aligning pretrained video generation models with user intent, controllability requirements, and domain-specific constraints, without retraining or altering core model architectures.
- By integrating instruction tuning, domain adaptation, and multi-conditional supervision, supervised fine-tuning enables control over semantics, motion, camera behavior, and spatial layout beyond text-only guidance.
- Lightweight adaptation, combined with structured and synthetic data pipelines, supports identity preservation, personalization, and robust alignment under limited supervision.

Supervised fine-tuning adapts pretrained video generation models to specialized domains, better aligns them with user intent, and enables fine-grained control and personalization through targeted supervision signals. Supervised fine-tuning methods are typically categorized according to the form of supervision they employ and the alignment capabilities they provide.

3.1. Instruction and Prompt-following Fine-tuning

A primary goal of supervised fine-tuning is to improve a model's ability to follow user instructions. While large-scale pretraining provides video models with powerful generative priors, their responses to natural-language prompts often remain coarse, ambiguous, or inconsistent over time. Instruction and prompt-following fine-tuning addresses this gap by explicitly aligning textual directives, such as editing commands, compositional constraints, and multi-step instructions, with the corresponding video outputs, typically using relatively small, curated instruction datasets.

Direct Instruction-to-Video Supervision. Early efforts in instruction-following video generation focus on directly aligning user instructions with video outputs through explicit fine-tuning of pretrained generative models [213, 235, 236]. Tune-A-Video [237] studies one-shot text-to-video generation by adapting a pre-trained text-to-image diffusion model using only a single text-video pair instead of large-scale video datasets. Concretely, the spatial U-Net is extended with a sparse causal spatio-temporal attention mechanism and coupled with DDIM inversion for structure-guided sampling, allowing the image diffusion prior to be reused for coherent motion synthesis across video frames. Similarly, ShowMe [32] unifies instructional image and video generation by repurposing a pretrained video diffusion model as an action-object state transformer that handles both state manipulation (image editing) and state prediction (video rollout). A two-stage tuning strategy decouples and selectively activates spatial and temporal components with task-specific LoRA adapters, while structure-consistency and motion-consistency rewards are introduced to enhance spatial fidelity and temporal coherence in instruction-guided visual transformations.

Beyond strictly paired instruction-video supervision, more recent work explores scalable optimization strategies, ranging from joint image-video fine-tuning to inference-time adaptation that relaxes reliance on exhaustive video data while improving instruction adherence [238–240]. GigaVideo-1 [33] proposes an automatic dataset synthesis pipeline for video diffusion model fine-tuning that emphasizes physical and temporal consistency without relying on large-scale curated external datasets. The method leverages LLM-augmented prompt generation and a reward-guided optimization strategy, where feedback from a frozen multimodal large language model (MLLM) is used to adaptively reweight synthesized training samples during fine-tuning. VIMI [34] further extends instruction supervision to multimodal settings by introducing a multimodal instruction pretraining framework for grounded video generation. The framework constructs a large-scale multimodal prompt-video dataset via retrieval-augmented in-context examples and employs a two-stage pipeline of multimodal conditional video pretraining and multimodal instruction tuning, leveraging MLLMs to unify text-to-video, subject-driven video generation, and video prediction within a single model.

Image Guided Video Generation. While direct instruction-video supervision aligns generation with user intent, it often suffers from ambiguity and instability when synthesizing long or complex dynamics. Image-guided video generation mitigates this issue by introducing visual context as an additional grounding signal [35, 241]. AID [36] adapts an image-to-video diffusion model for instruction-guided video prediction by leveraging Stable Video Diffusion's dynamics priors and integrating multimodal textual control through an MLLM and a Dual Query Transformer to fuse visual and textual conditions. ATI [37] proposes a trajectory-based motion control framework that unifies local, object-level, and camera movements within a pre-trained image-to-video diffusion model, using a Gaussian-based motion injector to encode user-specified trajectories for fine-grained and continuous control. Beyond motion-centric conditioning, image guidance is further extended to support higher-level semantic and interaction-driven video generation, in which static visual context grounds instruction execution in complex scenes. Populate-A-Scene [38] repurposes a pre-trained text-to-video model as an affordance-aware human-world interaction simulator that conditions video generation on a scene image together with prompts describing a person's appearance and action.

Instruction-guided Video Editing. Instruction-guided video editing aims to modify existing video content according to user instructions while preserving temporal coherence and physical consistency, posing challenges beyond those in unconditional or image-based editing [39, 40]. VEGGIE [41] addresses this with an end-to-end framework that integrates video concept editing, grounding, and reasoning based on user instructions. The system employs an MLLM to interpret user intents into frame-specific queries and uses a curriculum learning strategy, along with a pipeline that transforms static image data into dynamic video-editing samples. Similarly, OmniV2V [42] explores a unified dynamic content manipulation module to integrate various scenario-based operations. It incorporates a LLaVA-based visual-text instruction module [242] to understand content correspondence and utilizes a multi-task data processing system to efficiently handle data overlap and augmentation. Beyond semantic and structural edits, maintaining physical plausibility during instruction-guided motion transfer remains a key challenge. FlowV2V [43] explicitly targets this issue by employing optical flow to model complex motion dynamics and mitigate failures caused by shape deformation. The approach combines first-frame editing with conditional generation by simulating a pseudo-flow sequence aligned with the deformed shape, enabling physically consistent video editing under user instructions.

Parameter-efficient and Efficiency-aware Instruction Fine-tuning. Beyond expanding the scope of instruction-aligned behaviors, a complementary line of supervised fine-tuning work explores efficiency from both the parameter and architectural perspectives to enable instruction- or task-specific specialization under the high computational costs of video generation models. On the parameter side, parameter-efficient fine-tuning strategies update only a small subset of model parameters while keeping the pretrained backbone frozen [44, 243–245]. PanoLora [45] exemplifies this direction by framing panoramic video generation as a specialization problem and proposing a LoRA-based fine-tuning strategy, supported by analysis showing that low-rank updates suffice to model the transformation. Beyond reducing the number of trainable parameters, several works further address the prohibitive computational overhead of video generation through architectural and attention-level optimizations [46, 47]. MobileVD [48] reduces memory usage by lowering frame resolution and applying channel-wise and temporal block pruning, and further compresses the denoising process into a single step via adversarial training. At the transformer level, Attention Surgery [49] introduces hybrid attention mechanisms guided by a cost-aware block-rate strategy that balances expressiveness and efficiency across layers based on the observation that different blocks exhibit varying reconstruction errors under different token sample ratios.

3.2. Domain Adaptation and Specialization

General-purpose video generation models, while powerful in open-domain settings, often encounter significant performance degradation when applied to specialized fields such as healthcare, industrial physics, or long-form storytelling. These failures typically arise from domain shifts, where the target data distribution differs significantly from the web-scale data used during pretraining. Unlike instruction and prompt-following fine-tuning, which primarily aligns models with user intent, domain adaptation focuses on aligning a pretrained model's internal representations and spatiotemporal priors with a new target distribution. Consequently, supervised fine-tuning for domain adaptation has shifted from simple fine-tuning toward approaches that emphasize domain-specific supervision, robustness to distribution shift, and efficient specialization.

Navigating Domain Shifts and Data Scarcity. We begin by characterizing the types of domain shifts that motivate specialization in video generation models, which often arise when foundation models fail to generalize to specialized visual distributions or operate reliably under data scarcity. Beyond appearance-level shifts, many specialization scenarios impose structural requirements that deviate substantially from web-scale training data [20]. A representative example is long-form storytelling, in which the target distribution requires scene-level coherence and long-range temporal consistency rather than short, loosely connected clips. LCT [50] addresses this shift by adapting a pretrained single-shot video generator to longer context windows through supervised long-context tuning, improving long-horizon consistency under extended generation settings.

In contrast, in high-stakes domains such as healthcare, the challenge is data scarcity rather than visual fidelity. Mission Balance [51] tackles the "long-tail" problem in medical imaging where rare pathological events are underrepresented. It introduces a two-stage fine-tuning approach that decouples spatial fidelity from temporal dynamics, allowing the model to synthesize high-fidelity surgical videos even with limited training examples. More broadly, domain shifts also arise in specialized content distributions such as automotive driving scenes [52] and panoramic video generation [45], as well as settings where models must better adhere to physical commonsense under distribution shift [53, 246].

Domain-Specific Supervision Signals. To effectively transfer models to specialized domains, researchers increasingly employ domain-specific supervision signals that go beyond generic text instructions. Such signals often encode task-relevant structure, physical constraints, or relational cues that are underrepresented in web-scale video-text data [54, 55]. In domains governed by physical laws, VideoREPA [56] improves the physical commonsense of text-to-video generation by aligning video models with relational and physics-relevant cues distilled from foundation models, providing an implicit yet domain-aligned supervision signal for physically plausible dynamics. Pushing beyond generic plausibility toward actionable interactions, RoboScape [57] introduces a physics-informed world model for embodied AI. Instead of relying solely on RGB pixel loss, it jointly learns video generation and auxiliary physics prediction tasks. This form of implicit physical supervision encourages the model to respect 3D geometry and object interactions, producing video simulations suitable for robotic policy training.

Robust Fine-Tuning and Catastrophic Forgetting. A central challenge in domain adaptation is catastrophic forgetting: naïvely fine-tuning a pretrained video generator on a narrow in-domain dataset can improve domain-specific fidelity while degrading general prompt-following behavior or disrupting learned spatiotemporal priors outside the adapted distribution. This phenomenon reflects a fundamental tension between specialization and retention in video generation models. Recent supervised fine-tuning methods, therefore, focus on retention-aware objectives that stabilize adaptation under distribution shift and data scarcity [58]. PYoCo [59] introduces a noise prior that preserves temporal correlations during fine-tuning, mitigating motion-structure collapse when adapting to new domains. CREPA [60] complements this direction by explicitly enforcing cross-frame representation consistency, reducing temporal degradation, and improving robustness during adaptation.

Beyond robustness, structured tuning can also improve transfer efficiency in low-data regimes. Acuaviva et al. [243] show that appropriately designed supervised fine-tuning can induce emergent few-shot generalization, which can reduce over-specialization to narrow in-domain cues. TIC-FT [244] further leverages temporally structured conditioning during training as a regularizer, promoting generalizable control and stable adaptation across diverse generation settings. Together, these works highlight that effective domain adaptation requires not only stronger in-domain supervision but also objectives that preserve the pretrained model's general spatiotemporal priors.

Efficient Adaptation for Specialization. Beyond robustness, practical deployment introduces additional constraints on scalability and efficiency. Domain adaptation often requires reusing a single pretrained model across multiple specialized domains, where full fine-tuning is both prohibitively expensive and prone to overfitting under limited in-domain data. In this context, parameter-efficient post-training methods become a practical necessity rather than a mere optimization choice. Adapter-based approaches such as SimDA [62] enable specialization by updating only a small fraction of parameters while preserving shared spatiotemporal priors. Similarly, LoRA-style adaptation has proven effective for small-data specialization, including in the work of Akarsu et al. [247], by acquiring domain-specific characteristics with minimal parameter updates. For long-form or multi-shot specialization, ShotAdapter [63] further demonstrates how lightweight adaptation can extend a pretrained single-shot generator to longer-horizon settings without incurring the cost of full retraining.

3.3. Multi-conditioning and Controllability

As video generation models are increasingly deployed in interactive and task-driven settings, aligning models with user intent through instruction fine-tuning alone often proves insufficient for precise and reliable control. Multi-conditioning and controllability methods address this limitation by explicitly training pretrained models to accept structured control signals. These signals expose controllable interfaces over motion, spatial layout, and temporal evolution during generation, enabling fine-grained and reliable manipulation beyond implicit instruction execution.

Motion and Camera Control. A prominent direction in controllable video generation focuses on explicit motion and camera control, where pretrained models are guided by structured temporal signals to regulate dynamics beyond their learned motion priors. Early efforts in this direction emphasize decoupling motion dynamics from visual appearance, enabling controllable motion manipulation while preserving content fidelity. For example, MotionBooth [64] introduces motion-aware customization by disentangling motion representations from appearance, allowing users to inject motion styles without compromising visual consistency. Complementarily, CoMo [65] formulates motion control as a compositional problem, decomposing complex motions into reusable primitives that can be flexibly recombined under textual guidance. These approaches establish a foundational principle for motion control: separating temporal dynamics from appearance facilitates fine-grained manipulation while maintaining identity stability.

Building upon this principle, a large body of work introduces explicit motion-related conditions to more directly regulate temporal evolution. One common strategy is to guide video generation using structured motion signals such as trajectories [66–73, 248–250], poses [74–77, 83, 251–259], or other temporally aligned control cues. By injecting these conditions into the temporal modeling components, such methods enable precise control over subject movement while preserving appearance-related content. VideoComposer [21], for instance, treats motion vectors as first-class temporal signals that explicitly guide inter-frame dynamics, supporting motion transfer and user-specified trajectories without retraining the base model.

In addition to subject motion, camera controllability has emerged as a critical aspect of video generation [35, 78, 79, 260–272]. Camera-aware methods explicitly model viewpoint evolution by conditioning diffusion transformers on camera paths or 3D camera parameters. VD3D [80] encodes per-frame camera poses as spatiotemporal embeddings, enabling accurate camera motion control while maintaining visual fidelity. Related approaches similarly leverage explicit camera trajectories to regulate viewpoint changes and scene geometry, allowing users to specify cinematic motion patterns with precise control. Across both motion- and camera-controlled generation, enforcing cross-frame consistency becomes central, as it directly impacts identity preservation, temporal smoothness, and geometric coherence.

Object-, Part-, and Spatial-level Control. Complementary to global motion control, another line of work focuses on achieving object- and spatial-level controllability by explicitly binding generation to specific entities, regions, or parts within a scene [31, 162, 170, 185, 273–284]. These approaches typically rely on structured spatial conditions such as masks, bounding boxes, instance tokens, or 3D proxies to preserve object identity and spatial consistency across frames while enabling localized manipulation. At the object level, FACTOR [81] introduces fine-grained control by conditioning video generation on entity-specific appearance and spatial context. By jointly encoding object descriptions, sparse bounding-box trajectories, and reference images, FACTOR enables localized manipulation of multiple objects while maintaining consistent identities and spatial layouts over time.

Beyond individual object binding, relational multi-entity control further models spatial dependencies among interacting entities. DragEntity [82] represents each object as a latent entity and explicitly incorporates relative spatial relationships when applying trajectory guidance. This entity-centric formulation enables simultaneous control of multiple objects while preserving structural integrity and reducing the distortions commonly observed in pixel-level dragging approaches. At an even finer granularity, part-level control targets the internal structure and articulation of objects. Puppet-Master [107] binds sparse drag signals to specific object parts through dedicated drag tokens, enabling fine-grained internal dynamics such as articulation and deformation while maintaining overall object identity and spatial coherence across frames.

Programmatic and Latent Control. Beyond direct conditioning, some approaches treat controllable video generation as the execution of an explicit plan or program derived from high-level instructions. In these methods, natural language prompts are first translated into structured intermediate representations, such as scripts, trajectories, or action graphs, which are then executed or iteratively refined by video generation models to support long-horizon consistency and interpretable control [285–289].

Within this paradigm, VideoStudio [84] casts video generation as a script-driven process by leveraging a large language model to convert an input prompt into a structured multi-scene program. The resulting script explicitly specifies scene-level events, entities, and camera movements, which are then executed by a diffusion model to generate each scene sequentially, enabling consistent content and coherent long-horizon video generation. While VideoStudio focuses on executing a fixed, LLM-generated program, VideoAgent [85] further extends this execution-centric perspective by treating generated videos as intermediate plans rather than final outputs. By iteratively refining and selecting video plans prior to execution, VideoAgent introduces an explicit plan selection and execution interface that separates high-level programmatic control from direct conditioning, thereby enabling controllability at the level of long-horizon behavior rather than frame-wise appearance.

3.4. Personalization and Style Adaptation

Personalization and style adaptation aim to customize video generation models to produce subject-consistent outputs that reflect specific identities, appearances, or stylistic preferences. Unlike generic controllability mechanisms that regulate motion or camera dynamics, personalization focuses on preserving identity fidelity across diverse motions, viewpoints, and conditioning signals, often under limited supervision or reference data. Recent advances explore lightweight post-training strategies that adapt pretrained video diffusion models to individual subjects, characters, or application-specific requirements, while maintaining the original model's generative capacity and temporal coherence. In this subsection, we review representative approaches from three complementary perspectives: identity-preserving personalization mechanisms, modality- and character-centric scenarios, and application-driven customization for humans and products.

Identity-Preserving Personalization via Lightweight Conditioning and Adapters. Several works explore supervised fine-tuning strategies to enhance identity fidelity in personalized video generation while minimizing disruption to motion dynamics and semantic alignment. Broadly, these approaches focus on either conditioning-based personalization or explicit disentanglement of identity and motion representations. MagicMirror [22] exemplifies the conditioning-based paradigm by introducing identity-preserving conditioning mechanisms within video diffusion transformers, enabling subject-specific generation without modifying the core architecture. In contrast, DualReal [86] addresses the identity–motion trade-off through adaptive training that jointly optimizes disentangled identity and motion representations, achieving faithful integration of appearance and dynamics. Related efforts further investigate identity preservation under sparse conditioning signals or in multi-character interaction scenarios [87, 88, 290].

Portrait-, Character-, and Multimodal-Centric Personalization. A related line of work focuses on portrait-, character-, and multimodal-centric video personalization, where preserving subject identity across pose variation, expression changes, and modality shifts is particularly critical. In the portrait domain, SVP [89] improves temporal stability by explicitly modeling long-range facial consistency, reducing identity drift over extended sequences, while HunyuanPortrait [90] leverages lightweight adapter layers to generate consistent portrait animation. To enhance robustness under partial occlusion and large head motion, facelet-based compensation [291] decomposes facial representations into localized components and adaptively corrects occluded regions during talking-head generation. MirrorMe [91] further explores audio-driven portrait animation, enabling identity-preserving facial motion synchronized with speech.

Beyond face-centric settings, personalization has been extended to character animation and multimodal generation. FairyGen [92] enables character-consistent video generation from drawn references, while UniAnimate-DiT [93] employs a large-scale video diffusion transformer to generate coherent human motion conditioned on reference images. In multimodal scenarios, HunyuanVideo-

Avatar [94] and HunyuanCustom [95] support subject-consistent generation driven by audio, images, and video inputs, facilitating expressive and controllable character animation across modalities.

Application-Driven Personalization for Humans and Products. Beyond generic identity customization, personalization in video generation is often driven by application-specific requirements involving humans and products. A prominent line of work focuses on human-product interaction scenarios. DreamVVT [96] targets realistic virtual try-on by introducing a stage-wise diffusion transformer framework that progressively aligns garment appearance with human motion and body structure under in-the-wild conditions. Similarly, DreamActor-H1 [97] addresses human–product demonstration videos by designing motion-aware diffusion transformers that generate high-fidelity interactions while preserving both human dynamics and product details.

In addition to interaction-centric applications, personalization has also been explored under deployment- and efficiency-driven constraints. MobileVidFactory [98] adapts diffusion-based video generation to mobile social media applications through a supervised pipeline optimized for computational efficiency and stylistic consistency, enabling automated personalized video creation from text prompts under resource-constrained settings.

3.5. Data Construction and Curation Pipelines

Recent progress in video generation is strongly driven by advanced data construction pipelines that actively shape model training. Moving beyond raw video-text pairs, modern approaches design structured supervision and scalable labeling mechanisms to introduce intermediate semantic representations and automatically generated signals. These pipelines bridge user intent, object dynamics, and temporal coherence, thereby facilitating controllable and robust video generation.

Structured Semantic Supervision. Several pipelines enrich training data with intermediate semantic representations that explicitly guide temporal modeling and compositional generation. DreamVE [100] and VC4VG [99] emphasize instruction-style and optimized textual supervision to encode editing intent and compositional constraints. Beyond textual supervision, several data construction pipelines automatically derive object- and interaction-aware signals that serve as structured supervision during training. LoVoRA [101] and MATRIX [102] construct temporally aligned object localization and mask tracks to provide cross-frame object-level supervision without manual annotation.

Similarly, Phantom [103] and Puppet-Master [107] build part-level motion and cross-modal alignment representations through automated parsing pipelines, enabling disentangled supervision over appearance, structure, and dynamics. Human-centric pipelines such as HuMo [104] and TASTE-Rob [105] further extract pose and hand–object interaction cues as intermediate supervision signals to support structured motion learning. In audio-driven scenarios, recent pipelines proposed by Zhang et al. [292], EchoShot [106], and SkyReels-Audio [108] incorporate fine-grained audio–visual alignment signals to enable temporally synchronized motion generation.

Synthetic Data and Scalable Labeling. To alleviate the high cost and limited scalability of dense video annotation, many pipelines adopt automated data generation and labeling strategies that function as implicit supervision mechanisms. LinkTo-Anime [109] demonstrates the effectiveness of synthetic data by generating accurate motion supervision through rendered optical flow, enabling precise temporal guidance without manual labeling. VideoScore [205] introduces a learned automatic feedback signal that approximates fine-grained human judgments and can be incorporated into training pipelines for scalable supervision and model selection. Related efforts also explore organizing supervision around realistic user intents rather than exhaustive frame-level annotations, reducing annotation overhead while preserving semantic alignment [110].

Table 1 | Summary of supervised fine-tuning methods for video generation.

Model	# Stages	Base Model	GPU Model	# GPUs	Venue	Year	Link
Tune-A-Video [237]	1	Stable Diffusion	A100	–	ICCV	2023	
VideoComposer [21]	2	Stable Diffusion	–	–	NeurIPS	2023	
DreamPose [76]	2	Stable Diffusion	A100	2	ICCV	2023	
PYoCo [59]	4	eDiff-I	–	–	ICCV	2023	–
SparseCtrl [293]	1	Stable Diffusion	–	–	ECCV	2024	
VideoDirectorGPT [294]	1	ModelScopeT2V	A6000	8	COLM	2024	
SimDA [62]	1	Stable Diffusion	A100	8	CVPR	2024	
MotionBooth [64]	1	Zeroscope LaVie	A100	1	NeurIPS	2024	
CMD [47]	2	Stable Diffusion	A100	1	ICLR	2024	–
Follow-Your-Pose [83]	2	Stable Diffusion	A100	8	AAAI	2024	
VD3D [80]	1	SnapVideo	A100 (40G)	64	ICLR	2024	
VideoStudio [84]	2	Stable Diffusion	A100	64	ECCV	2024	
DriveDreamer-2 [295]	2	Stable Diffusion	A800	8	AAAI	2025	
FlipSketch [296]	1	ModelScope	–	–	CVPR	2025	
MinT [297]	1	OpenSora	A100	–	CVPR	2025	–
CTRL-Adapter [20]	1	I2VGen-XL Stable Video Diffusion Latte Hotshot-XL	A100	–	ICLR	2025	
I2VControl [298]	1	MagicVideo-V2	–	–	ICCV	2025	–
CustomCrafter [299]	1	VideoCrafter2	A100	4	AAAI	2025	
TrackGo [279]	1	Stable Video Diffusion	A100	8	AAAI	2025	–
Puppet-Master [107]	1	Stable Video Diffusion	A6000	1	ICCV	2025	
ReCapture [300]	2	Stable Video Diffusion	A100	1	CVPR	2025	–
MagicStick [301]	1	Stable Diffusion	RTX 3090Ti	1	WACV	2025	
FACTOR [81]	2	Phenaki	–	–	WACV	2025	–
EDG [302]	3	DynamiCrafter	A100	8	CVPR	2025	
Go-with-the-Flow [78]	1	Stable Diffusion CogVideoX	A100	8	CVPR	2025	
GS-DiT [303]	2	CogVideoX	A100	8	CVPR	2025	
FramePack [286]	1	HunyuanVideo	A100	8	NeurIPS	2025	
HunyuanPortrait [90]	1	Stable Video Diffusion	A100	128	CVPR	2025	
LCT [50]	2	MMDiT	H800	128	ICCV	2025	–
TASTE-Rob [105]	3	DynamiCrafter	A6000	1	CVPR	2025	
Phantom [103]	2	MMDiT	A100	–	ICCV	2025	
RealCam-I2V [265]	1	DynamiCrafter	–	–	ICCV	2025	
VideoREPA [56]	1	CogVideoX	A100	8	NeurIPS	2025	
WISA [54]	1	CogVideoX	A100	8	NeurIPS	2025	
DiffPhy [58]	1	Wan2.1	H100	4	ICLR	2025	
MoAlign [55]	2	CogVideoX	H100	4	ICLR	2025	–
HunyuanVideo-Avatar [94]	2	HunyuanVideo	A100 (96G)	160	arXiv	2025	
TIC-FT [244]	1	CogVideoX Wan2.1	H100	1	NeurIPS	2025	
RoboScape [57]	1	–	A800	32	NeurIPS	2025	
EchoShot [106]	1	Wan2.1	A100	–	NeurIPS	2025	
LTD [304]	1	Wan2.1	H20	8	ICASSP	2026	–
ALIVE [305]	6	Waver1.0	–	–	arXiv	2026	

4. Self-training and Knowledge Distillation Methods

Takeaways

- Self-training and test-time training reuse self-generated outputs or intermediate representations as supervision, enabling iterative refinement and inference-time adaptation under limited or no external annotations, with particular benefits for temporal consistency, controllability, and long-context video generation.
- Knowledge distillation transfers capabilities from large or computationally expensive teacher models to more efficient students by encouraging consistency between teacher and student generation behaviors, thereby reducing inference cost while maintaining video quality and temporal coherence.

Self-training and knowledge distillation improve video generation models by reusing model-generated signals or transferring knowledge from stronger teachers. By treating the model or a teacher as the source of supervision rather than relying on costly external annotations, these approaches are particularly attractive for video generation, where dense temporal labels are scarce. Specifically, self-training and test-time training enable iterative refinement or online adaptation during inference. Knowledge distillation instead focuses on transferring capabilities from large or computationally expensive teacher models to more efficient student models, thereby improving inference speed and deployment efficiency. Together, these techniques complement earlier post-training paradigms by enabling scalable improvement and practical deployment in modern video generation pipelines.

4.1. Self-training and Test-time Training

Self-training and test-time training optimize video generation models by leveraging the model's own outputs or intermediate representations as supervision. By using self-generated signals rather than externally curated labels, these methods enable iterative refinement or online adaptation during inference. By leveraging feedback implicit in the generation process itself, this paradigm supports continual improvement under limited external supervision.

Inference-time Parameter Adaptation. A prominent line of work applies test-time training by adapting a small set of model parameters during inference to improve temporal consistency or task-specific performance. Zhang et al. [26] propose Zo3T, where a LoRA module is optimized at inference time to align intermediate feature representations across frames, enforcing trajectory-level consistency through feature-space alignment. CustomTTT [27] extends this paradigm to customized video generation by decoupling appearance and action modeling. Separate LoRA adapters are trained for figure and action, and merged at inference time via self-supervised distillation to mitigate conflicts introduced by joint optimization. Similarly, Jeong et al. [111] adapt LoRA parameters on the input video using pseudo-labels derived from a self-supervised formulation, enabling test-time fine-tuning for video viewpoint transformation. Together, these methods demonstrate that lightweight, inference-time parameter adaptation provides a flexible mechanism for improving controllability and consistency without requiring offline retraining.

Inference-time Temporal Modeling via Test-Time Training. Beyond parameter-efficient adaptation, test-time training has also been integrated directly into temporal modeling architectures to address long-context video generation. Dalal et al. [113] propose a hybrid architecture in which test-time training (TTT) layers are embedded as recurrent modules within the model's temporal computation. By performing self-supervised reconstruction of low-rank token representations at inference time, these TTT layers capture long-range dependencies without relying on global self-attention or extended attention windows. This formulation reframes test-time training as an integral component of temporal modeling rather than a post-hoc adaptation strategy, enabling efficient long-context video generation without offline fine-tuning.

Offline Self-training with Self-generated Supervision. Complementary to test-time adaptation, self-training methods improve video generation models through offline optimization on self-generated signals. VideoAgent [85] adopts a rejection-sampling-based self-training framework for embodied control, where successful trajectories collected during real-world robot execution are reused as training data to iteratively refine the video generation model. SVI [112] addresses error accumulation in long video generation through iterative error recycling. The method estimates generation errors by approximating diffusion trajectories with one-step integration, injects these errors into subsequent training inputs, and explicitly trains the model to correct its accumulated deviations. These approaches illustrate how self-training enables continual improvement under limited supervision by transforming model failures or successes into learning signals.

4.2. Knowledge Distillation

Knowledge distillation has been widely adopted in video generation as an effective approach for accelerating inference, transferring or consolidating model capabilities, and improving overall generation quality. Existing work can be broadly categorized by the form of supervision and the role of the teacher model, with diffusion-based video generation as the primary focus.

Diffusion-to-Autoregressive Distillation. A major line of research distills slow but expressive diffusion-based teacher models into fast autoregressive student models capable of long-horizon video generation. Representative works [28, 114, 115, 306] typically adopt Distribution Matching Distillation (DMD), which aligns the output distributions of teacher and student models to enable efficient generation while preserving visual fidelity. By compressing iterative diffusion sampling into a small number of steps, these methods substantially reduce inference latency without retraining models from scratch.

Trajectory-Level and Continuous-Time Distillation. Unlike distillation methods that match distributions only at the final state, a line of work provides denser supervision by aligning diffusion trajectories over time. SwiftVideo [116] introduces Continuous-Time Consistency Distillation (CCD) based on a flow-matching formulation, directly aligning the velocity fields predicted by the teacher and student models at each timestep. In this formulation, the teacher velocity serves as the primary supervision signal, enabling strong temporal consistency under few-step sampling. In parallel, Luo et al. [117] extend distribution matching from single-step alignment to trajectory-level consistency by enforcing alignment at multiple intermediate diffusion states, which reduces error accumulation and stabilizes long-horizon generation. rCM [118] also targets continuous-time modeling but adopts a different supervision role assignment, treating student self-consistency as the primary objective and introducing teacher score (velocity) distillation as an auxiliary regularizer rather than a timestep-wise regression target, thereby preserving generation diversity while mitigating error accumulation and detail degradation in few-step regimes.

Adversarial and Hybrid Distillation Objectives. Beyond distribution matching and consistency-based objectives, another line of work augments diffusion model distillation with adversarial learning to improve generation quality and training stability. In this paradigm, a discriminator provides additional supervision by distinguishing between teacher- and student-generated predictions or representations, sometimes reusing pretrained teacher components within the discriminator architecture. SF-V [119] fine-tunes a student initialized from the teacher model and employs a discriminator built upon a frozen teacher encoder with trainable spatial and temporal heads to assess generation quality.

Similarly, NFD [120] introduces adversarial supervision after a score-consistency-based warm-up phase, using a discriminator initialized from the teacher model. More recent approaches further combine adversarial learning with distribution matching or consistency objectives to mitigate error accumulation and distribution mismatch. For example, works by Cheng et al. [121] and Xue et al. [4] integrate GAN-based supervision into DMD, with different emphases on overall distributional quality and motion dynamics. DOLLAR [122] adopts a hybrid formulation that combines distribution matching, consistency-based distillation, and latent reward optimization to alleviate mode collapse and fidelity degradation in few-step video generation.

Training Strategies for Distillation. Beyond architectural and objective-level innovations, training data organization and distillation-related strategies also play an important role in scalable video generation. Seedance 1.0 [126] adopts a progressive training scheme that gradually increases video resolution and temporal complexity, facilitating more stable optimization. Other works modify the distillation process itself; for example,self-forcing methods mitigate exposure bias by conditioning training on self-generated histories [115, 123]. In addition, text encoder distillation has emerged as a complementary technique that significantly reduces model size and inference cost while preserving semantic alignment [124]. Collectively, these studies highlight the importance of coordinated distillation objectives, data curricula, and auxiliary supervision in efficient video generation.

Table 2 | Summary of self-training and knowledge distillation methods for video generation.

Model	# Stages	Base Model	GPU Model	# GPUs	Venue	Year	Link
SF-V [119]	1	SVD	A100	8	NeurIPS	2024	
CausVid [28]	2	Wan2.1	–	64	CVPR	2025	
CustomTTT [27]	3	CogVideoX	A6000	1	AAAI	2025	
DOLLAR [122]	3	DiT OpenSora LDM	A100	8	ICCV	2025	
TDM [117]	1	Stable Diffusion	–	–	arXiv	2025	
Reangle-A-Video [111]	2	CogVideoX	–	–	ICCV	2025	
One-Minute Video [113]	1	CogVideoX	H100	256	CVPR	2025	
Self Forcing [307]	1	Wan2.1	H100	64	NeurIPS	2025	–
NFD [120]	3	–	A100	–	arXiv	2025	–
ADM [308]	1	CogVideoX SDXL SD3	–	–	ICCV	2025	–
V.I.P. [123]	1	VideoCrafter2 AnimateDiff	A100	4	ICCV	2025	–
SwiftVideo [116]	3	Wan2.1	A100	8	arXiv	2025	–
V-PAE [121]	2	Wan2.1	H20	32	arXiv	2025	–
Zo3T [26]	–	Stable Video Diffusion	A100	1	arXiv	2025	–
SVI [112]	1	Wan2.1	–	–	arXiv	2025	
Neodragon [124]	4	Pyramidal Flow DiT	H100	–	arXiv	2025	
VideoTPO [309]	–	Wan2.1 Kling	–	–	arXiv	2025	
MoGAN [4]	2	Wan2.1	H200	16	arXiv	2025	–
Causal Forcing [125]	3	Wan2.1	H100	–	arXiv	2026	
EchoTorrent [310]	4	InfiniteTalk	A100	64	arXiv	2026	–
AMD [311]	2	Wan2.1	H800	8	arXiv	2026	–

5. Preference- and Reward-based Methods

Takeaways

- Reinforcement learning aligns video generation by explicitly modeling generation as a long-horizon decision process, enabling the enforcement of temporal consistency, physical plausibility, and structured constraints through trajectory-level optimization.
- Preference-based optimization directly optimizes relative comparisons between generated videos, with recent advances introducing temporally structured, physically grounded, and stability-aware preference objectives tailored to diffusion-based video models.
- Video reward modeling underpins both reinforcement learning and preference-based alignment by decomposing human preferences into multi-dimensional, identity-aware, and physics- or reasoning-aware signals that capture video-specific quality beyond frame-level appearance.

Preference- and reward-based methods align video generation models by leveraging evaluative signals, such as preference comparisons, learned reward models, or verifiable outcome-based criteria, rather than relying solely on fixed supervised targets. These signals enable models to better align with human intent, temporal coherence, physical plausibility, and safety constraints.

5.1. Preliminaries: Optimization Paradigms

Three optimization paradigms are representative in preference-based and reinforcement learning for video generation models: Proximal Policy Optimization (PPO), Direct Preference Optimization (DPO), and Group Relative Policy Optimization (GRPO). These paradigms form the methodological basis for the alignment methods reviewed in this section. Although they originate from language modeling and image generation, we present them in a unified formulation applicable to both autoregressive and diffusion-based video generation models.

We use x to denote the multimodal conditioning signal (e.g., text, images, or control inputs), y to denote a generated video or its latent representation, and τ to denote a generation trajectory. Depending on the model, τ may correspond to a sequence of discrete tokens or a continuous diffusion trajectory over denoising timesteps. For diffusion-based generators, we interpret each denoising step as an action and define the trajectory likelihood $\log \pi_\theta(\tau \mid x)$ as the sum of step-wise conditional log-densities (or a training-time surrogate), since the marginal likelihood of the final generated video is generally intractable.

PPO-style Reinforcement Learning (RLHF and RLAIF). Reinforcement Learning with Human Feedback (RLHF) [312] aligns a generative policy by first training a reward model (RM) and then optimizing the policy using PPO [226] under a constraint that limits deviation from a reference model π_{ref} (e.g., an SFT or pretrained model). The reward model is typically trained on preference pairs (x, y^+, y^-) using a Bradley–Terry objective [313],

$$\mathcal{L}_{\text{RM}}(\phi) = -\mathbb{E}_{(x,y^+,y^-)} \log \sigma\big(r_\phi(x, y^+) - r_\phi(x, y^-)\big) , \tag{5}$$

where $r_\phi(x, y)$ denotes a scalar reward and $\sigma(\cdot)$ is the logistic function. Given a fixed reward model, PPO optimizes the policy by maximizing a clipped policy-gradient objective augmented with a KL regularization term relative to the reference policy. Let $r_t(\theta) = \frac{\pi_\theta(y_t \mid x, y_{<t})}{\pi_{\theta_{\text{old}}}(y_t \mid x, y_{<t})}$ denote the probability ratio and $\hat{A}_t$ an advantage estimator, commonly implemented by broadcasting a sequence-level reward

across individual timesteps. The PPO objective is

$$\mathcal{L}_{\mathrm{PPO}}(\theta) = -\mathbb{E}\left[\sum_t \min\Big(r_t(\theta)\,\hat{A}_t,\ \mathrm{clip}\big(r_t(\theta),\ 1-\epsilon,\ 1+\epsilon\big)\,\hat{A}_t\Big)\right] \ +\ \beta\,\mathrm{KL}(\pi_\theta(\cdot|x)\,\|\,\pi_{\mathrm{ref}}(\cdot|x))\,. \qquad (6)$$

Reinforcement Learning with AI Feedback (RLAIF) [314] follows the same optimization procedure but replaces human annotations with AI-generated rewards or preferences. Although PPO-style reinforcement learning provides a principled framework for trajectory-level credit assignment, explicit RLHF or RLAIF is relatively uncommon in video generation due to the difficulty of designing stable and dense reward signals over high-dimensional, long-horizon video trajectories.

Direct Preference Optimization (DPO). Direct Preference Optimization (DPO) [227] eliminates the need for an explicit reward model by directly optimizing the policy to match observed preferences relative to a fixed reference policy. Given preference pairs (x, y^+, y^-) and a temperature parameter $\beta > 0$, the DPO objective is

$$\mathcal{L}_{\mathrm{DPO}}(\theta) = -\,\mathbb{E}\,\log\sigma\Big(\beta\big[\log\pi_\theta(y^+|x) - \log\pi_{\mathrm{ref}}(y^+|x) - \log\pi_\theta(y^-|x) + \log\pi_{\mathrm{ref}}(y^-|x)\big]\Big). \qquad (7)$$

This formulation can be interpreted as implicitly inducing a reward proportional to the log-probability ratio between the policy and the reference model, thereby combining preference alignment and KL regularization into a single contrastive objective. DPO-style optimization has proven particularly attractive for video generation models, where training high-quality reward models is challenging, and preference supervision can be applied at varying temporal granularities.

Group Relative Policy Optimization (GRPO). Group Relative Policy Optimization (GRPO) [315] provides an alternative alignment paradigm that replaces learned rewards or explicit preference pairs with verifiable outcome-level signals. For a given conditioning input x, GRPO samples a group of K trajectories $\{\tau^{(k)}\}_{k=1}^K$ from the current policy $\pi_{\theta_{\mathrm{old}}}$ and evaluates each trajectory using a verifiable scoring function $r^{(k)} \in [0,1]$, such as correctness checks, temporal consistency criteria, or task-specific rules. A group baseline $\bar{r} = \frac{1}{K}\sum_{j=1}^K r^{(j)}$ is computed, and group-relative advantages are defined as

$$A^{(k)} = r^{(k)} - \mathrm{stopgrad}(\bar{r}), \qquad \ell^{(k)}(\theta) = \sum_{t\in\tau^{(k)}} \log\pi_\theta(y_t\,|\,x, y_{<t}). \qquad (8)$$

The GRPO objective is then

$$\mathcal{L}_{\mathrm{GRPO}}(\theta) = -\frac{1}{K}\sum_{k=1}^K A^{(k)}\,\ell^{(k)}(\theta) \ +\ \beta\,\mathrm{KL}(\pi_\theta(\cdot|x)\,\|\,\pi_{\mathrm{ref}}(\cdot|x))\,. \qquad (9)$$

By relying on relative comparisons within a sampled group, GRPO avoids explicit reward modeling and reduces sensitivity to absolute score calibration. This property is particularly appealing for video generation, where designing reliable scalar rewards is difficult, but outcome-level verification or heuristic constraints are often available.

5.2. Reinforcement Learning for Video Generation

Reinforcement learning (RL) aligns video generation models by treating generation as a sequential decision-making process optimized under long-horizon video-level objectives. Compared to supervised post-training and preference-based optimization, RL explicitly models the interaction between generation actions and delayed rewards, making it well-suited to enforcing temporal consistency, physical plausibility, and structured constraints. Existing approaches apply RL at different levels of the generation pipeline, ranging from system-level alignment to diffusion-level optimization and constraint-aware training strategies.

Reinforcement Learning as an End-to-End Alignment Framework. Several works apply RL as an end-to-end alignment framework at the system level, treating the entire video generation pipeline as a policy optimized with respect to long-horizon video-level objectives [127, 155, 316, 317]. In this setting, RL serves as an end-to-end post-training mechanism that directly aligns model behavior beyond supervised fine-tuning. VANS [128] exemplifies this paradigm by applying RL to jointly align a vision-language model and a video diffusion model for video next-event prediction. Through a unified Joint-GRPO strategy, VANS optimizes both components under a shared reward, enabling system-level coordination between semantic reasoning and video generation. Similarly, Seedance 1.0 [126] incorporates video-specific RL from human feedback as a system-level alignment component, directly maximizing multi-dimensional reward signals to jointly improve prompt adherence, motion plausibility, and visual fidelity in large-scale video generation. From a more explicit decision-making perspective, RLIR [318] formulates video generation as a sequential policy optimization problem by recovering verifiable reward signals from generated videos, demonstrating how RL can be cleanly applied when suitable action-reward representations are available.

Reinforcement Learning for Optimizing the Generation Process. Beyond end-to-end alignment, another line of work applies RL directly to the video generation process itself, intervening at the level of diffusion sampling and generation trajectories [132, 157]. Instead of treating RL solely as a high-level post-training objective, these methods integrate RL signals into intermediate stages of generation, enabling fine-grained control over temporal dynamics and physically grounded motion. Phys-AR [130] reformulates diffusion-based video generation as a token-level sequential decision process by introducing diffusion timestep tokens that explicitly represent evolving physical states. RL is applied to optimize reasoning trajectories under rule-based physical rewards, enabling the model to enforce motion consistency and generalize to out-of-distribution physical conditions beyond data-driven interpolation. Complementarily, CamVerse [131] treats the video diffusion model as a stochastic policy and applies online RL to optimize camera-controlled video generation. By designing a verifiable geometric reward that provides dense, segment-level feedback on camera-trajectory alignment, CamVerse directly guides the generation process toward geometrically consistent and controllable camera motion.

Reinforcement Learning for Stability, Efficiency, and Structured Constraints. Applying RL to video generation poses challenges in training stability, controllability, and enforcing task-specific constraints. Recent work extends RL beyond generic policy optimization through curriculum-style training and structured objectives that improve robustness [135, 137]. To improve optimization stability, Self-Paced GRPO [133] proposes a competence-aware RL framework in which reward supervision co-evolves with the generator. By progressively shifting the reward function's emphasis from coarse visual quality to temporal coherence and semantic alignment, self-paced GRPO mitigates reward saturation and stabilizes long-horizon policy optimization. Beyond training dynamics, RL has also been used to impose structured constraints. PhysMaster [138] adopts a top-down strategy that optimizes a physics-aware representation as an explicit conditioning signal, while Identity-GRPO [136] enforces identity consistency through specialized rewards and GRPO optimization. Together, these approaches illustrate how RL can be adapted to address stability concerns and enforce structured objectives in video generation, extending its role beyond generic alignment.

5.3. Preference-based Optimization for Video Alignment

Preference-based optimization aligns video generation models by directly optimizing relative-preference objectives over generated samples. Unlike reinforcement learning, which requires complex trajectory-level credit assignment, these methods rely on simpler pairwise or relative comparisons to shape model behavior, making them particularly suitable for high-dimensional video diffusion models.

Recent work adapts DPO and its variants to the video domain by designing scalable mechanisms for constructing reliable preference signals under limited or fully automated supervision.

Preference-based Optimization as a Direct Alignment Objective. A growing line of work formulates video alignment as direct optimization over preference signals, avoiding explicit reward modeling and policy-based reinforcement learning [23–25]. VideoDPO [18] pioneers the adaptation of Direct Preference Optimization to video diffusion models by introducing OmniScore, a multi-dimensional scoring function that jointly evaluates visual quality and text–video semantic alignment, and automatically constructs preference pairs by ranking multiple generated videos per prompt. While VideoDPO derives preferences from score-induced comparisons among generated samples, DF-DPO [139] addresses the cost and ambiguity of such comparisons by using real videos as winning samples and their edited counterparts with explicit temporal or spatial artifacts as losing samples, providing unambiguous and scalable preference supervision. Beyond score- or artifact-based preference construction, SePPO [140] further extends preference-based optimization through a semi-policy framework that constructs preference pairs using historical model checkpoints as reference policies, enabling reward-model-free alignment while stabilizing training via an anchor-based adaptive flipper.

Fine-Grained and Structured Preference Supervision. Beyond video-level binary preferences, effective preference-based alignment for video generation requires structuring preference signals across finer temporal and semantic dimensions [141–144, 319, 320]. DenseDPO [145] addresses the motion bias of vanilla DPO by constructing structurally aligned video pairs from partially noised real videos and collecting segment-level preference labels, enabling dense temporal supervision that localizes artifacts while preserving global motion dynamics. Similarly, AlignHuman [146] exploits the observation that different denoising timesteps control distinct generation attributes and proposes timestep-segment preference optimization to decouple motion and fidelity alignment. Preference data are partitioned across denoising intervals, with specialized LoRA experts activated within their corresponding time-step ranges to target motion naturalness and visual fidelity in a divide-and-conquer manner. Beyond temporal structuring, PhysHPO [147] generalizes fine-grained preference optimization by organizing preferences across hierarchical semantic levels, including instance, state, motion, and semantic alignment. By applying Direct Preference Optimization across multiple abstraction levels, PhysHPO enables the generation of physically plausible videos that extend beyond surface appearance.

Stability, Efficiency, and Hybrid Preference Optimization. While preference objectives effectively guide alignment, applying them to video diffusion models often introduces instability, high cost, and scalability issues; recent work therefore focuses on making preference optimization more robust and efficient [148, 149]. Diffusion-NPO [150] trains a complementary negative-preference model to explicitly teach the generator what to avoid, which improves classifier-free guidance behavior and reduces undesirable outputs without requiring new datasets. To reduce optimization cost and variance, BranchGRPO [151] restructures GRPO rollouts into a branching tree with depth-wise reward fusion and pruning, amortizing shared computation, producing denser step-level advantages, and substantially accelerating and stabilizing training. To preserve diversity while scaling preference optimization, DPP-GRPO [152] formulates set-level policy optimization using a Determinantal Point Process term with GRPO, turning diversity into an explicit objective so that models learn to generate diverse video sets without sacrificing prompt fidelity.

5.4. Video Reward Modeling

Effective alignment of video generation models critically depends on the availability of reliable reward signals that reflect human preferences. Unlike images or text, video reward modeling must account for high-dimensional factors such as temporal dynamics, motion consistency, and long-range coherence, which substantially increase the difficulty of reward design. Recent work on video reward modeling can be broadly categorized by the aspects of video quality they emphasize, including multi-dimensional quality assessment, identity and temporal consistency, and physics- and reasoning-aware evaluation.

Multi-Dimensional Video Quality Assessment. A central direction in video reward modeling is to decompose human preference into multiple quality dimensions, recognizing that video alignment cannot be captured by a single scalar score [33, 126]. VideoReward [155] establishes a large-scale, human-annotated preference dataset over modern video generation models and trains a multi-dimensional reward model that separately evaluates visual quality, motion quality, and text-video alignment. By explicitly modeling these dimensions under a Bradley–Terry-with-ties formulation, VideoReward provides a robust reward backbone for preference-based and reinforcement learning alignment in video generation. While VideoReward targets open-domain video generation, AnimeReward [149] shows that generic video reward models fail to capture domain-specific quality criteria in anime generation, particularly appearance stylization and character consistency. To address this gap, AnimeReward constructs the first anime-specific multi-dimensional reward dataset and employs specialized vision-language models for different evaluation dimensions, demonstrating that domain-aware reward decomposition is critical for aligning stylized video generation with human preferences.

Identity, Consistency, and Temporal Coherence Rewards. Beyond overall quality assessment, a central challenge in video generation is preserving subject identity and maintaining coherent appearance and motion over time, motivating reward designs that explicitly target video-specific consistency failures. PersonalVideo [156] addresses this problem by introducing an Identity Consistency Reward that evaluates whether generated frames preserve the reference identity, together with a complementary Semantic Consistency Reward that constrains the semantic distribution of generated videos to remain aligned with the original text-to-video model, thereby avoiding dynamic and semantic degradation during identity injection.

Similarly, IPRO [137] formulates identity preservation as direct optimization with a differentiable facial identity reward. By backpropagating the identity reward through the final denoising steps of the diffusion process and regularizing deviation from the base model, IPRO effectively suppresses identity drift across frames while maintaining temporal coherence. Beyond identity preservation, AR-Drag [157] introduces a trajectory-based reward model that explicitly evaluates motion paths in autoregressive generation. This reward provides fine-grained supervision over temporal dynamics and controllability, enabling stable and coherent motion generation in long-horizon, few-step autoregressive-controlled diffusion video generation.

Physics- and Reasoning-Aware Reward Modeling. Beyond perceptual quality and temporal consistency, recent work explores reward designs that explicitly encode physical laws and reasoning structure, aiming to align video generation with objective physical plausibility rather than subjective visual cues. NewtonRewards [158] introduces the first physics-grounded post-training framework based on verifiable rewards that extracts measurable proxies, such as optical flow and visual appearance features, to enforce Newtonian constraints, including constant-acceleration dynamics and mass conservation, thereby penalizing physically implausible motion.

Similarly, PhysCorr [141] proposes PhysicsRM, a dedicated physics reward model that jointly evaluates intra-object stability and inter-object interactions, providing a structured assessment of physical consistency that goes beyond frame-level aesthetics. Beyond explicit physical rules, PRFL [159] reframes reward modeling as a process-aware task by repurposing video-generation models as latent reward models, enabling timestep-aware reward evaluation directly in the noisy latent space and supporting reasoning about motion and structure throughout the denoising trajectory. Together, these approaches highlight physics- and reasoning-aware reward modeling as a crucial component for enforcing causal and physical faithfulness in video generation.

Table 3 | Summary of preference-based and reinforcement learning methods for video generation.

Model	# Stages	Base Model	GPU Model	# GPUs	Venue	Year	Link
InstructVideo [25]	1	ModelScopeT2V	A100	4	CVPR	2024	
T2V-Turbo [321]	1	VideoCrafter2 ModelScopeT2V	A100	8	NeurIPS	2024	
VADER [322]	1	VideoCrafter OpenSora ModelScopeT2V Stable Video Diffusion	A6000	2	arXiv	2024	
Prompt-A-Video [24]	2	OpenSora CogVideoX	–	–	ICCV	2024	
PersonalVideo [156]	1	HunyuanVideo AnimateDiff	A800	1	ICCV	2025	
VideoDPO [18]	–	VideoCrafter2 T2V-Turbo CogVideo	A100	4	CVPR	2025	
VideoReward [155]	3	–	A800	8	NeurIPS	2025	
MagicID [143]	1	HunyuanVideo	H100	1	ICCV	2025	
DF-DPO [139]	1	CogVideoX	H100	8	arXiv	2025	–
AnimeReward [149]	3	CogVideoX	A800	8	arXiv	2025	
Phys-AR [130]	3	Llama3.1	A800	32	arXiv	2025	–
DiffusionNPO [150]	1	VideoCrafter2	–	–	ICLR	2025	
DenseDPO [145]	1	MAGVIT-v2	A100	64	NeurIPS	2025	–
Seedance 1.0 [126]	4	DiT	–	–	arXiv	2025	–
RDPO [142]	3	LTX-Video	H100	32	arXiv	2025	–
BranchGRPO [151]	1	FLUX.1-Dev Wan2.1	H200	16	arXiv	2025	
RLGF [132]	1	MagicDrive-V2	A100	8	NeurIPS	2025	–
PhysMaster [138]	3	DiT	A800	8	arXiv	2025	
IdentityGRPO [136]	2	VACE	A100	8	arXiv	2025	
Epipolar-DPO [144]	1	Wan2.1	A6000	4	arXiv	2025	
PhysCorr [141]	2	Wan2.1	A800	4	arXiv	2025	–
Ar-Drag [157]	2	Wan2.1	H200	8	arXiv	2025	–
McSc [134]	3	VideoCrafter2 Wan2.1	A100	8	arXiv	2025	
ID-Crafter [135]	1	Wan	H20	16	arXiv	2025	
BPGO [148]	1	Wan2.1 Wan2.2	H100	16	arXiv	2025	–
DPP-GRPO [152]	2	Wan2.1 CogVideoX	L40S	4	arXiv	2025	–
Self-paced GRPO [133]	1	Wan2.1 HunyuanVideo	H100	16	arXiv	2025	–
NewtonRewards [158]	2	OpenSora	H100	8	arXiv	2025	
IC-World [129]	2	Wan2.1	H20	8	arXiv	2025	
Euphonium [316]	1	HunyuanVideo	H800	40	arXiv	2026	
HuDA [323]	1	Wan2.1	H100	32	arXiv	2026	–
PhysRVG [324]	2	Wan2.2	H20	32	arXiv	2026	–

6. Inference-Time Methods via Post-trained Signals

Takeaways

- Inference-time alignment operationalizes post-trained signals by steering video generation during sampling, allowing alignment objectives to be enforced without further parameter updates.
- Guidance-based methods modify denoising trajectories using learned alignment signals or auxiliary models to control semantics, structure, motion, and physical plausibility while preserving the pretrained generative prior.
- Iterative refinement and self-editing regulate video generation through inference-time feedback loops or multi-stage refinement, enabling error correction, long-horizon consistency, and fine-grained control via closed-loop inference alone.

While Sections 3–5 focus on how video generation models are aligned through post-training procedures, alignment does not cease once training is complete. In practice, post-training produces a variety of alignment artifacts, such as reward models, learned critics, auxiliary guidance modules, and alignment-specific adapters, whose influence on model behavior is ultimately realized during inference-time generation. At inference time, these learned alignment signals are operationalized to steer, constrain, or refine video generation without further updating model parameters. Rather than defining new alignment objectives or training paradigms, such mechanisms govern how post-trained signals are consumed during sampling, shaping generation trajectories, enforcing semantic or physical constraints, and regulating trade-offs among competing alignment objectives.

6.1. Guidance-based Alignment

Guidance-based alignment directs the video generation process by injecting auxiliary control signals into the denoising trajectory at inference time. This approach influences generation towards specified semantics, structures, or dynamics without modifying the underlying model parameters. Guidance-based methods provide careful control over objects, motion, and style by shaping intermediate latent states throughout the denoising process, while maintaining the generative prior of the pretrained model.

Direct Trajectory Guidance. Direct trajectory guiding directs video generation by altering the denoising trajectory during inference, usually by direct modulation of latent variables or conditioning signals, while maintaining fixed model parameters [29, 160, 166, 325]. InstanceV [30] examples this paradigm by implementing spatially aware unconditional guiding that directly modifies the denoising process to maintain instance-level consistency, hence reducing the loss or distortion of small objects during sampling. ALG [161] further shows that trajectory perturbation can directly target motion dynamics, selectively filtering high-frequency conditioning signals at early denoising steps to prevent static generation and encourage more expressive motion. More fine-grained control is achieved by selective latent intervention methods such as Masked Latent Adaptation [162], which applies learned masks to restrict guidance to task-relevant latent regions, allowing targeted alignment of motion or appearance while preserving the pretrained generative prior.

Semantic and Structural Steering via Auxiliary Models. In addition to direct trajectory perturbation, an alternative approach directs video generation using auxiliary models that provide high-level semantic or structural signals during inference. CSVC [163] illustrates this paradigm by utilizing vision-language models (VLMs) to establish semantic or counterfactual objectives that then guide the denoising trajectory towards intended causal or semantic results without modifying the generator parameters. SynMotion [164] introduces semantic guidance by first breaking down textual descriptions into motion-relevant components with the help of an auxiliary semantic model. This decomposition allows finer control over motion patterns during sampling. In a more constrained setting, DiffPhy [165] takes a different approach: it relies on external physical rule checkers to evaluate intermediate latent states and filters out trajectories that violate inferred physical laws. Together, these methods illustrate how auxiliary models can act as validators that shape generation behavior at inference time.

6.2. Iterative Refinement and Self-editing

Complementing guidance-based steering, inference-time iterative refinement optimizes generation through feedback-driven adjustments. By applying cyclic updates to intermediate representations without modifying model parameters, these approaches improve temporal consistency, motion accuracy, and structural coherence. This perspective highlights the possibility of regulating generation behavior through closed-loop refinement alone.

Inference-Time Iterative Refinement and Self-Correction. Inference-time iterative refinement operates through multi-step feedback loops that progressively revise intermediate representations or generation plans during sampling. By iteratively correcting intermediate states, these methods address motion errors, temporal artifacts, and structural inconsistencies without updating model parameters. Feedback may be applied directly in latent space or at a higher-level planning and verification stage, enabling refinement at both fine-grained spatiotemporal details and long-horizon generation behavior.

At the latent level, DragVideo [31] uses iterative motion supervision on noisy latents to align generated motion with user-defined point trajectories, enabling precise spatiotemporal control. DFVEdit [167] instead relies on cyclic latent updates for zero-shot video editing, where latent representations are repeatedly refined to achieve the desired edits without any model retraining. FlashI2V [168] takes a complementary direction by revisiting the initialization strategy. By gradually shifting the initial noise distribution during inference, it mitigates conditional image leakage and produces smoother, more consistent motion. Beyond latent manipulation, self-correction mechanisms introduce recursive feedback at the planning or reasoning level. MotionAgent [169] adopts an agentic framework in which the model performs a “rethinking” step to verify motion alignment and iteratively adjust generation plans, enabling more robust long-horizon motion consistency through closed-loop inference-time feedback.

Cascaded and Multi-Stage Refinement. Cascaded and multi-stage refinement structures inference into successive stages, where early stages establish coarse motion and layout and later stages focus on refining local interactions and visual details [326, 327]. By separating global dynamics from fine-grained refinement, this design reduces the accumulation of early motion errors that often degrade long and complex video generations [328]. iDiT-HOI [170] exemplifies this paradigm with a two-stage diffusion transformer that first captures coarse motion patterns and then refines complex hand–object interactions, leading to improved temporal coherence and physical plausibility. More generally, cascaded refinement architectures assign distinct semantic or temporal roles to different stages, allowing later stages to condition on stabilized intermediate representations rather than raw noise, which improves robustness in long-horizon generation. Compared to iterative refinement methods that rely on cyclic feedback and correction, cascaded refinement adopts a feed-forward, stage-wise inference paradigm, trading iterative flexibility for improved stability and more predictable computational cost. Overall, staging inference in this way provides coarse-to-fine control without sacrificing inference-time efficiency.

Table 4 | Summary of inference-time methods via post-trained signals for video generation.

Model	Category	Base Model	Venue	Year	Link
Gen-1 [325]	Direct Trajectory Guidance	Stable Diffusion	ICCV	2023	
MotionAgent [169]	Iterative Refinement and Self-Correction	Stable Video Diffusion	ICCV	2025	
AICL [29]	Direct Trajectory Guidance	VideoCrafter VideoCrafter2 LVDM	ACM MM	2025	–
PAHA [160]	Direct Trajectory Guidance	VLDM	arXiv	2025	–
InstanceV [30]	Direct Trajectory Guidance	Wan	arXiv	2025	–
ALG [161]	Direct Trajectory Guidance	CogVideoX Wan 2.1 HunyuanVideo LTX	arXiv	2025	
CSVC [163]	Semantic and Structural Steering	Stable Diffusion	arXiv	2025	
SynMotion [164]	Semantic and Structural Steering	HunyuanVideo	arXiv	2025	–
DiffPhy [165]	Semantic and Structural Steering	Wan2.1	arXiv	2025	–
DFVEdit [167]	Iterative Refinement and Self-Correction	CogvideoX Wan2.1	arXiv	2025	
FlashI2V [168]	Iterative Refinement and Self-Correction	Wan2.1	arXiv	2025	
iDiT-HOI [170]	Cascaded and Multi-Stage Refinement	Wan FLUX.1-Dev	arXiv	2025	–
Raccoon [327]	Cascaded and Multi-Stage Refinement	–	arXiv	2025	–
WMReward [166]	Direct Trajectory Guidance	MAGI-1 VLDM	arXiv	2026	–

7. Datasets, Benchmarks, and Evaluation Protocols

Takeaways

- Post-training and alignment datasets encode alignment objectives explicitly, providing targeted supervision for instruction following, temporal consistency, identity preservation, physical plausibility, and preference modeling beyond large-scale pretraining data.
- Benchmarks for video generation alignment are increasingly organized by alignment dimensions, separating instruction adherence, long-horizon temporal coherence, and physical plausibility to enable more diagnostic and complementary evaluation.
- Evaluation protocols are commonly grouped into three categories: automated metrics, learned evaluators, and human judgments, each serving distinct roles within the evaluation pipeline.

While post-training methods determine how video generation models are optimized, datasets, benchmarks, and evaluation protocols define what it means for a model to be aligned in practice. Datasets encode alignment objectives through structured supervision, benchmarks translate these objectives into concrete evaluation targets, and evaluation protocols specify how aligned behavior is measured and compared. Rather than serving as passive resources, these components shape how post-trained video generation models are developed, diagnosed, and evaluated.

7.1. Post-training Datasets

Post-training and alignment of video generation models depend not only on optimization methods, but also critically on the datasets that encode alignment signals. Unlike large-scale pretraining corpora that prioritize coverage and diversity, datasets used for post-training emphasize specific alignment objectives. These datasets are typically curated to reflect deployment-time requirements, providing structured supervision that shapes model behavior beyond likelihood-based learning and ultimately determines the types of alignment models can reliably achieve. Accordingly, they can be categorized by the type of alignment signal they provide, including instruction-following supervision, temporal consistency and identity preservation, physics- and reasoning-oriented constraints, and preference signals derived from either synthetic or real videos.

Table 5 | Datasets used for training in video generation post-training and alignment.

Name	Size	Tasks	Link
ChronoMagic-Pro [194]	460,000	High resolution time-lapse video.	🤗
SafeSora [182]	57,333	Human preference text-video pairs for safety and value alignment.	🤗
CookGen [178]	200,000	Long-form narrative generation in the cooking domain.	🤗
HOIGen-1M [177]	1,000,000	Human-object interaction videos.	🤗
TIP-I2V [174]	1,700,000	User-driven text-image prompt dataset for image-to-video generation.	🤗
SynFMC [329]	62,000	Camera-object motion control for video generation.	🤗
PhyWorld [330]	6,000,000	Physics-simulated video prediction dataset.	🤗
OpenS2V-5M [176]	5,000,000	High resolution subject-text-video triples.	🤗
EgoVid-5M [180]	5,000,000	Egocentric videos with action annotations.	
VideoUFO [172]	1,091,712	User-focused topic-aligned text-video pairs for text-to-video generation.	🤗
WISA-80K [54]	79,500	Physics-aware text-to-video generation.	🤗
CI-VID [173]	340,000	Coherent sequence of video clips with text captions.	🤗
OpenHumanVid [179]	52,300,000	Human-centric text-video pairs with fine-grained appearance and motion.	–
TalkCuts [195]	164,000	Multi-shot human speech videos.	–
GRADEO-Instruct [331]	3,300	Human-annotated video-rationale-score triples.	–
MMVideo [175]	350,000	Hybrid real-and-synthetic dataset aligned across modalities and captions.	–
Dprim [181]	32,000	Primitive-level embodied video prediction for robotic world modeling.	–
DAVID-X [183]	747	Defect-annotated explainable AI-generated video detection dataset with spatiotemporal evidence and rationales.	–
PairFS-4K [196]	4,000	Two-person figure skating video dataset.	–
PNData [171]	296,960	Prompt-random-noise-refined-noise triples.	–

Instruction-following datasets. Instruction-following datasets aim to ensure that generated videos accurately reflect user intent expressed through textual descriptions and, when available, structured conditions [171–173]. Representative examples include TIP-I2V [174], which collects millions of real-world text and image prompts from user interactions, capturing realistic prompt distributions that differ substantially from those of curated captions. Such datasets are particularly valuable for aligning image-to-video models with user intent, as they expose failure modes arising from incomplete, underspecified, or noisy prompts. Beyond purely textual supervision, MMVideo [175] pairs text prompts with densely aligned multimodal annotations covering geometry, appearance, and semantics. This form of supervision translates instructions into executable constraints, enabling post-training methods to improve semantic adherence, controllability, and robustness under diverse instruction formulations. Together, these datasets support post-training strategies that improve instruction adherence while maintaining robustness under diverse prompt formulations.

Temporal consistency and identity datasets. A second class of datasets focuses on inherently temporal alignment objectives, such as long-range coherence, motion stability, and identity preservation [176–178]. Because small frame-level errors can quickly accumulate into perceptual artifacts, these datasets are designed to stress-test temporal consistency and provide supervision that penalizes such failures. OpenHumanVid [179] exemplifies this category by focusing on human-centric videos that require consistent appearance and articulation across diverse motions and viewpoints. EgoVid-5M [180] refines the temporal alignment objective to egocentric video generation, where first-person camera motion is tightly linked to action dynamics. EgoVid-5M reveals temporal failure patterns frequently overlooked in more generic datasets through careful kinematic control signals, detailed action annotations, and comprehensive data cleaning. ViMoGen-228K [332] offers an alternative perspective by emphasizing motion diversity and generalization, promoting expressive dynamics while ensuring temporal stability. Together, these datasets serve as important sources of alignment supervision for post-training methods aimed at reducing temporal drift while preserving motion realism and identity consistency.

Physics and reasoning datasets. Beyond perceptual coherence, an emerging class of datasets targets physical plausibility and causal consistency, reflecting the growing interest in video generation models as world simulators. These datasets encode alignment objectives that extend beyond appearance and motion, emphasizing whether generated videos adhere to basic physical laws, object interactions, and cause-and-effect relationships. Datasets such as WISA-80K [54] introduce physics-aware supervision by constructing videos that reflect structured world dynamics, which helps post-training methods better align generated outputs with physical constraints. In more embodied, domain-specific scenarios, datasets such as Dprim [181] go a step further by linking video generation to action-conditioned world transitions. In these settings, physical consistency is evaluated alongside downstream tasks such as robotics. These datasets play a crucial role in supporting reinforcement learning and preference-based alignment methods that rely on verifiable, rule-based signals rather than purely subjective judgments.

Synthetic versus real preference datasets. Finally, a distinct category of datasets provides preference-based alignment signals by contrasting synthetic and real videos, often with explicit annotations of failure modes. Rather than specifying how a video should be generated, these datasets define what constitutes undesirable or unacceptable outcomes, making them particularly useful for alignment diagnosis, evaluation, and preference optimization. SafeSora [182] exemplifies this direction by collecting human preference annotations focused on safety and value alignment in text-to-video generation. More diagnostically oriented datasets, such as DAVID-X [183], pair AI-generated and real videos with fine-grained spatio-temporal defect annotations and natural language rationales. Although such datasets are rarely used to directly train video generators, they provide valuable preference signals that inform post-training strategies, reward modeling, and evaluation protocols. By annotating identity inconsistencies, motion anomalies, and physically implausible behaviors, synthetic-versus-real datasets facilitate alignment of automated training objectives with human judgment.

7.2. Benchmarks by Alignment Dimensions

Unlike datasets, which mainly specify the source and structure of supervision, benchmarks translate alignment goals into concrete evaluation targets. In video generation, benchmarks are increasingly designed around specific alignment dimensions rather than comprehensive quality assessment. As a result, they tend to isolate particular aspects of aligned behavior, such as instruction adherence, temporal coherence, or physical plausibility. This dimension-oriented perspective clarifies how different benchmarks capture complementary aspects of alignment and enables more meaningful comparisons across methods.

Instruction-following and controllability benchmarks. Benchmarks in this category emphasize semantic correctness and controllability, assessing whether generated videos adhere to certain instructions about subjects, actions, scene setup, and audio outputs [41, 185–190]. This class of benchmarks is especially important for post-training evaluation, since instruction-following failures often persist even when visual fidelity appears high. Representative benchmarks include OpenS2V-Eval [176], which evaluates subject-to-video generation by measuring whether models preserve subject identity and attributes specified in the input. In addition, domain-structured benchmarks such as RecipeGen [191] and CineTechBench [192] assess instruction execution under procedural or cinematic constraints, while TAVGBench [193] extends instruction-following evaluation to multimodal settings by jointly considering audio and video outputs. Together, these benchmarks frame instruction adherence as a distinct alignment dimension, making it possible to analyze how effectively models translate intent into controlled video generation.

Temporal consistency and identity preservation benchmarks. Temporal alignment benchmarks evaluate whether video generation models preserve coherent structure over long durations. Rather than prioritizing immediate semantic accuracy, these benchmarks emphasize temporal consistency and examine whether models maintain stable dynamics across frames, shots, or narrative segments. This perspective is reflected in benchmarks that target different forms of long-horizon coherence. ChronoMagic-Bench [194] evaluates text-to-time-lapse generation under strong physical priors, such as biological growth or physical transformations. It measures metamorphic amplitude and temporal coherence, rather than appearance stability alone. For multi-character interaction settings, DanceTogether [196] introduces TogetherVideoBench. This benchmark specifically evaluates identity-action binding, assessing the model's ability to maintain distinct identities during complex, extended interactions. Together, these benchmarks highlight long-horizon temporal alignment as a multi-faceted objective, spanning physical progression, interaction stability, and narrative coherence.

Physical plausibility and world-model benchmarks. Physical plausibility benchmarks target alignment objectives that extend beyond perceptual coherence. Rather than asking whether a video looks consistent over time, these benchmarks assess whether the depicted dynamics match real-world expectations [197–199]. This line of work is motivated by viewing video generation models as implicit world simulators. Representative benchmarks in this category are often grounded in task-oriented or embodied settings. WorldSimBench [200] evaluates whether generated videos support world simulation. It combines human feedback with downstream video-to-action or agent-centric evaluations to test whether the dynamics are actionable and physically meaningful. In a similar spirit, Drive&Gen [201] evaluates physical plausibility through domain-specific tasks such as autonomous driving. In these settings, violations of physical consistency directly degrade downstream performance. Unlike preference-based or purely perceptual benchmarks, these evaluations rely on structured criteria and verifiable outcomes. This makes them particularly compatible with reinforcement learning and verification-driven alignment methods.

Table 6 | Representative benchmarks used for video generation post-training and alignment evaluation.

Name	Size	Tasks	Link
FETV [333]	618	Fine-grained and temporal-aware evaluation of text-to-video generation.	🤗
StoryBench [334]	6,000	Story-driven text-to-video generation evaluation	–
VBench [184]	–	Multi-dimensional video generation evaluation.	
ChronoMagic-Bench [194]	1,649	Time-lapse T2V generation; temporal coherence and metamorphic change evaluation.	🤗
EvalCrafter [202]	700	Text-to-video generation across diverse prompt types and multi-dimensional quality criteria.	🤗
TAVGBench [193]	1,700,000	Text to Audible-Video Generation.	🤗
T2VSafetyBench [335]	4,400	Text-to-video model safety assessment.	–
MTBench [336]	100	Motion transfer task evaluation.	🤗
FiVE [337]	100	Fine-grained text-guided video editing evaluation.	🤗
StoryEval [338]	423	Story-level multi-event text-to-video generation evaluation.	
MJ-BENCH-VIDEO [339]	10,842	Fine-grained video preference evaluation.	🤗
OpenS2V-Eval [176]	180	Subject-consistent video generation.	🤗
Doc2Present [190]	30	Document-to-presentation video generation.	🤗
VideoPhy [198]	688	Physical commonsense for real-world activities assessment.	🤗
T2V-CompBench [189]	700	Compositional text-to-video generation.	🤗
VEG-Bench [41]	132	Instructional video editing.	🤗
VMBench [188]	1,050	Human perception-aligned motion evaluation.	🤗
VidCapBench [187]	643	Text-to-video generation video caption evaluation.	🤗
VideoGen-RewardBench [155]	26,500	Annotated prompt-video pairs for reward model evaluation.	🤗
Verse-Bench [340]	600	Joint audio-video generation evaluation.	🤗
AIGC-LipSync [341]	615	Audio-driven video lip synchronization evaluation.	🤗
DisenStudioBench [275]	1,500	Customized multi-subject text-to-video generation.	–
TC-Bench [197]	270	Temporal Compositionality of video generation assessment.	–
HVEval [342]	20,000	Human-centric videos generation.	–
PhyGenBench [234]	160	Evaluate physical commonsense correctness in text-to-video generation.	–
Video-Bench [186]	419	Human-aligned video generation.	–
AIGVQA-DB [208]	36,576	Text-to-video model capability assessment.	–
ETVABench [209]	2,000	Textvideo alignment evaluation.	–

7.3. Evaluation Protocols and Metrics

While benchmarks define which aspects of alignment are evaluated, evaluation protocols and metrics determine how those aspects are measured and compared. In video generation, evaluation typically draws on multiple sources of evidence, ranging from automated metrics to learned evaluators and human judgments. Each approach rests on different assumptions about perceptual quality, semantic correctness, and temporal coherence. As a result, no single evaluation protocol is sufficient to cover all alignment dimensions.

Automated metrics. Automated metrics form the backbone of evaluation protocols in video generation, offering scalable and reproducible evaluation. Early methods largely borrowed metrics from image generation and video compression, such as FID [343], SSIM [344], PSNR [344], and LPIPS [345]. These metrics measure visual similarity or reconstruction quality at the frame level. They are effective for evaluating low-level visual quality, but struggle to capture semantic correctness and temporal coherence. In particular, they either ignore cross-frame dynamics or treat videos as collections of independent frames. To better account for temporal structure, video-specific metrics such as FVD [346] are introduced. FVD is more sensitive to motion and temporal inaccuracies because it assesses distributions of video features rather than individual frames. However, it remains focused on visual realism and does not explicitly measure alignment with conditioning signals.

As video generation models increasingly emphasize text conditioning, later metrics incorporate vision-language representations [202–204]. These metrics assess text-video correspondence and overall perceptual quality at the clip level. VBench [184] exemplifies this direction by combining multiple automated signals into a unified evaluation pipeline. For example, visual quality is measured using perceptual similarity in learned feature spaces, semantic alignment is assessed through vision-language representations, and temporal consistency is approximated with motion-sensitive video features. In this setting, automated metrics are rarely used in isolation. Instead, they are combined into standardized pipelines that capture complementary aspects of video-generation quality.

More recent approaches, such as VideoScore [205] and VideoScore2 [206], further extend automated evaluation through learned scoring models. These models are trained to approximate human judgments across multiple perceptual and semantic dimensions. Unlike hand-crafted similarity metrics, they aggregate heterogeneous cues into a single evaluative signal, while still being used primarily for evaluation rather than optimization. In addition to general-purpose metrics, some benchmarks introduce task-specific automated indicators to capture alignment dimensions that are poorly reflected by standard scores. For example, ChronoMagic-Bench [194] proposes MTScore and CHScore to assess metamorphic amplitude and long-range temporal coherence in time-lapse generation. They demonstrate that domain-specific metrics can supplement generic evaluation metrics when alignment targets are defined.

Learned evaluators. Beyond fixed metrics and aggregation-based protocols, recent work increasingly adopts learned evaluators to approximate human judgment in video generation. These evaluators vary in supervision sources, output formats, and inference mechanisms. However, they share a common goal: capturing alignment properties that are difficult to express with hand-crafted similarity measures. Existing approaches can be broadly divided into two categories. One group consists of reward-style evaluators that produce scalar scores [200, 347]. The other relies on LLM-based evaluators that provide explicit reasoning or generative feedback. Representative examples of reward-style evaluators is AnimeReward [149] and VideoReward [155], which are trained on human preference data to produce scalar scores reflecting perceptual quality and alignment. These models aggregate heterogeneous visual and semantic cues into a single reward signal, enabling scalable evaluation that correlates more closely with human judgment than traditional metrics. Although originally introduced for optimization and post-training, such reward models are also widely reused as evaluators due to their simplicity and effectiveness. However, their evaluations remain largely opaque, as alignment is summarized through numerical scores without explicit reasoning or interpretability.

More recent approaches leverage MLLMs as evaluators, treating evaluation as a reasoning or generation task rather than pure scoring [198, 207, 208, 337]. ETVA [209] exemplifies a reasoning-based LLM evaluator. It assesses text-video alignment through question-driven evaluation, enabling fine-grained checks of semantic attributes such as object existence, relations, and physical consistency beyond similarity-based metrics. Complementary to reasoning-based methods, AIGVE-MACS [210] represents a generative MLLM evaluator. It produces both aspect-wise scores and natural language feedback, framing evaluation as structured generation. This design improves interpretability and supports more diagnostic analysis of alignment quality.

Human evaluation and hybrid protocols. Despite advances in automated metrics and learned evaluators, human evaluation remains the reference standard for assessing alignment in video generation. This is especially true for visual realism, semantic precision, and overall preference. Common human evaluation protocols include absolute rating, pairwise comparison, and ranking-based judgments [202, 333]. In addition, recent work explores more efficient methods for scaling human feedback. Arena-style frameworks, such as K-Sort Arena [211], improve the robustness of preference-based benchmarking through structured comparison and probabilistic ranking. Importantly, these methods do not rely on automated evaluators. In practice, evaluation pipelines increasingly adopt hybrid protocols. Automated metrics and learned evaluators are used for large-scale screening and diagnostic analysis, while human evaluation is reserved for validation and final comparison. This hybrid paradigm balances scalability with reliability and reflects current best practices for evaluating aligned video generation models.

8. Challenges and Future Directions

Takeaways

- Supervised fine-tuning is limited by tightly coupled spatial and temporal representations and weak long-horizon reasoning, making it difficult to jointly scale identity preservation, motion adaptation, and multi-stage instruction alignment.
- Self-training and distillation risk error accumulation and trajectory mismatch, requiring verification and temporally aware objectives for robustness.
- Preference-based and reinforcement learning suffer from coarse reward design and unstable optimization, often leading to conservative motion and limited scalability.
- Inference-time alignment suffers from limited generality, guidance conflicts, and computational overhead, motivating more structured, reasoning-driven control.
- Evaluation and benchmarking lack systematic cross-paradigm comparison and fail to thoroughly analyze objective-induced trade-offs and dynamic safety failures.

Despite rapid progress in post-training and alignment techniques for video generation models, significant challenges remain before these systems can achieve robust, controllable, and trustworthy deployment. Building on the methodological taxonomy presented in Sections 3–7, we outline key open problems and promising research directions across supervised fine-tuning, preference learning and reinforcement learning, self-training and distillation, inference-time alignment, as well as cross-cutting challenges in evaluation, benchmarking, and safety.

8.1. Supervised Fine-tuning

Decoupled Appearance and Motion Representation. Supervised fine-tuning often exposes a tight coupling between appearance and motion representations in video generation models. Adapting models to new motion patterns can unintentionally alter identity and visual consistency [22, 65]. To mitigate this issue, existing work seeks to decouple motion dynamics from appearance, typically via disentangled representations or specialized conditioning mechanisms [64, 86]. However, these approaches reveal an inherent trade-off between preserving identity and learning new motion patterns. In modern video generators, spatiotemporal latent structures are highly shared. As a result, motion, texture, and identity remain tightly coupled during optimization, making full disentanglement difficult in practice. Future research may explore dual-branch or multi-stream fine-tuning methods that separate appearance and motion pathways.

Long-horizon and Multi-stage Instruction Alignment. Current supervised fine-tuning methods are typically optimized for short video clips of a few to tens of seconds. Consequently, models often experience challenges with following long-horizon or multi-stage instructions [294]. Over time, movements start to drift, details get blurry, and the background shifts unnaturally [112, 286]. This makes the video inconsistent and causes it to ignore the original prompt. Prior work has explored hierarchical planning or long-context tuning to extend temporal coherence [50, 84, 85]. However, these methods are often computationally expensive and struggle with complex causal reasoning and long-term temporal dependencies. One possible direction is to use MLLMs as high-level planners that translate user intent into structured generation plans. In parallel, more efficient generation mechanisms are needed to scale to long videos. For instance, dynamic token routing could help reduce redundancy during long-context generation.

8.2. Self-training and Distillation

Model Collapse and Error Accumulation in Self-training. Self-training on model-generated videos carries a risk of model collapse. Repeatedly training on self-produced data can reduce diversity and reinforce existing biases. In long video generation, this problem becomes more severe. Small errors introduced early in the video can accumulate over time and gradually dominate later frames [112]. A promising direction is verifier-guided self-evolution, where self-training is paired with an explicit verification stage [85, 169]. In this setting, generated videos in earlier steps are first evaluated by a verifier, such as a critic or world model, and only verified or corrected samples are reused for training [130, 165]. This closed-loop design refines model behavior while limiting error accumulation.

Temporal Consistency Loss in Distillation. Distilling diffusion-based video generators from hundreds of denoising steps to only a few inference steps often harms temporal smoothness and motion continuity. This typically leads to flickering artifacts or abrupt motion changes. Recent studies suggest that this issue arises because many distillation methods rely on distribution-level matching and overlook the temporal trajectories of the generation process [116, 118]. One promising direction is trajectory-aware adversarial distillation, which combines adversarial supervision on temporal coherence with trajectory- or flow-based matching objectives. Beyond accelerating inference, future distillation methods should aim to teach the student model a temporally consistent velocity field that stays aligned with the teacher's generation dynamics [119].

8.3. Preference-based and Reinforcement Learning

Reward Limitations for Motion and Physical Plausibility. Preference-based and reinforcement learning methods reveal systematic limitations in reward design for video generation. Existing reward formulations tend to favor visually sharp and temporally stable outputs, implicitly encouraging conservative generation behaviors that suppress motion dynamics in video generation models [18]. In addition, most reward signals operate at the video level and rely on holistic preference judgments, which limit their sensitivity to localized temporal failures [145]. As a result, subtle but important physical errors, such as sliding motion or object interpenetration, are often missed, even though they greatly reduce physical plausibility [138, 141, 158, 234]. A promising direction is to incorporate physics-grounded reward signals that complement perceptual rewards [141]. These rewards can be derived from auxiliary physical constraints or verification signals. In parallel, supervision can move from holistic video-level labels to finer segment-level rewards [147]. This shift enables more localized credit assignment during optimization. With finer-grained rewards, optimization algorithms can penalize a specific time step rather than the entire generated video.

Inefficiency and Instability in Video Reinforcement Learning. Preference-based and reinforcement learning methods also face fundamental challenges in sample efficiency and training stability when applied to video generation. Video generation is inherently expensive. Constructing paired samples for preference optimization or performing online sampling for reinforcement learning quickly becomes prohibitive at scale [150]. In addition, video trajectories are long and high-dimensional. This often leads to high optimization variance, resulting in unstable training or convergence to overly conservative solutions [133, 151]. Together, these issues make it difficult to directly apply standard reinforcement learning pipelines to video generation at scale. One promising direction is to improve the efficiency of reinforcement learning, for example, by reusing previously generated samples via off-policy optimization or by operating directly in the latent space to reduce decoding cost [155].

8.4. Inference-Time Alignment

Limited Generality of Gradient-based Inference Guidance. Most inference-time guidance methods steer video generation using classifier-based gradients. This usually requires training task-specific evaluators, which limits flexibility. It also makes it hard to incorporate high-level semantic reasoning or physical constraints. Some recent work explores gradient-free alternatives, such as using VLMs for semantic steering or external rule checkers for physical filtering [163, 165]. However, these approaches are often limited to predefined constraints or simple rejection strategies. A more promising direction is agent-like control with powerful VLMs that monitor intermediate states and intervene during generation, adjusting prompts, attention, or motion plans in real time to enforce semantic and physical consistency without task-specific fine-tuning [348, 349].

Guidance Conflicts and Computational Overhead. Inference-time alignment methods often combine multiple forms of guidance, including text, object-level cues, and physical constraints. In practice, however, these signals can conflict with one another, which may lead to unstable generation or even collapsed outputs [27, 161]. In addition, many inference-time control strategies depend on iterative latent updates or repeated sampling, substantially increasing inference latency and limiting their use in interactive or real-time settings [30, 31]. One potential direction is test-time search and planning, which approaches alignment through explicit lookahead or structured exploration instead of local gradient-based guidance. Rather than directly optimizing latent variables, future systems could first generate high-level plans or keyframes, assess their semantic and physical validity, and then selectively refine intermediate frames via backtracking or branching search [85]. Such a strategy may offer more stable multi-objective alignment while keeping computational costs under control.

8.5. Evaluation and Benchmarking Challenges

Cross-Paradigm Comparisons and Design Principles. While post-training and alignment methods have shown promise, their relative strengths remain poorly understood. Most studies evaluate these paradigms in isolation, making it difficult to derive general design principles for video alignment [112, 126, 151, 237]. Although recent benchmarks such as Video-Bench [186], VMBench [188], and VideoGen-RewardBench [155] attempt to standardize evaluation protocols, they are rarely used for systematic cross-paradigm comparisons across supervised, preference-based, and reinforcement learning approaches. Such analyses would enable principled choices among alignment strategies based on task requirements, temporal horizon, and available supervision, ultimately yielding more reliable and interpretable video-generation models.

Failure Modes and Trade-offs. Beyond evaluation, post-training methods for video alignment introduce systematic trade-offs that are often insufficiently analyzed. Objectives that strongly penalize temporal inconsistency can lead models to minimize motion altogether, producing overly static or rigid videos [18, 145]. Similarly, identity-preservation losses that tightly constrain appearance across frames can limit compositional generalization, making it difficult for models to handle novel interactions or scene changes [22, 86, 136].

Preference- and reward-based optimization introduces additional trade-offs. Optimizing for a learned reward can bias generation toward a narrow set of reward-aligned patterns, reducing diversity and, in some cases, leading to reward exploitation where perceptual quality improves while long-term coherence or realism degrades [141, 147, 150]. These issues are not isolated failures but recurring behaviors induced by the optimization objectives themselves, including over-regularized motion, gradual temporal drift in long-horizon videos, and overspecialization to the proxy metrics used during training [294, 350]. Progress in video alignment, therefore, requires not only reporting aggregate performance gains but also systematically analyzing failure cases and regressions introduced by alignment objectives, especially outside controlled benchmark settings.

Temporal Safety and Dynamic Misalignment Analysis. Safety alignment in video generation extends beyond static content filtering and must account for temporal misalignment, where unsafe behaviors emerge gradually over time. Existing safety mechanisms are often prompt- or frame-centric and fail to capture delayed harms [182], evolving interactions, or cumulative violations that only become apparent through extended generation [68, 335]. We therefore highlight the need for temporally grounded safety objectives and evaluation protocols that monitor behavior across the full generation horizon. This includes detecting the gradual escalation of harmful content, identity misuse over time [92, 94], and temporal jailbreaks, in which models initially comply with constraints but later deviate [186, 200]. Treating safety as a dynamic alignment problem is essential for the reliable deployment of video generation models in real-world applications.

9. Conclusion

This survey reviews the emergence of post-training as a new trend in video generation, where the focus has gradually shifted from pure pre-training to targeted alignment and optimization. By combining supervised fine-tuning, preference- and reward-based methods, self-training, and inference-time control, recent approaches have improved controllability, temporal coherence, and alignment with user intent. Despite this progress, several fundamental challenges remain. Key issues include the tight coupling between appearance and motion during post-training, limitations in long-horizon and multi-stage instruction-following, reward and preference signals that fail to capture motion and physical plausibility, and the inefficiency and instability of preference-based and reinforcement learning on high-dimensional video trajectories. Future research will likely depend on more efficient optimization algorithms, stronger grounding signals, and a tighter integration between training-time alignment and inference-time computation. Progress along these directions will be crucial for building more robust and general-purpose video intelligence.

References

[1] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators, 2024.

[2] Le Zhuo, Ruoyi Du, Han Xiao, Yangguang Li, Dongyang Liu, Rongjie Huang, Wenze Liu, Lirui Zhao, Fu-Yun Wang, Zhanyu Ma, Xu Luo, Zehan Wang, Kaipeng Zhang, Xiangyang Zhu, Si Liu, Xiangyu Yue, Dingning Liu, Wanli Ouyang, Ziwei Liu, Yu Qiao, Hongsheng Li, and Peng Gao. Lumina-next: Making lumina-t2x stronger and faster with next-dit. In *Proceedings of the Thirty-Eighth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

[3] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. In *Proceedings of the Thirty Annual Conference on Neural Information Processing Systems*, 2016.

[4] Haotian Xue, Qi Chen, Zhonghao Wang, Xun Huang, Eli Shechtman, Jinrong Xie, and Yongxin Chen. Mogan: Improving motion quality in video diffusion via few-step motion-aware adversarial post-training. *arXiv preprint arXiv:2511.21592*, 2025.

[5] Emily Denton and Rob Fergus. Stochastic video generation with a learned prior. In *Proceedings of the Thirty-Fifth International Conference on Machine Learning (ICML)*, 2018.

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 2020.

[7] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proceedings of the Thirty-Fourth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[8] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

[9] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*, 2024.

[10] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2023.

[11] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. In *Proceedings of the Thirty-Sixth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

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

[13] Bin Lin, Yunyang Ge, Xinhua Cheng, Zongjian Li, Bin Zhu, Shaodong Wang, Xianyi He, Yang Ye, Shenghai Yuan, Liuhan Chen, et al. Open-sora plan: Open-source large video generation model. *arXiv preprint arXiv:2412.00131*, 2024.

[14] Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong MU, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. In *Proceedings of the Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

[15] 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. In *Proceedings of the Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

[16] 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 (TMLR)*, 2025.

[17] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2023.

[18] Runtao Liu, Haoyu Wu, Ziqiang Zheng, Chen Wei, Yingqing He, Renjie Pi, and Qifeng Chen. Videodpo: Omni-preference alignment for video diffusion generation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[19] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *Proceedings of the Tenth International Conference on Learning Representations (ICLR)*, 2022.

[20] Han Lin, Jaemin Cho, Abhay Zala, and Mohit Bansal. Ctrl-adapter: An efficient and versatile framework for adapting diverse controls to any diffusion model. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

[21] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. In *Proceedings of the Thirty-Seventh Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[22] Yuechen Zhang, Yaoyang Liu, Bin Xia, Bohao Peng, Zexin Yan, Eric Lo, and Jiaya Jia. Magicmirror: Id-preserved video generation in video diffusion transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2025.

[23] Ruojun Xu, Yu Kai, Xuhua Ren, Jiaxiang Cheng, Bing Ma, Tianxiang Zheng, and Qinhlin Lu. Beyond reward margin: Rethinking and resolving likelihood displacement in diffusion models via video generation. *arXiv preprint arXiv:2511.19049*, 2025.

[24] Yatai Ji, Jiacheng Zhang, Jie Wu, Shilong Zhang, Shoufa Chen, Chongjian Ge, Peize Sun, Weifeng Chen, Wenqi Shao, Xuefeng Xiao, et al. Prompt-a-video: Prompt your video diffusion model via preference-aligned llm. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2025.

[25] Hangjie Yuan, Shiwei Zhang, Xiang Wang, Yujie Wei, Tao Feng, Yining Pan, Yingya Zhang, Ziwei Liu, Samuel Albanie, and Dong Ni. InstructVideo: Instructing video diffusion models with human feedback. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[26] Ruicheng Zhang, Jun Zhou, Zunnan Xu, Zihao Liu, Jiehui Huang, Mingyang Zhang, Yu Sun, and Xiu Li. Zo3t: Zero-shot 3d-aware trajectory-guided image-to-video generation via test-time training. *arXiv preprint arXiv:2509.06723*, 2025.

[27] Xiuli Bi, Jian Lu, Bo Liu, Xiaodong Cun, Yong Zhang, Weisheng Li, and Bin Xiao. Customttt: Motion and appearance customized video generation via test-time training. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2025.

[28] Tianwei Yin, Qiang Zhang, Richard Zhang, William T. Freeman, Frédo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[29] Jianzhi Liu, Junchen Zhu, Pengpeng Zeng, Lianli Gao, Heng Tao Shen, and Jingkuan Song. Aicl: Action in-context learning for text-to-video generation. In *Proceedings of the 33rd ACM International Conference on Multimedia (ACM MM)*, 2025.

[30] Yuheng Chen, Teng Hu, Jiangning Zhang, Zhucun Xue, Ran Yi, and Lizhuang Ma. Instancev: Instance-level video generation. *arXiv preprint arXiv:2511.23146*, 2025.

[31] Yufan Deng, Ruida Wang, Yuhao Zhang, Yu-Wing Tai, and Chi-Keung Tang. Dragvideo: Interactive drag-style video editing. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

[32] Yujiang Pu, Zhanbo Huang, Vishnu Boddeti, and Yu Kong. Show me: Unifying instructional image and video generation with diffusion models. *arXiv preprint arXiv:2511.17839*, 2025.

[33] Xiaoyi Bao, Jindi Lv, Xiaofeng Wang, Zheng Zhu, Xinze Chen, YuKun Zhou, Jiancheng Lv, Xingang Wang, and Guan Huang. Gigavideo-1: Advancing video generation via automatic feedback with 4 gpu-hours fine-tuning. *arXiv preprint arXiv:2506.10639*, 2025.

[34] Yuwei Fang, Willi Menapace, Aliaksandr Siarohin, Tsai-Shien Chen, Kuan-Chieh Wang, Ivan Skorokhodov, Graham Neubig, and Sergey Tulyakov. Vimi: Grounding video generation through multi-modal instruction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

[35] Luis Denninger, Sina Mokhtarzadeh Azar, and Juergen Gall. Camc2v: Context-aware controllable video generation. In *Proceedings of the Thirteenth International Conference on 3D Vision (3DV)*, 2026.

[36] Zhen Xing, Qi Dai, Zejia Weng, Zuxuan Wu, and Yu-Gang Jiang. Aid: Adapting image2video diffusion models for instruction-guided video prediction. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2025.

[37] Angtian Wang, Haibin Huang, Jacob Zhiyuan Fang, Yiding Yang, and Chongyang Ma. Ati: Any trajectory instruction for controllable video generation. *arXiv preprint arXiv:2505.22944*, 2025.

[38] Mengyi Shan, Zecheng He, Haoyu Ma, Felix Juefei-Xu, Peizhao Zhang, Tingbo Hou, and Ching-Yao Chuang. Populate-a-scene: Affordance-aware human video generation. *arXiv preprint arXiv:2507.00334*, 2025.

[39] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2023.

[40] Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. Video-p2p: Video editing with cross-attention control. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[41] Shoubin Yu, Difan Liu, Ziqiao Ma, Yicong Hong, Yang Zhou, Hao Tan, Joyce Chai, and Mohit Bansal. Veggie: Instructional editing and reasoning video concepts with grounded generation. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2025.

[42] Sen Liang, Zhentao Yu, Zhengguang Zhou, Teng Hu, Hongmei Wang, Yi Chen, Qin Lin, Yuan Zhou, Xin Li, Qinglin Lu, et al. Omniv2v: Versatile video generation and editing via dynamic content manipulation. *arXiv preprint arXiv:2506.01801*, 2025.

[43] Ge Wang, Songlin Fan, Hangxu Liu, Quanjian Song, Hewei Wang, and Jinfeng Xu. Consistent video editing as flow-driven image-to-video generation. *arXiv preprint arXiv:2506.07713*, 2025.

[44] Chenjian Gao, Lihe Ding, Xin Cai, Zhanpeng Huang, Zibin Wang, and Tianfan Xue. Lora-edit: Controllable first-frame-guided video editing via mask-aware lora fine-tuning. *arXiv preprint arXiv:2506.10082*, 2025.

[45] Zeyu Dong, Yuyang Yin, Yuqi Li, Eric Li, Hao-Xiang Guo, and Yikai Wang. Panolora: Bridging perspective and panoramic video generation with lora adaptation. *arXiv preprint arXiv:2509.11092*, 2025.

[46] Xingyang Li, Muyang Li, Tianle Cai, Haocheng Xi, Shuo Yang, Yujun Lin, Lvmin Zhang, Songlin Yang, Jinbo Hu, Kelly Peng, et al. Radial attention: $o(n \log n)$ sparse attention with energy decay for long video generation. *arXiv preprint arXiv:2506.19852*, 2025.

[47] Sihyun Yu, Weili Nie, De-An Huang, Boyi Li, Jinwoo Shin, and Anima Anandkumar. Efficient video diffusion models via content-frame motion-latent decomposition. *arXiv preprint arXiv:2403.14148*, 2024.

[48] Haitam Ben Yahia, Denis Korzhenkov, Ioannis Lelekas, Amir Ghodrati, and Amirhossein Habibian. Mobile video diffusion. *arXiv preprint arXiv:2412.07583*, 2024.

[49] Mohsen Ghafoorian, Denis Korzhenkov, and Amirhossein Habibian. Attention surgery: An efficient recipe to linearize your video diffusion transformer. *arXiv preprint arXiv:2509.24899*, 2025.

[50] Yuwei Guo, Ceyuan Yang, Ziyan Yang, Zhibei Ma, Zhijie Lin, Zhenheng Yang, Dahua Lin, and Lu Jiang. Long context tuning for video generation. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2025.

[51] Danush Kumar Venkatesh, Isabel Funke, Micha Pfeiffer, Fiona R Kolbinger, Hanna Maria Schmeiser, Jürgen Weitz, Marius Distler, and Stefanie Speidel. Mission balance: Generating under-represented class samples using video diffusion models. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2025.

[52] Björn Möller, Zhengyang Li, Malte Stelzer, Thomas Graave, Fabian Bettels, Muaaz Ataya, and Tim Fingscheidt. Openviga: Video generation for automotive driving scenes by streamlining and fine-tuning open source models with public data. *arXiv preprint arXiv:2509.15479*, 2025.

[53] Zijian Song, Sihan Qin, Tianshui Chen, Liang Lin, and Guangrun Wang. Physical autoregressive model for robotic manipulation without action pretraining. *arXiv preprint arXiv:2508.09822*, 2025.

[54] Jing Wang, Ao Ma, Ke Cao, Jun Zheng, Jiasong Feng, Zhanjie Zhang, Wanyuan Pang, and Xiaodan Liang. WISA: World simulator assistant for physics-aware text-to-video generation. In *Proceedings of the Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

[55] Aritra Bhowmik, Denis Korzhenkov, Cees G. M. Snoek, Amirhossein Habibian, and Mohsen Ghafoorian. Moalign: Motion-centric representation alignment for video diffusion models. *arXiv preprint arXiv:2510.19022*, 2025.

[56] Xiangdong Zhang, Jiaqi Liao, Shaofeng Zhang, Fanqing Meng, Xiangpeng Wan, Junchi Yan, and Yu Cheng. Videorepa: Learning physics for video generation through relational alignment with foundation models. *arXiv preprint arXiv:2505.23656*, 2025.

[57] Yu Shang, Xin Zhang, Yinzhou Tang, Lei Jin, Chen Gao, Wei Wu, and Yong Li. Roboscape: Physics-informed embodied world model. *arXiv preprint arXiv:2506.23135*, 2025.

[58] Ke Zhang, Cihan Xiao, Jiacong Xu, Yiqun Mei, and Vishal M. Patel. Think before you diffuse: Llms-guided physics-aware video generation. *arXiv preprint arXiv:2505.21653*, 2025.

[59] Songwei Ge, Seungjun Nah, Guilin Liu, Tyler Poon, Andrew Tao, Bryan Catanzaro, David Jacobs, Jia-Bin Huang, Ming-Yu Liu, and Yogesh Balaji. Preserve your own correlation: A noise prior for video diffusion models. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2023.

[60] Sungwon Hwang, Hyojin Jang, Kinam Kim, Minho Park, and Jaegul Choo. Cross-frame representation alignment for fine-tuning video diffusion models. *arXiv preprint arXiv:2506.09229*, 2025.

[61] Kinam Kim, Junha Hyung, and Jaegul Choo. Temporal in-context fine-tuning with temporal reasoning for versatile control of video diffusion models. In *Proceedings of the Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

[62] Zhen Xing, Qi Dai, Han Hu, Zuxuan Wu, and Yu-Gang Jiang. Simda: Simple diffusion adapter for efficient video generation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[63] Ozgur Kara, Krishna Kumar Singh, Feng Liu, Duygu Ceylan, James M. Rehg, and Tobias Hinz. Shotadapter: Text-to-multi-shot video generation with diffusion models. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[64] Jianzong Wu, Xiangtai Li, Yanhong Zeng, Jiangning Zhang, Qianyu Zhou, Yining Li, Yunhai Tong, and Kai Chen. Motionbooth: Motion-aware customized text-to-video generation. In *Proceedings of the Thirty-eighth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

[65] Youcan Xu, Zhen Wang, Jiaxin Shi, Kexin Li, Feifei Shao, Jun Xiao, Yi Yang, Jun Yu, and Long Chen. Como: Compositional motion customization for text-to-video generation. *arXiv preprint arXiv:2510.23007*, 2025.

[66] Zuhao Yang, Jiahui Zhang, Yingchen Yu, Shijian Lu, and Song Bai. Versatile transition generation with image-to-video diffusion. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2025.

[67] Jingyun Liang, Jingkai Zhou, Shikai Li, Chenjie Cao, Lei Sun, Yichen Qian, Weihua Chen, and Fan Wang. Realismotion: Decomposed human motion control and video generation in the world space. *arXiv preprint arXiv:2508.08588*, 2025.

[68] Jiawei Zhou, Linye Lyu, Zhuotao Tian, Cheng Zhuo, and Yu Li. Safemvdrive: Multi-view safety-critical driving video synthesis in the real world domain. *arXiv preprint arXiv:2505.17727*, 2025.

[69] Quanhao Li, Zhen Xing, Rui Wang, Hui Zhang, Qi Dai, and Zuxuan Wu. Magicmotion: Controllable video generation with dense-to-sparse trajectory guidance. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2025.

[70] Ruihang Chu, Yefei He, Zhekai Chen, Shiwei Zhang, Xiaogang Xu, Bin Xia, Dingdong Wang, Hongwei Yi, Xihui Liu, Hengshuang Zhao, et al. Wan-move: Motion-controllable video generation via latent trajectory guidance. In *Proceedings of the Thirty-Ninth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

[71] Pengxiang Li, Kai Chen, Zhili Liu, Ruiyuan Gao, Lanqing Hong, Dit-Yan Yeung, Huchuan Lu, and Xu Jia. Trackdiffusion: Tracklet-conditioned video generation via diffusion models. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, 2025.

[72] Kwan Yun, Seokhyeon Hong, Chaelin Kim, and Junyong Noh. Anymole: Any character motion in-betweening leveraging video diffusion models. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[73] Yanbo Ding, Xirui Hu, Zhizhi Guo, Chi Zhang, and Yali Wang. Mtvcrafter: 4d motion tokenization for open-world human image animation. In *Proceedings of the Fourteenth International Conference on Learning Representations (ICLR)*, 2026.

[74] Zujin Guo, Size Wu, Zhongang Cai, Wei Li, and Chen Change Loy. Controllable human-centric keyframe interpolation with generative prior. In *Proceedings of the Thirty-Ninth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

[75] Jeongho Kim, Min-Jung Kim, Junsoo Lee, and Jaegul Choo. Tcan: Animating human images with temporally consistent pose guidance using diffusion models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

[76] Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. Dreampose: Fashion image-to-video synthesis via stable diffusion. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2023.

[77] Mingtao Guo, Guanyu Xing, and Yanli Liu. High-fidelity relightable monocular portrait animation with lighting-controllable video diffusion model. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[78] Ryan Burgert, Yuancheng Xu, Wenqi Xian, Oliver Pilarski, Pascal Clausen, Mingming He, Li Ma, Yitong Deng, Lingxiao Li, Mohsen Mousavi, et al. Go-with-the-flow: Motion-controllable video diffusion models using real-time warped noise. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[79] Xi Wang, Robin Courant, Marc Christie, and Vicky Kalogeiton. Akira: Augmentation kit on rays for optical video generation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[80] Sherwin Bahmani, Ivan Skorokhodov, Aliaksandr Siarohin, Willi Menapace, Guocheng Qian, Michael Vasilkovsky, Hsin-Ying Lee, Chaoyang Wang, Jiaxu Zou, Andrea Tagliasacchi, et al. Vd3d: Taming large video diffusion transformers for 3d camera control. In *Proceedings of the Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

[81] Hsin-Ping Huang, Yu-Chuan Su, Deqing Sun, Lu Jiang, Xuhui Jia, Yukun Zhu, and Ming-Hsuan Yang. Fine-grained controllable video generation via object appearance and context. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, 2025.

[82] Zhang Wan, Sheng Tang, Jiawei Wei, Ruize Zhang, and Juan Cao. Dragentity:trajectory guided video generation using entity and positional relationships. In *Proceedings of the 32nd ACM International Conference on Multimedia (ACM MM)*, 2024.

[83] Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Siran Chen, Xiu Li, and Qifeng Chen. Follow your pose: Pose-guided text-to-video generation using pose-free videos. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2024.

[84] Fuchen Long, Zhaofan Qiu, Ting Yao, and Tao Mei. Videostudio: Generating consistent-content and multi-scene videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

[85] Achint Soni, Sreyas Venkataraman, Abhranil Chandra, Sebastian Fischmeister, Percy Liang, Bo Dai, and Sherry Yang. Videoagent: Self-improving video generation. *arXiv preprint arXiv:2410.10076*, 2025.

[86] Wenchuan Wang, Mengqi Huang, Yijing Tu, and Zhendong Mao. Dualreal: Adaptive joint training for lossless identity-motion fusion in video customization. *arXiv preprint arXiv:2505.02192*, 2025.

[87] Shen Sang, Tiancheng Zhi, Tianpei Gu, Jing Liu, and Linjie Luo. Lynx: Towards high-fidelity personalized video generation. *arXiv preprint arXiv:2509.15496*, 2025.

[88] Jingxuan He, Busheng Su, and Finn Wong. Posegen: In-context lora finetuning for pose-controllable long human video generation. *arXiv preprint arXiv:2508.05091*, 2025.

[89] Mirela Ostrek and Justus Thies. Stable video portraits. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

[90] Zunnan Xu, Zhentao Yu, Zixiang Zhou, Jun Zhou, Xiaoyu Jin, Fa-Ting Hong, Xiaozhong Ji, Junwei Zhu, Chengfei Cai, Shiyu Tang, Qin Lin, Xiu Li, and Qinglin Lu. Hunyuanportrait: Implicit condition control for enhanced portrait animation. *arXiv preprint arXiv:2503.18860*, 2025.

[91] Dechao Meng, Steven Xiao, Xindi Zhang, Guangyuan Wang, Peng Zhang, Qi Wang, Bang Zhang, and Liefeng Bo. Mirrorme: Towards realtime and high fidelity audio-driven halfbody animation. *arXiv preprint arXiv:2506.22065*, 2025.

[92] Jiayi Zheng and Xiaodong Cun. Fairygen: Storied cartoon video from a single child-drawn character. In *Proceedings of the SIGGRAPH Asia Conference Papers*, 2025.

[93] Xiang Wang, Shiwei Zhang, Longxiang Tang, Yingya Zhang, Changxin Gao, Yuehuan Wang, and Nong Sang. Unianimate-dit: Human image animation with large-scale video diffusion transformer. *arXiv preprint arXiv:2504.11289*, 2025.

[94] 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.

[95] 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.

[96] Tongchun Zuo, Zaiyu Huang, Shuliang Ning, Ente Lin, Chao Liang, Zerong Zheng, Jianwen Jiang, Yuan Zhang, Mingyuan Gao, and Xin Dong. Dreamvvt: Mastering realistic video virtual try-on in the wild via a stage-wise diffusion transformer framework. *arXiv preprint arXiv:2508.02807*, 2025.

[97] Lizhen Wang, Zhurong Xia, Tianshu Hu, Pengrui Wang, Pengfei Wei, Zerong Zheng, Ming Zhou, Yuan Zhang, and Mingyuan Gao. Dreamactor-h1: High-fidelity human-product demonstration video generation via motion-designed diffusion transformers. *arXiv preprint arXiv:2506.10568*, 2025.

[98] Junchen Zhu, Huan Yang, Wenjing Wang, Huiguo He, Zixi Tuo, Yongsheng Yu, Wen-Huang Cheng, Lianli Gao, Jingkuan Song, Jianlong Fu, et al. Mobilevidfactory: Automatic diffusion-based social media video generation for mobile devices from text. In *Proceedings of the 31st ACM International Conference on Multimedia (ACM MM)*, 2023.

[99] Yang Du, Zhuoran Lin, Kaiqiang Song, Biao Wang, Zhicheng Zheng, Tiezheng Ge, Bo Zheng, and Qin Jin. Vc4vg: Optimizing video captions for text-to-video generation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025.

[100] Bin Xia, Jiyang Liu, Yuechen Zhang, Bohao Peng, Ruihang Chu, Yitong Wang, Xinglong Wu, Bei Yu, and Jiaya Jia. Dreamve: Unified instruction-based image and video editing. *arXiv preprint arXiv:2508.06080*, 2025.

[101] Zhihan Xiao, Lin Liu, Yixin Gao, Xiaopeng Zhang, Haoxuan Che, Songping Mai, and Qi Tian. Lovora: Text-guided and mask-free video object removal and addition with learnable object-aware localization. *arXiv preprint arXiv:2512.02933*, 2025.

[102] Siyoon Jin, Seongchan Kim, Dahyun Chung, Jaeho Lee, Hyunwook Choi, Jisu Nam, Jiyoung Kim, and Seungryong Kim. Matrix: Mask track alignment for interaction-aware video generation. *arXiv preprint arXiv:2510.07310*, 2025.

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

[104] Liyang Chen, Tianxiang Ma, Jiawei Liu, Bingchuan Li, Zhuowei Chen, Lijie Liu, Xu He, Gen Li, Qian He, and Zhiyong Wu. Humo: Human-centric video generation via collaborative multi-modal conditioning. *arXiv preprint arXiv:2509.08519*, 2025.

[105] Hongxiang Zhao, Xingchen Liu, Mutian Xu, Yiming Hao, Weikai Chen, and Xiaoguang Han. Taste-rob: Advancing video generation of task-oriented hand-object interaction for generalizable robotic manipulation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[106] Jiahao Wang, Hualian Sheng, Sijia Cai, Weizhan Zhang, Caixia Yan, Yachuang Feng, Bing Deng, and Jieping Ye. Echoshot: Multi-shot portrait video generation. In *Proceedings of the Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

[107] Ruining Li, Chuanxia Zheng, Christian Rupprecht, and Andrea Vedaldi. Puppet-master: Scaling interactive video generation as a motion prior for part-level dynamics. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2025.

[108] Zhengcong Fei, Hao Jiang, Di Qiu, Baoxuan Gu, Youqiang Zhang, Jiahua Wang, Jialin Bai, Debang Li, Mingyuan Fan, Guibin Chen, et al. Skyreels-audio: Omni audio-conditioned talking portraits in video diffusion transformers. *arXiv preprint arXiv:2506.00830*, 2025.

[109] Xiaoyi Feng, Kaifeng Zou, Caichun Cen, Tao Huang, Hui Guo, Zizhou Huang, Yingli Zhao, Mingqing Zhang, Ziyuan Zheng, Diwei Wang, Yuntao Zou, and Dagang Li. Linkto-anime: A 2d animation optical flow dataset from 3d model rendering. *arXiv preprint arXiv:2506.02733*, 2025.

[110] Shuting Wang, Yunqi Liu, Zixin Yang, Ning Hu, Zhicheng Dou, and Chenyan Xiong. Respond beyond language: A benchmark for video generation in response to realistic user intents. *arXiv preprint arXiv:2506.01689*, 2025.

[111] Hyeonho Jeong, Suhyeon Lee, and Jong Chul Ye. Reangle-a-video: 4d video generation as video-to-video translation. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2025.

[112] Wuyang Li, Wentao Pan, Po-Chien Luan, Yang Gao, and Alexandre Alahi. Stable video infinity: Infinite-length video generation with error recycling. In *Proceedings of the Fourteenth International Conference on Learning Representations (ICLR)*, 2026.

[113] Karan Dalal, Daniel Koceja, Guohao Hussein, Jiarui Xu, Yitu Zhao, Yizhi Song, and Song Han. One-minute video generation with test-time training. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[114] Ryan Po, Eric Ryan Chan, Changan Chen, and Gordon Wetzstein. Bagger: Backwards aggregation for mitigating drift in autoregressive video diffusion models. *arXiv preprint arXiv:2512.12080*, 2025.

[115] Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time. *arXiv preprint arXiv:2509.25161*, 2025.

[116] Yanxiao Sun, Jiafu Wu, Yun Cao, Chengming Xu, Yabiao Wang, Weijian Cao, Donghao Luo, Chengjie Wang, and Yanwei Fu. Swiftvideo: A unified framework for few-step video generation through trajectory-distribution alignment. *arXiv preprint arXiv:2508.06082*, 2025.

[117] Yihong Luo, Tianyang Hu, Jiacheng Sun, Yujun Cai, and Jing Tang. Learning few-step diffusion models by trajectory distribution matching. *arXiv preprint arXiv:2503.06674*, 2025.

[118] Kaiwen Zheng, Yuji Wang, Qianli Ma, Huayu Chen, Jintao Zhang, Yogesh Balaji, Jianfei Chen, Ming-Yu Liu, Jun Zhu, and Qinsheng Zhang. Large scale diffusion distillation via score-regularized continuous-time consistency. *arXiv preprint arXiv:2510.08431*, 2025.

[119] Zhixing Zhang, Yanyu Li, Yushu Wu, Yanwu Xu, Anil Kag, Ivan Skorokhodov, Willi Menapace, Aliaksandr Siarohin, Junli Cao, Dimitris Metaxas, Sergey Tulyakov, and Jian Ren. Sf-v: Single forward video generation model. In *Proceedings of the Thirty-Eighth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

[120] Xinle Cheng, Tianyu He, Jiayi Xu, Junliang Guo, Di He, and Jiang Bian. Playing with transformer at 30+ fps via next-frame diffusion. *arXiv preprint arXiv:2506.01380*, 2025.

[121] Jiaxiang Cheng, Bing Ma, Xuhua Ren, Hongyi Jin, Kai Yu, Peng Zhang, Wenyue Li, Yuan Zhou, Tianxiang Zheng, and Qinglin Lu. Phased one-step adversarial equilibrium for video diffusion models. *arXiv preprint arXiv:2508.21019*, 2025.

[122] Zihan Ding, Chi Jin, Difan Liu, Haitian Zheng, Krishna Kumar Singh, Qiang Zhang, Yan Kang, Zhe Lin, and Yuchen Liu. Dollar: Few-step video generation via distillation and latent reward optimization. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2025.

[123] Jisoo Kim, Wooseok Seo, Junwan Kim, Seungho Park, Sooyeon Park, and Youngjae Yu. Vip: Iterative online preference distillation for efficient video diffusion models. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2025.

[124] Animesh Karnewar, Denis Korzhenkov, Ioannis Lelekas, Noor Fathima, Adil Karjauv, Hanwen Xiong, Vancheeswaran Vaidyanathan, Will Zeng, Rafael Esteves, Tushar Singhal, Fatih Porikli, Mohsen Ghafoorian, and Amirhossein Habibian. Neodragon: Mobile video generation using diffusion transformer. *arXiv preprint arXiv:2511.06055*, 2025.

[125] Hongzhou Zhu, Min Zhao, Guande He, Hang Su, Chongxuan Li, and Jun Zhu. Causal forcing: Autoregressive diffusion distillation done right for high-quality real-time interactive video generation. *arXiv preprint arXiv:2602.02214*, 2026.

[126] Yu Gao, Haoyuan Guo, Tuyen Hoang, Weilin Huang, Lu Jiang, Fangyuan Kong, Huixia Li, Jiashi Li, Liang Li, Xiaojie Li, et al. Seedance 1.0: Exploring the boundaries of video generation models. *arXiv preprint arXiv:2506.09113*, 2025.

[127] Vladimir Arkhipkin, Vladimir Korviakov, Nikolai Gerasimenko, Denis Parkhomenko, Viacheslav Vasilev, Alexey Letunovskiy, Nikolai Vaulin, Maria Kovaleva, Ivan Kirillov, Lev Novitskiy, Denis Koposov, Nikita Kiselev, Alexander Varlamov, Dmitrii Mikhailov, Vladimir Polovnikov, Andrey Shutkin, Julia Agafonova, Ilya Vasiliev, Anastasiia Kargapoltseva, Anna Dmitrienko, Anastasia Maltseva, Anna Averchenkova, Olga Kim, Tatiana Nikulina, and Denis Dimitrov. Kandinsky 5.0: A family of foundation models for image and video generation. *arXiv preprint arXiv:2511.14993*, 2025.

[128] Junhao Cheng, Liang Hou, Xin Tao, and Jing Liao. Video-as-answer: Predict and generate next video event with joint-grpo. *arXiv preprint arXiv:2511.16669*, 2025.

[129] Fan Wu, Jiacheng Wei, Ruibo Li, Yi Xu, Junyou Li, Deheng Ye, and Guosheng Lin. Ic-world: In-context generation for shared world modeling. *arXiv preprint arXiv:2512.02793*, 2025.

[130] Wang Lin, Liyu Jia, Wentao Hu, Kaihang Pan, Zhongqi Yue, Wei Zhao, Jingyuan Chen, Fei Wu, and Hanwang Zhang. Reasoning physical video generation with diffusion timestep tokens via reinforcement learning. *arXiv preprint arXiv:2504.15932*, 2025.

[131] Zhaoqing Wang, Xiaobo Xia, Zhuolin Bie, Jinlin Liu, Dongdong Yu, Jia-Wang Bian, and Changhu Wang. Taming camera-controlled video generation with verifiable geometry reward. *arXiv preprint arXiv:2512.02870*, 2025.

[132] Tianyi Yan, Wencheng Han, Xia Zhou, Xueyang Zhang, Kun Zhan, Cheng-zhong Xu, and Jianbing Shen. Rlgf: Reinforcement learning with geometric feedback for autonomous driving video generation. In *Proceedings of the Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

[133] Rui Li, Yuanzhi Liang, Ziqi Ni, Haibing Huang, Chi Zhang, and Xuelong Li. Growing with the generator: Self-paced grpo for video generation. *arXiv preprint arXiv:2511.19356*, 2025.

[134] Qiushi Yang, Yingjie Chen, Yuan Yao, Yifang Men, Huaizhuo Liu, and Miaomiao Cui. Mcsc: Motion-corrective preference alignment for video generation with self-critic hierarchical reasoning. *arXiv preprint arXiv:2511.22974*, 2025.

[135] Panwang Pan, Jingjing Zhao, Yuchen Lin, Chenguo Lin, Chenxin Li, Hengyu Liu, Tingting Shen, and Yadong MU. Id-crafter: Vlm-grounded online rl for compositional multi-subject video generation. *arXiv preprint arXiv:2511.00511*, 2025.

[136] Xiangyu Meng, Zixian Zhang, Zhenghao Zhang, Junchao Liao, Long Qin, and Weizhi Wang. Identity-grpo: Optimizing multi-human identity-preserving video generation via reinforcement learning. *arXiv preprint arXiv:2510.14256*, 2025.

[137] Liao Shen, Wentao Jiang, Yiran Zhu, Jiahe Li, Tiezheng Ge, Zhiguo Cao, and Bo Zheng. Identity-preserving image-to-video generation via reward-guided optimization. *arXiv preprint arXiv:2510.14255*, 2025.

[138] Sihui Ji, Xi Chen, Xin Tao, Pengfei Wan, and Hengshuang Zhao. Physmaster: Mastering physical representation for video generation via reinforcement learning. *arXiv preprint arXiv:2510.13809*, 2025.

[139] Haoran Cheng, Qide Dong, Liang Peng, Zhizhou Sha, Weiguo Feng, Jinghui Xie, Zhao Song, Shilei Wen, Xiaofei He, and Boxi Wu. Discriminator-free direct preference optimization for video diffusion. *arXiv preprint arXiv:2504.08542*, 2025.

[140] Daoan Zhang, Guangchen Lan, Dong-Jun Han, Wenlin Yao, Xiaoman Pan, Hongming Zhang, Mingxiao Li, Pengcheng Chen, Yu Dong, Christopher Brinton, et al. Seppo: Semi-policy preference optimization for diffusion alignment. *arXiv preprint arXiv:2410.05255*, 2024.

[141] Peiyao Wang, Weining Wang, and Qi Li. Physcorr: Dual-reward dpo for physics-constrained text-to-video generation with automated preference selection. *arXiv preprint arXiv:2511.03997*, 2025.

[142] Wenxu Qian, Chaoyue Wang, Hou Peng, Zhiyu Tan, Hao Li, and Anxiang Zeng. Rdpo: Real data preference optimization for physics consistency video generation. *arXiv preprint arXiv:2506.18655*, 2025.

[143] Hengjia Li, Lifan Jiang, Xi Xiao, Tianyang Wang, Hongwei Yi, Boxi Wu, and Deng Cai. Magicid: Hybrid preference optimization for id-consistent and dynamic-preserved video customization. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2025.

[144] Orest Kupyn, Fabian Manhardt, Federico Tombari, and Christian Rupprecht. Epipolar geometry improves video generation models. *arXiv preprint arXiv:2510.21615*, 2025.

[145] Ziyi Wu, Anil Kag, Ivan Skorokhodov, Willi Menapace, Ashkan Mirzaei, Igor Gilitschenski, Sergey Tulyakov, and Aliaksandr Siarohin. Densedpo: Fine-grained temporal preference optimization for video diffusion models. *arXiv preprint arXiv:2506.03517*, 2025.

[146] Chao Liang, Jianwen Jiang, Wang Liao, Jiaqi Yang, Weihong Zeng, Han Liang, et al. Alignhuman: Improving motion and fidelity via timestep-segment preference optimization for audio-driven human animation. *arXiv preprint arXiv:2506.11144*, 2025.

[147] Harold Haodong Chen, Haojian Huang, Qifeng Chen, Harry Yang, and Ser-Nam Lim. Hierarchical fine-grained preference optimization for physically plausible video generation. In *Proceedings of the Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

[148] Ruiying Liu, Yuanzhi Liang, Haibin Huang, Tianshu Yu, and Chi Zhang. Learning what to trust: Bayesian prior-guided optimization for visual generation. *arXiv preprint arXiv:2511.18919*, 2025.

[149] Bingwen Zhu, Yudong Jiang, Baohan Xu, Siqian Yang, Mingyu Yin, Yidi Wu, Huyang Sun, and Zuxuan Wu. Aligning anime video generation with human feedback. *arXiv preprint arXiv:2504.10044*, 2025.

[150] Fu-Yun Wang, Yunhao Shui, Jingtan Piao, Keqiang Sun, and Hongsheng Li. Diffusion-npo: Negative preference optimization for better preference aligned generation of diffusion models. In *Proceedings of the Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

[151] Yuming Li, Yikai Wang, Yuying Zhu, Zhongyu Zhao, Ming Lu, Qi She, and Shanghang Zhang. Branchgrpo: Stable and efficient grpo with structured branching in diffusion models. *arXiv preprint arXiv:2509.06040*, 2025.

[152] Tahira Kazimi, Connor Dunlop, and Pinar Yanardag. Diverse video generation with determinantal point process-guided policy optimization. *arXiv preprint arXiv:2511.20647*, 2025.

[153] Akhil Agnihotri, Rahul Jain, Deepak Ramachandran, and Zheng Wen. Multi-objective preference optimization: Improving human alignment of generative models. *arXiv preprint arXiv:2505.10892*, 2025.

[154] Kyungmin Lee, Xiahong Li, Qifei Wang, Junfeng He, Junjie Ke, Ming-Hsuan Yang, Irfan Essa, Jinwoo Shin, Feng Yang, and Yinxiao Li. Calibrated multi-preference optimization for aligning diffusion models. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[155] Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Menghan Xia, Xintao Wang, Xiaohong Liu, Fei Yang, Pengfei Wan, Di ZHANG, Kun Gai, Yujiu Yang, and Wanli Ouyang. Improving video generation with human feedback. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

[156] Hengjia Li, Haonan Qiu, Shiwei Zhang, Xiang Wang, Yujie Wei, Zekun Li, Yingya Zhang, Boxi Wu, and Deng Cai. Personalvideo: High id-fidelity video customization without dynamic and semantic degradation. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2025.

[157] Kesen Zhao, Jiaxin Shi, Beier Zhu, Junbao Zhou, Xiaolong Shen, Yuan Zhou, Qianru Sun, and Hanwang Zhang. Real-time motion-controllable autoregressive video diffusion. *arXiv preprint arXiv:2510.08131*, 2025.

[158] Minh-Quan Le, Yuanzhi Zhu, Vicky Kalogeiton, and Dimitris Samaras. What about gravity in video generation? post-training newton's laws with verifiable rewards. *arXiv preprint arXiv:2512.00425*, 2025.

[159] Xiaoyue Mi, Wenqing Yu, Jiesong Lian, Shibo Jie, Ruizhe Zhong, Zijun Liu, Guozhen Zhang, Zixiang Zhou, Zhiyong Xu, Yuan Zhou, et al. Video generation models are good latent reward models. *arXiv preprint arXiv:2511.21541*, 2025.

[160] S. Z. Zhou, Y. B. Wang, J. F. Wu, T. Hu, and J. N. Zhang. A unit enhancement and guidance framework for audio-driven avatar video generation. *arXiv preprint arXiv:2505.03603*, 2025.

[161] June Suk Choi, Kyungmin Lee, Sihyun Yu, Yisol Choi, Jinwoo Shin, and Kimin Lee. Enhancing motion dynamics of image-to-video models via adaptive low-pass guidance. *arXiv preprint arXiv:2506.08456*, 2025.

[162] Jiyang Zheng, Siqi Pan, Yu Yao, Zhaoqing Wang, Dadong Wang, and Tongliang Liu. Aligning what matters: Masked latent adaptation for text-to-audio-video generation. In *Proceedings of the Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

[163] Nikos Spyrou, Athanasios Vlontzos, Paraskevas Pegios, Thomas Melistas, Nefeli Gkouti, Yannis Panagakis, Giorgos Papanastasiou, and Sotirios A. Tsaftaris. Causally steered diffusion for automated video counterfactual generation. *arXiv preprint arXiv:2506.14404*, 2025.

[164] Shuai Tan, Biao Gong, Yujie Wei, Shiwei Zhang, Zhuoxin Liu, Dandan Zheng, Jingdong Chen, Yan Wang, Hao Ouyang, Kecheng Zheng, and Yujun Shen. Synmotion: Semantic-visual adaptation for motion customized video generation. *arXiv preprint arXiv:2506.23690*, 2025.

[165] Ke Zhang, Cihan Xiao, Jiacong Xu, Yiqun Mei, and Vishal M. Patel. Think before you diffuse: Infusing physical rules into video diffusion. *arXiv preprint arXiv:2505.21653*, 2025.

[166] Jianhao Yuan, Xiaofeng Zhang, Felix Friedrich, Nicolas Beltran-Velez, Melissa Hall, Reyhane Askari-Hemmat, Xiaochuang Han, Nicolas Ballas, Michal Drozdzal, and Adriana Romero-Soriano. Inference-time physics alignment of video generative models with latent world models. *arXiv preprint arXiv:2601.10553*, 2026.

[167] Lingling Cai, Kang Zhao, Hangjie Yuan, Xiang Wang, Yingya Zhang, and Kejie Huang. Dfvedit: Conditional delta flow vector for zero-shot video editing. *arXiv preprint arXiv:2506.20967*, 2025.

[168] Yunyang Ge, Xinhua Cheng, Chengshu Zhao, Xianyi He, Shenghai Yuan, Bin Lin, Bin Zhu, and Li Yuan. Flashi2v: Fourier-guided latent shifting prevents conditional image leakage in image-to-video generation. *arXiv preprint arXiv:2509.25187*, 2025.

[169] Xinyao Liao, Xianfang Zeng, Liao Wang, Gang Yu, Guosheng Lin, and Chi Zhang. Motionagent: Fine-grained controllable video generation via motion field agent. In *Proceedings of 2025 International Conference on Computer Vision (ICCV)*, 2025.

[170] Zhelun Shen, Chenming Wu, Junsheng Zhou, Chen Zhao, Kaisiyuan Wang, Hang Zhou, Yingying Li, Haocheng Feng, Wei He, and Jingdong Wang. idit-hoi: Inpainting-based hand object interaction reenactment via video diffusion transformer. *arXiv preprint arXiv:2506.12847*, 2025.

[171] Chengyu Bai, Yuming Li, Zhongyu Zhao, Jintao Chen, Peidong Jia, Qi She, Ming Lu, and Shanghang Zhang. Fastinit: Fast noise initialization for temporally consistent video generation. *arXiv preprint arXiv:2506.16119*, 2025.

[172] Wenhao Wang and Yi Yang. Videoufo: A million-scale user-focused dataset for text-to-video generation. In *Proceedings of the Thirty-Ninth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

[173] Yiming Ju, Jijin Hu, Zhengxiong Luo, Haoge Deng, hanyu Zhao, Li Du, Chengwei Wu, Donglin Hao, Xinlong Wang, and Tengfei Pan. Ci-vid: A coherent interleaved text-video dataset. *arXiv preprint arXiv:2507.01938*, 2025.

[174] Wenhao Wang and Yi Yang. Tip-i2v: A million-scale real text and image prompt dataset for image-to-video generation. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2025.

[175] Dianbing Xi, Jiepeng Wang, Yuanzhi Liang, Xi Qiu, Jialun Liu, Hao Pan, Yuchi Huo, Rui Wang, Haibin Huang, Chi Zhang, and Xuelong Li. Ctrlvdiff: Controllable video generation via unified multimodal video diffusion. *arXiv preprint arXiv:2511.21129*, 2025.

[176] Shenghai Yuan, Xianyi He, Yufan Deng, Yang Ye, Jinfa Huang, Bin Lin, Jiebo Luo, and Li Yuan. Opens2v-nexus: A detailed benchmark and million-scale dataset for subject-to-video generation. In *Proceedings of the Thirty-Ninth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

[177] Kun Liu, Qi Liu, Xinchen Liu, Jie Li, Yongdong Zhang, Jiebo Luo, Xiaodong He, and Wu Liu. Hoigen-1m: A large-scale dataset for human-object interaction video generation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[178] Junfei Xiao, Feng Cheng, Lu Qi, Liangke Gui, Jiepeng Cen, Zhibei Ma, Alan Yuille, and Lu Jiang. Videoauteur: Towards long narrative video generation. *arXiv preprint arXiv:2501.06173*, 2025.

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

[180] Xiaofeng Wang, Kang Zhao, Feng Liu, Jiayu Wang, Guosheng Zhao, Xiaoyi Bao, Zheng Zhu, Yingya Zhang, and Xingang Wang. Egovid-5m: A large-scale video-action dataset for egocentric video generation. In *Proceedings of the Thirty-Ninth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

[181] Qiao Sun, Liujia Yang, Wei Tang, Wei Huang, Kaixin Xu, Yongchao Chen, Mingyu Liu, Jiange Yang, Haoyi Zhu, Yating Wang, Tong He, Yilun Chen, Xili Dai, Nanyang Ye, and Qinying Gu. Learning primitive embodied world models: Towards scalable robotic learning. *arXiv preprint arXiv:2508.20840*, 2025.

[182] Juntao Dai, Tianle Chen, Xuyao Wang, Ziran Yang, Taiye Chen, Jiaming Ji, and Yaodong Yang. Safesora: Towards safety alignment of text2video generation via a human preference dataset. In *Proceedings of the Thirty-Eighth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

[183] Yifeng Gao, Yifan Ding, Hongyu Su, Juncheng Li, Yunhan Zhao, Lin Luo, Zixing Chen, Li Wang, Xin Wang, Yixu Wang, Xingjun Ma, and Yu-Gang Jiang. David-xr1: Detecting ai-generated videos with explainable reasoning. *arXiv preprint arXiv:2506.14827*, 2025.

[184] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[185] Kien T. Pham, Yingqing He, Yazhou Xing, Qifeng Chen, and Long Chen. Spa2v: Harnessing spatial auditory cues for audio-driven spatially-aware video generation. In *Proceedings of the 33rd ACM International Conference on Multimedia (ACM MM)*, 2025.

[186] Hui Han, Siyuan Li, Jiaqi Chen, Yiwen Yuan, Yuling Wu, Chak Tou Leong, Hanwen Du, Junchen Fu, Youhua Li, Jie Zhang, Chi Zhang, Li jia Li, and Yongxin Ni. Video-bench: Human-aligned video generation benchmark. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[187] Xinlong Chen, Yuanxing Zhang, Chongling Rao, Yushuo Guan, Jiaheng Liu, Fuzheng Zhang, Chengru Song, Qiang Liu, Di Zhang, and Tieniu Tan. Vidcapbench: A comprehensive benchmark of video captioning for controllable text-to-video generation. In *Proceedings of the Association for Computational Linguistics (ACL)*, 2025.

[188] Xinran Ling, Chen Zhu, Meiqi Wu, Hangyu Li, Xiaokun Feng, Cundian Yang, Aiming Hao, Jiashu Zhu, Jiahong Wu, and Xiangxiang Chu. Vmbench: A benchmark for perception-aligned video motion generation. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2025.

[189] Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. T2v-compbench: A comprehensive benchmark for compositional text-to-video generation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[190] Jingwei Shi, Zeyu Zhang, Biao Wu, Yanjie Liang, Meng Fang, Ling Chen, and Yang Zhao. PresentAgent: Multimodal agent for presentation video generation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025.

[191] Ruoxuan Zhang, Jidong Gao, Bin Wen, Hongxia Xie, Chenming Zhang, Hong-Han Shuai, and Wen-Huang Cheng. Recipegen: A step-aligned multimodal benchmark for real-world recipe generation. In *Proceedings of the 33rd ACM International Conference on Multimedia (ACM MM)*, 2025.

[192] Xinran Wang, Songyu Xu, Xiangxuan Shan, Yuxuan Zhang, Muxi Diao, Xueyan Duan, Yanhua Huang, Kongming Liang, and Zhanyu Ma. Cinetechbench: A benchmark for cinematographic technique understanding and generation. In *Proceedings of the Thirty-Ninth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

[193] Yuxin Mao, Xuyang Shen, Jing Zhang, Zhen Qin, Jinxing Zhou, Mochu Xiang, Yiran Zhong, and Yuchao Dai. Tavgbench: Benchmarking text to audible-video generation. In *Proceedings of the 32nd ACM International Conference on Multimedia (ACM MM)*, 2024.

[194] Shenghai Yuan, Jinfa Huang, Yongqi Xu, Yaoyang Liu, Shaofeng Zhang, Yujun Shi, Ruijie Zhu, Xinhua Cheng, Jiebo Luo, and Li Yuan. Chronomagic-bench: A benchmark for metamorphic evaluation of text-to-time-lapse video generation. In *Proceedings of the Thirty-Eighth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

[195] Jiaben Chen, Zixin Wang, Ailing Zeng, Yang Fu, Xueyang Yu, Siyuan Cen, Julian Tanke, Yihang Chen, Koichi Saito, Yuki Mitsufuji, and Chuang Gan. Talkcuts: A large-scale dataset for multi-shot human speech video generation. *arXiv preprint arXiv:2510.07249*, 2025.

[196] Junhao Chen, Mingjin Chen, Jianjin Xu, Xiang Li, Junting Dong, Mingze Sun, Puhua Jiang, Hongxiang Li, Yuhang Yang, Hao Zhao, Xiaoxiao Long, and Ruqi Huang. Dancetogether! identity-preserving multi-person interactive video generation. *arXiv preprint arXiv:2505.18078*, 2025.

[197] Weixi Feng, Jiachen Li, Michael Saxon, Tsu-Jui Fu, Wenhu Chen, and William Yang Wang. TC-bench: Benchmarking temporal compositionality in conditional video generation. In *Proceedings of the Association for Computational Linguistics (ACL)*, 2025.

[198] Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. Videophy: Evaluating physical commonsense for video generation. In *Proceedings of the Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

[199] Yiheng Zhang, Zhaofan Qiu, Qi Cai, Yehao Li, Fuchen Long, Yingwei Pan, Ting Yao, and Tao Mei. Identity-preserving video generation challenge. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025.

[200] Yiran Qin, Zhelun Shi, Jiwen Yu, Xijun Wang, Enshen Zhou, Lijun Li, Zhenfei Yin, Xihui Liu, Lu Sheng, Jing Shao, LEI BAI, and Ruimao Zhang. Worldsimbench: Towards video generation models as world simulators. In *Proceedings of the Forty-Second International Conference on Machine Learning (ICML)*, 2025.

[201] Jiahao Wang, Zhenpei Yang, Yijing Bai, Yingwei Li, Yuliang Zou, Bo Sun, Abhijit Kundu, Jose Lezama, Luna Yue Huang, Zehao Zhu, et al. Drive&gen: Co-evaluating end-to-end driving and video generation models. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2025.

[202] Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[203] SP Sharan, Minkyu Choi, Sahil Shah, Harsh Goel, Mohammad Omama, and Sandeep Chinchali. Neuro-symbolic evaluation of text-to-video models using formal verification. In *Proceedings of the Conference onComputer Vision and Pattern Recognition (CVPR)*, 2025.

[204] Nabyl Quignon, Baptiste Chopin, Yaohui Wang, and Antitza Dantcheva. Theval. evaluation framework for talking head video generation. *arXiv preprint arXiv:2511.04520*, 2025.

[205] Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, et al. Videoscore: Building automatic metrics to simulate fine-grained human feedback for video generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

[206] Xuan He, Dongfu Jiang, Ping Nie, Minghao Liu, Zhengxuan Jiang, Mingyi Su, Wentao Ma, Junru Lin, Chun Ye, Yi Lu, et al. Videoscore2: Think before you score in generative video evaluation. *arXiv preprint arXiv:2509.22799*, 2025.

[207] Xiele Wu, Zicheng Zhang, Mingtao Chen, Yixian Liu, Yiming Liu, Shushi Wang, Zhichao Hu, Yuhong Liu, Guangtao Zhai, and Xiaohong Liu. Q-save: Towards scoring and attribution for generated video evaluation. *arXiv preprint arXiv:2511.18825*, 2025.

[208] Jiarui Wang, Huiyu Duan, Guangtao Zhai, Juntong Wang, and Xiongkuo Min. Aigv-assessor: Benchmarking and evaluating the perceptual quality of text-to-video generation with lmm. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[209] Kaisi Guan, Zhengfeng Lai, Yuchong Sun, Peng Zhang, Wei Liu, Kieran Liu, Meng Cao, and Ruihua Song. Etva: Evaluation of text-to-video alignment via fine-grained question generation and answering. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2025.

[210] Xiao Liu and Jiawei Zhang. Aigve-macs: Unified multi-aspect commenting and scoring model for ai-generated video evaluation. *arXiv preprint arXiv:2507.01255*, 2025.

[211] Zhikai Li, Xuewen Liu, Dongrong Joe Fu, Jianquan Li, Qingyi Gu, Kurt Keutzer, and Zhen Dong. K-sort arena: Efficient and reliable benchmarking for generative models via k-wise human preferences. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

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

[213] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.

[214] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

[215] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *Proceedings of the International conference on medical image computing and computer-assisted intervention (MICCAI)*, 2016.

[216] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

[217] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2023.

[218] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

[219] Lijun Yu, José Lezama, Nitesh B. Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Vighnesh Birodkar, Agrim Gupta, Xiuye Gu, Alexander G. Hauptmann, Boqing Gong, Ming-Hsuan Yang, Irfan Essa, David A. Ross, and Lu Jiang. Language model beats diffusion – tokenizer is key to visual generation. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*, 2024.

[220] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 2024.

[221] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *Proceedings of the Eleventh International Conference on Learning Representations (ICLR)*, 2023.

[222] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *Proceedings of the Eleventh International Conference on Learning Representations (ICLR)*, 2023.

[223] Xingchao Liu, Xiwen Zhang, Jianzhu Ma, Jian Peng, et al. Instaflow: One step is enough for high-quality diffusion-based text-to-image generation. In *Proceedings of the Eleventh International Conference on Learning Representations (ICLR)*, 2023.

[224] Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming-Chang Chiu, et al. Videopoet: A large language model for zero-shot video generation. In *Proceedings of the Forty-First International Conference on Machine Learning (ICML)*, 2024.

[225] Hu Yu, Biao Gong, Hangjie Yuan, DanDan Zheng, Weilong Chai, Jingdong Chen, Kecheng Zheng, and Feng Zhao. Videomar: Autoregressive video generation with continuous tokens. In *Proceedings of the Thirty-Ninth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

[226] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[227] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Proceedings of the Thirty-Seventh Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[228] Agrim Gupta, Lijun Yu, Kihyuk Sohn, Xiuye Gu, Meera Hahn, Fei-Fei Li, Irfan Essa, Lu Jiang, and José Lezama. Photorealistic video generation with diffusion models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

[229] Ze Ma, Daquan Zhou, Xue-She Wang, Chun-Hsiao Yeh, Xiuyu Li, Huanrui Yang, Zhen Dong, Kurt Keutzer, and Jiashi Feng. Magic-me: Identity-specific video customized diffusion. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

[230] Eyal Molad, Eliahu Horwitz, Dani Valevski, Alex Rav Acha, Yossi Matias, Yael Pritch, Yaniv Leviathan, and Yedid Hoshen. Dreamix: Video diffusion models are general video editors. *arXiv preprint arXiv:2302.01329*, 2023.

[231] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[232] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *Proceedings of the SIGGRAPH Asia Conference Papers*, 2024.

[233] Shengming Yin, Chenfei Wu, Jian Liang, Jie Shi, Houqiang Li, Gong Ming, and Nan Duan. Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory. *arXiv preprint arXiv:2308.08089*, 2023.

[234] Fanqing Meng, Jiaqi Liao, Xinyu Tan, Wenqi Shao, Quanfeng Lu, Kaipeng Zhang, Yu Cheng, Dianqi Li, Yu Qiao, and Ping Luo. Towards world simulator: Crafting physical commonsense-based benchmark for video generation. In *Proceedings of the Forty-Second International Conference on Machine Learning (ICML)*, 2025.

[235] Fu-Yun Wang, Wenshuo Chen, Guanglu Song, Han-Jia Ye, Yu Liu, and Hongsheng Li. Gen-l-video: Multi-text to long video generation via temporal co-denoising. *arXiv preprint arXiv:2305.18264*, 2023.

[236] Qinghe Wang, Xiaoyu Shi, Baolu Li, Weikang Bian, Quande Liu, Huchuan Lu, Xintao Wang, Pengfei Wan, Kun Gai, and Xu Jia. Multishotmaster: A controllable multi-shot video generation framework. *arXiv preprint arXiv:2512.03041*, 2025.

[237] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2023.

[238] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2023.

[239] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[240] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *International Journal of Computer Vision*, 133(5):3059–3078, 2025.

[241] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[242] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Proceedings of the Thirty-Seventh Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[243] Pablo Acuaviva, Aram Davtyan, Mariam Hassan, Sebastian Stapf, Ahmad Rahimi, Alexandre Alahi, and Paolo Favaro. From generation to generalization: Emergent few-shot learning in video diffusion models. *arXiv preprint arXiv:2506.07280*, 2025.

[244] Kinam Kim, Junha Hyung, and Jaegul Choo. Temporal in-context fine-tuning for versatile control of video diffusion models. In *Proceedings of the Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

[245] Rameen Abdal, Or Patashnik, Ekaterina Deyneka, Hao Chen, Aliaksandr Siarohin, Sergey Tulyakov, Daniel Cohen-Or, and Kfir Aberman. Zero-shot dynamic concept personalization with grid-based lora. In *Proceedings of the SIGGRAPH Asia Conference Papers*, 2025.

[246] Varun Varma Thozhiyoor, Shivam Tripathi, Venkatesh Babu Radhakrishnan, and Anand Bhattad. Objects in generated videos are slower than they appear: Models suffer sub-earth gravity and don't know galileo's principle...for now. *arXiv preprint arXiv:2512.02016*, 2025.

[247] Kerem Çatay, Sedat Bin Vedat, Meftun Akarsu, Enes Kutay Yarkan, İlke Şentürk, Arda Sar, Dafne Ekşioğlu, and Meltem Vargı. Fine-tuning open video generators for cinematic scene synthesis: A small-data pipeline with lora and wan2.1 i2v. *arXiv preprint arXiv:2510.27364*, 2025.

[248] Zeqi Xiao, Wenqi Ouyang, Yifan Zhou, Shuai Yang, Lei Yang, Jianlou Si, and Xingang Pan. Trajectory attention for fine-grained video motion control. In *Proceedings of the Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

[249] Yao-Chih Lee, Zhoutong Zhang, Jiahui Huang, Jui-Hsien Wang, Joon-Young Lee, Jia-Bin Huang, Eli Shechtman, and Zhengqi Li. Generative video motion editing with 3d point tracks. *arXiv preprint arXiv:2512.02015*, 2025.

[250] Ryan Burgert, Charles Herrmann, Forrester Cole, Michael S Ryoo, Neal Wadhwa, Andrey Voynov, and Nataniel Ruiz. Motionv2v: Editing motion in a video. *arXiv preprint arXiv:2511.20640*, 2025.

[251] Mingtao Guo, Guanyu Xing, Yanci Zhang, and Yanli Liu. Navigating large-pose challenge for high-fidelity face reenactment with video diffusion model. *Computers & Graphics*, 2025.

[252] Ming Chen, Liyuan Cui, Wenyuan Zhang, Haoxian Zhang, Yan Zhou, Xiaohan Li, Songlin Tang, Jiwen Liu, Borui Liao, Hejia Chen, et al. Midas: Multimodal interactive digital-human synthesis via real-time autoregressive video generation. *arXiv preprint arXiv:2508.19320*, 2025.

[253] Peng Liu, Xiaoming Ren, Fengkai Liu, Qingsong Xie, Quanlong Zheng, Yanhao Zhang, Haonan Lu, and Yujiu Yang. Dynamic-i2v: Exploring image-to-video generaion models via multimodal llm. *arXiv preprint arXiv:2505.19901*, 2025.

[254] Enrico Pallotta, Sina Mokhtarzadeh Azar, Lars Doorenbos, Serdar Ozsoy, Umar Iqbal, and Juergen Gall. Egocontrol: Controllable egocentric video generation via 3d full-body poses. *arXiv preprint arXiv:2511.18173*, 2025.

[255] Mohammad Mahdi, Yuqian Fu, Nedko Savov, Jiancheng Pan, Danda Pani Paudel, and Luc Van Gool. Exo2egosyn: Unlocking foundation video generation models for exocentric-to-egocentric video synthesis. *arXiv preprint arXiv:2511.20186*, 2025.

[256] Dian Shao, Mingfei Shi, Shengda Xu, Haodong Chen, Yongle Huang, and Binglu Wang. Finephys: Fine-grained human action generation by explicitly incorporating physical laws for effective skeletal guidance. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[257] Shuyuan Tu, Zhen Xing, Xintong Han, Zhi-Qi Cheng, Qi Dai, Chong Luo, Zuxuan Wu, and Yu-Gang Jiang. Stableanimator++: Overcoming pose misalignment and face distortion for human image animation. *arXiv preprint arXiv:2507.15064*, 2025.

[258] Xianghao Kong, Qiaosong Qi, Yuanbin Wang, Anyi Rao, Biaolong Chen, Aixi Zhang, Si Liu, and Hao Jiang. Profashion: Prototype-guided fashion video generation with multiple reference images. *arXiv preprint arXiv:2505.06537*, 2025.

[259] Yuhao Liu, Tengfei Wang, Fang Liu, Zhenwei Wang, and Rynson WH Lau. Shape-for-motion: Precise and consistent video editing with 3d proxy. In *Proceedings of the SIGGRAPH Asia Conference Papers*, 2025.

[260] Yunpeng Bai, Shaoheng Fang, Chaohui Yu, Fan Wang, and Qixing Huang. Geovideo: Introducing geometric regularization into video generation model. *arXiv preprint arXiv:2512.03453*, 2025.

[261] Zirui Pan, Xin Wang, Yipeng Zhang, Hong Chen, Kwan Man Cheng, Yaofei Wu, and Wenwu Zhu. Modular-cam: Modular dynamic camera-view video generation with llm. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2025.

[262] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2025.

[263] Soon Yau Cheong, Duygu Ceylan, Armin Mustafa, Andrew Gilbert, and Chun-Hao Paul Huang. Boosting camera motion control for video diffusion transformers. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2025.

[264] Duolikun Danier, Ge Gao, Steven McDonagh, Changjian Li, Hakan Bilen, and Oisin Mac Aodha. View-consistent diffusion representations for 3d-consistent video generation. *arXiv preprint arXiv:2511.18991*, 2025.

[265] Teng Li, Guangcong Zheng, Rui Jiang, Shuigen Zhan, Tao Wu, Yehao Lu, Yining Lin, Chuanyun Deng, Yepan Xiong, Min Chen, Lin Cheng, and Xi Li. Realcam-i2v: Real-world image-to-video generation with interactive complex camera control. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2025.

[266] Hyeonho Jeong, Suhyeon Lee, and Jong Chul Ye. Reangle-a-video: 4d video generation as video-to-video translation. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2025.

[267] Yue Ma, Kunyu Feng, Xinhua Zhang, Hongyu Liu, David Junhao Zhang, Jinbo Xing, Yinhan Zhang, Ayden Yang, Zeyu Wang, and Qifeng Chen. Follow-your-creation: Empowering 4d creation through video inpainting. *arXiv preprint arXiv:2506.04590*, 2025.

[268] Shihan Cheng, Nilesh Kulkarni, David Hyde, and Dmitriy Smirnov. Less is more: Data-efficient adaptation for controllable text-to-video generation. *arXiv preprint arXiv:2511.17844*, 2025.

[269] Haoyu Wu, Diankun Wu, Tianyu He, Junliang Guo, Yang Ye, Yueqi Duan, and Jiang Bian. Geometry forcing: Marrying video diffusion and 3d representation for consistent world modeling. In *Proceedings of the Fourteenth International Conference on Learning Representations (ICLR)*, 2026.

[270] Zun Wang, Jaemin Cho, Jialu Li, Han Lin, Jaehong Yoon, Yue Zhang, and Mohit Bansal. Epic: Efficient video camera control learning with precise anchor-video guidance. *arXiv preprint arXiv:2505.21876*, 2025.

[271] Pooja Guhan, Divya Kothandaraman, Geonsun Lee, Tsung-Wei Huang, Guan-Ming Su, and Dinesh Manocha. I want it that way! specifying nuanced camera motions in video editing. *arXiv preprint arXiv:2504.09472*, 2025.

[272] Dong-Yu Chen, Yixin Guo, Shuojin Yang, Tai-Jiang Mu, and Shi-Min Hu. Beyond inpainting: Unleash 3d understanding for precise camera-controlled video generation. *arXiv preprint arXiv:2601.10214*, 2026.

[273] Xingchang Huang, Ashish Kumar Singh, Florian Dubost, Cristina Nader Vasconcelos, Sakar Khattar, Liang Shi, Christian Theobalt, Cengiz Oztireli, and Gurprit Singh. Restereo: Diffusion stereo video generation and restoration. *arXiv preprint arXiv:2506.06023*, 2025.

[274] Peng Hu, Yu Gu, Liang Luo, and Fuji Ren. Ssg-dit: A spatial signal guided framework for controllable video generation. *arXiv preprint arXiv:2508.17062*, 2025.

[275] Hong Chen, Xin Wang, Yipeng Zhang, Yuwei Zhou, Zeyang Zhang, Siao Tang, and Wenwu Zhu. Disenstudio: Customized multi-subject text-to-video generation with disentangled spatial control. In *Proceedings of the 32nd ACM International Conference on Multimedia (ACM MM)*, 2024.

[276] Fangyuan Mao, Aiming Hao, Jintao Chen, Dongxia Liu, Xiaokun Feng, Jiashu Zhu, Meiqi Wu, Chubin Chen, Jiahong Wu, and Xiangxiang Chu. Omni-effects: Unified and spatially-controllable visual effects generation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2026.

[277] Qing Chang, Yao-Xiang Ding, and Kun Zhou. Enhancing identity-deformation disentanglement in stylegan for one-shot face video re-enactment. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence (AAAI)*, 2025.

[278] Rui Xie, Yinhong Liu, Penghao Zhou, Chen Zhao, Jun Zhou, Kai Zhang, Zhenyu Zhang, Jian Yang, Zhenheng Yang, and Ying Tai. Star: Spatial-temporal augmentation with text-to-video models for real-world video super-resolution. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2025.

[279] Haitao Zhou, Chuang Wang, Rui Nie, Jinlin Liu, Dongdong Yu, Qian Yu, and Changhu Wang. Trackgo: a flexible and efficient method for controllable video generation. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence (AAAI)*, 2025.

[280] Weitao Wang, Zichen Wang, Hongdeng Shen, Yulei Lu, Xirui Fan, Suhui Wu, Jun Zhang, Haoqian Wang, and Hao Zhang. Dreamswapv: Mask-guided subject swapping for any customized video editing. In *Proceedings of the Fourteenth International Conference on Learning Representations (ICLR)*, 2026.

[281] Ziyao Huang, Zixiang Zhou, Juan Cao, Yifeng Ma, Yi Chen, Zejing Rao, Zhiyong Xu, Hongmei Wang, Qin Lin, Yuan Zhou, Qinglin Lu, and Fan Tang. Hunyuanvideo-homa: Generic human-object interaction in multimodal driven human animation. *arXiv preprint arXiv:2506.08797*, 2025.

[282] Yufan Deng, Yuanyang Yin, Xun Guo, Yizhi Wang, Jacob Zhiyuan Fang, Shenghai Yuan, Yiding Yang, Angtian Wang, Bo Liu, Haibin Huang, and Chongyang Ma. Magref: Masked guidance for any-reference video generation with subject disentanglement. *arXiv preprint arXiv:2505.23742*, 2025.

[283] Animesh Karnewar, Denis Korzhenkov, Ioannis Lelekas, Adil Karjauv, Noor Fathima, Hanwen Xiong, Vancheeswaran Vaidyanathan, Will Zeng, Rafael Esteves, Tushar Singhal, Fatih Porikli, Mohsen Ghafoorian, and Amirhossein Habibian. Neodragon: Mobile video generation using diffusion transformer. In *Proceedings of the Fourteenth International Conference on Learning Representations (ICLR)*, 2026.

[284] Lei lei Li, Jianwu Fang, Junbin Xiao, Shanmin Pang, Hongkai Yu, Chen Lv, Jianru Xue, and Tat-Seng Chua. Causal-entity reflected egocentric traffic accident video synthesis. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2025.

[285] Xiangjun Zhang, Litong Gong, Yinglin Zheng, Yansong Liu, Wentao Jiang, Mingyi Xu, Biao Wang, Tiezheng Ge, and Ming Zeng. Rise-t2v: Rephrasing and injecting semantics with llm for expansive text-to-video generation. *arXiv preprint arXiv:2511.04317*, 2025.

[286] Lvmin Zhang, Shengqu Cai, Muyang Li, Gordon Wetzstein, and Maneesh Agrawala. Frame context packing and drift prevention in next-frame-prediction video diffusion models. In *Proceedings of the Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

[287] Guosheng Zhao, Xiaofeng Wang, Zheng Zhu, Xinze Chen, Guan Huang, Xiaoyi Bao, and Xingang Wang. Drivedreamer-2: Llm-enhanced world models for diverse driving video generation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2025.

[288] Xunzhi Xiang, Yabo Chen, Guiyu Zhang, Zhongyu Wang, Zhe Gao, Quanming Xiang, Gonghu Shang, Junqi Liu, Haibin Huang, Yang Gao, Chi Zhang, Qi Fan, and Xuelong Li. Macro-from-micro planning for high-quality and parallelized autoregressive long video generation. *arXiv preprint arXiv:2508.03334*, 2025.

[289] Yuanyang Yin, Yufan Deng, Shenghai Yuan, Kaipeng Zhang, Xiao Yang, and Feng Zhao. Focal guidance: Unlocking controllability from semantic-weak layers in video diffusion models. *arXiv preprint arXiv:2601.07287*, 2026.

[290] Tingting Liao, Chongjian Ge, Guangyi Liu, Hao Li, and Yi Zhou. Character mixing for video generation. *arXiv preprint arXiv:2510.05093*, 2025.

[291] Yuhui Deng, Yuqin Lu, Yangyang Xu, Yongwei Nie, and Shengfeng He. Occlusion-insensitive talking head video generation via facelet compensation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2025.

[292] Lin Zhang, Zefan Cai, Yufan Zhou, Shentong Mo, Jinhong Lin, Cheng-En Wu, Yibing Wei, Yijing Zhang, Ruiyi Zhang, Wen Xiao, et al. Scaling up audio-synchronized visual animation: An efficient training paradigm. *arXiv preprint arXiv:2508.03955*, 2025.

[293] Yuwei Guo, Ceyuan Yang, Anyi Rao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Sparsectrl: Adding sparse controls to text-to-video diffusion models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

[294] Han Lin, Abhay Zala, Jaemin Cho, and Mohit Bansal. Videodirectorgpt: Consistent multi-scene video generation via llm-guided planning. In *Proceedings of the Conference on Language Modeling (COLM)*, 2024.

[295] Guosheng Zhao, Xiaofeng Wang, Zheng Zhu, Xinze Chen, Guan Huang, Xiaoyi Bao, and Xingang Wang. Drivedreamer-2: Llm-enhanced world models for diverse driving video generation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2025.

[296] Hmrishav Bandyopadhyay and Yi-Zhe Song. Flipsketch: Flipping static drawings to text-guided sketch animations. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[297] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[298] Wanquan Feng, Tianhao Qi, Jiawei Liu, Mingzhen Sun, Pengqi Tu, Tianxiang Ma, Fei Dai, Songtao Zhao, Siyu Zhou, and Qian He. I2vcontrol: Disentangled and unified video motion synthesis control. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2025.

[299] Tao Wu, Yong Zhang, Xintao Wang, Xianpan Zhou, Guangcong Zheng, Zhongang Qi, Ying Shan, and Xi Li. Customcrafter: Customized video generation with preserving motion and concept composition abilities. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2025.

[300] David Junhao Zhang, Roni Paiss, Shiran Zada, Nikhil Karnad, David E Jacobs, Yael Pritch, Inbar Mosseri, Mike Zheng Shou, Neal Wadhwa, and Nataniel Ruiz. Recapture: Generative video camera controls for user-provided videos using masked video fine-tuning. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[301] Yue Ma, Xiaodong Cun, Sen Liang, Jinbo Xing, Yingqing He, Chenyang Qi, Siran Chen, and Qifeng Chen. Magicstick: Controllable video editing via control handle transformations. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, 2025.

[302] Jie Tian, Xiaoye Qu, Zhenyi Lu, Wei Wei, Sichen Liu, and Yu Cheng. Extrapolating and decoupling image-to-video generation models: Motion modeling is easier than you think. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[303] Weikang Bian, Zhaoyang Huang, Xiaoyu Shi, Yijin Li, Fu-Yun Wang, and Hongsheng Li. Gs-dit: Advancing video generation with pseudo 4d gaussian fields through efficient dense 3d point tracking. *arXiv preprint arXiv:2501.02690*, 2025.

[304] Meiqi Wu, Bingze Song, Ruimin Lin, Chen Zhu, Xiaokun Feng, Jiahong Wu, Xiangxiang Chu, and Kaiqi Huang. Latent temporal discrepancy as motion prior: A loss-weighting strategy for dynamic fidelity in t2v. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2026.

[305] Ying Guo, Qijun Gan, Yifu Zhang, Jinlai Liu, Yifei Hu, Pan Xie, Dongjun Qian, Yu Zhang, Ruiqi Li, Yuqi Zhang, et al. Alive: Animate your world with lifelike audio-video generation. *arXiv preprint arXiv:2602.08682*, 2026.

[306] Ethan Chern, Zhulin Hu, Bohao Tang, Jiadi Su, Steffi Chern, Zhijie Deng, and Pengfei Liu. Livetalk: Real-time multimodal interactive video diffusion via improved on-policy distillation. *arXiv preprint arXiv:2512.23576*, 2025.

[307] Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *arXiv preprint arXiv:2506.08009*, 2025.

[308] Yanzuo Lu, Yuxi Ren, Xin Xia, Shanchuan Lin, Xing Wang, Xuefeng Xiao, Andy J Ma, Xiaohua Xie, and Jian-Huang Lai. Adversarial distribution matching for diffusion distillation towards efficient image and video synthesis. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2025.

[309] Harold Haodong Chen, Disen Lan, Wen-Jie Shu, Qingyang Liu, Zihan Wang, Sirui Chen, Wenkai Cheng, Kanghao Chen, Hongfei Zhang, Zixin Zhang, Rongjin Guo, Yu Cheng, and Ying-Cong Chen. Tivibench: Benchmarking think-in-video reasoning for video generative models. *arXiv preprint arXiv:2511.13704*, 2025.

[310] Rang Meng, Weipeng Wu, Yingjie Yin, Yuming Li, and Chenguang Ma. Echotorrent: Towards swift, sustained, and streaming multi-modal video generation. *arXiv preprint 2602.13669*, 2026.

[311] Lichen Bai, Zikai Zhou, Shitong Shao, Wenliang Zhong, Shuo Yang, Shuo Chen, Bojun Chen, and Zeke Xie. Optimizing few-step generation with adaptive matching distillation. *arXiv preprint arXiv:2602.07345*, 2026.

[312] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Proceedings of the Thirty-Sixth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[313] Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 1952.

[314] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

[315] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[316] Ruizhe Zhong, Jiesong Lian, Xiaoyue Mi, Zixiang Zhou, Yuan Zhou, Qinglin Lu, and Junchi Yan. Euphonium: Steering video flow matching via process reward gradient guided stochastic dynamics. *arXiv preprint arXiv:2602.04928*, 2026.

[317] Jin Wang, Jianxiang Lu, Guangzheng Xu, Comi Chen, Haoyu Yang, Linqing Wang, Peng Chen, Mingtao Chen, Zhichao Hu, Longhuang Wu, et al. Tagrpo: Boosting grpo on image-to-video generation with direct trajectory alignment. *arXiv preprint arXiv:2601.05729*, 2026.

[318] Yang Ye, Tianyu He, Shuo Yang, and Jiang Bian. Reinforcement learning with inverse rewards for world model post-training. *arXiv preprint arXiv:2509.23958*, 2025.

[319] Jiawei Liu, Junqiao Li, Jiangfan Deng, Gen Li, Siyu Zhou, Zetao Fang, Shanshan Lao, Zengde Deng, Jianing Zhu, Tingting Ma, et al. Dreamontage: Arbitrary frame-guided one-shot video generation. *arXiv preprint arXiv:2512.21252*, 2025.

[320] Yuanhao Cai, Kunpeng Li, Menglin Jia, Jialiang Wang, Junzhe Sun, Feng Liang, Weifeng Chen, Felix Juefei-Xu, Chu Wang, Ali Thabet, et al. Phygdpo: Physics-aware groupwise direct preference optimization for physically consistent text-to-video generation. *arXiv preprint arXiv:2512.24551*, 2025.

[321] Jiachen Li, Weixi Feng, Tsu-Jui Fu, Xinyi Wang, Sugato Basu, Wenhu Chen, and William Yang Wang. T2v-turbo: Breaking the quality bottleneck of video consistency model with mixed reward feedback. In *Proceedings of the Thirty-Eighth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

[322] Mihir Prabhudesai, Russell Mendonca, Zheyang Qin, Katerina Fragkiadaki, and Deepak Pathak. Video diffusion alignment via reward gradients. *arXiv preprint arXiv:2407.08737*, 2024.

[323] Kumar Ashutosh, XuDong Wang, Xi Yin, Kristen Grauman, Adam Polyak, Ishan Misra, and Rohit Girdhar. Human detectors are surprisingly powerful reward models. *arXiv preprint arXiv:2601.14037*, 2026.

[324] Qiyuan Zhang, Biao Gong, Shuai Tan, Zheng Zhang, Yujun Shen, Xing Zhu, Yuyuan Li, Kelu Yao, Chunhua Shen, and Changqing Zou. Physrvg: Physics-aware unified reinforcement learning for video generative models. *arXiv preprint arXiv:2601.11087*, 2026.

[325] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. *arXiv preprint arXiv:2302.03011*, 2023.

[326] Hanwen Shen, Jiajie Lu, Yupeng Cao, and Xiaonan Yang. Enhancing scene transition awareness in video generation via post-training. In *Proceedings of the 14th International Joint Conference on Natural Language Processing (IJCNLP)*, 2025.

[327] Zhiyu Tan, Junyan Wang, Hao Yang, Luozheng Qin, Hesen Chen, Qiang Zhou, and Hao Li. Raccoon: Multi-stage diffusion training with coarse-to-fine curating videos. *arXiv preprint arXiv:2502.21314*, 2025.

[328] Shanchuan Lin, Xin Xia, Yuxi Ren, Ceyuan Yang, Xuefeng Xiao, and Lu Jiang. Diffusion adversarial post-training for one-step video generation. *arXiv preprint arXiv:2501.08316*, 2025.

[329] Xincheng Shuai, Henghui Ding, Zhenyuan Qin, Hao Luo, Xingjun Ma, and Dacheng Tao. Free-form motion control: Controlling the 6d poses of camera and objects in video generation. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2025.

[330] Bingyi Kang, Yang Yue, Rui Lu, Zhijie Lin, Yang Zhao, Kaixin Wang, Gao Huang, and Jiashi Feng. How far is video generation from world model: A physical law perspective. In *Proceedings of the Forty-Second International Conference on Machine Learning (ICML)*, 2025.

[331] Zhun Mou, Bin Xia, Zhengchao Huang, Wenming Yang, and Jiaya Jia. GRADEO: Towards human-like evaluation for text-to-video generation via multi-step reasoning. In *Proceedings of Machine Learning Research (PMLR)*, 2025.

[332] Jing Lin, Ruisi Wang, Junzhe Lu, Ziqi Huang, Guorui Song, Ailing Zeng, Xian Liu, Chen Wei, Wanqi Yin, Qingping Sun, Zhongang Cai, Lei Yang, and Ziwei Liu. The quest for generalizable motion generation: Data, model, and evaluation. In *Proceedings of the Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

[333] Yuanxin Liu, Lei Li, Shuhuai Ren, Rundong Gao, Shicheng Li, Sishuo Chen, Xu Sun, and Lu Hou. Fetv: A benchmark for fine-grained evaluation of open-domain text-to-video generation. In *Proceedings of the Thirty-Seventh Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[334] Emanuele Bugliarello, Hernan Moraldo, Ruben Villegas, Mohammad Babaeizadeh, Mohammad Taghi Saffar, Han Zhang, Dumitru Erhan, Vittorio Ferrari, Pieter-Jan Kindermans, and Paul Voigtlaender. Storybench: A multifaceted benchmark for continuous story visualization. In *Proceedings of the Thirty-Seventh Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[335] Yibo Miao, Yifan Zhu, Yinpeng Dong, Lijia Yu, Jun Zhu, and Xiao-Shan Gao. T2vsafetybench: Evaluating the safety of text-to-video generative models. In *Proceedings of the Thirty-Eighth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

[336] Qingyu Shi, Jianzong Wu, Jinbin Bai, Jiangning Zhang, Lu Qi, Yunhai Tong, and Xiangtai Li. Decouple and track: Benchmarking and improving video diffusion transformers for motion transfer. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2025.

[337] Minghan Li, Chenxi Xie, Yichen Wu, Lei Zhang, and Mengyu Wang. Five-bench: A fine-grained video editing benchmark for evaluating emerging diffusion and rectified flow models. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2025.

[338] Yiping Wang, Xuehai He, Kuan Wang, Luyao Ma, Jianwei Yang, Shuohang Wang, Simon Shaolei Du, and Yelong Shen. Is your world simulator a good story presenter? a consecutive events-based benchmark for future long video generation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[339] Haibo Tong, Zhaoyang Wang, Zhaorun Chen, Haonian Ji, Shi Qiu, Siwei Han, Kexin Geng, Zhongkai Xue, Yiyang Zhou, Peng Xia, et al. Mj-video: Fine-grained benchmarking and rewarding video preferences in video generation. In *Proceedings of the Thirty-Ninth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

[340] Duomin Wang, Wei Zuo, Aojie Li, Ling-Hao Chen, Xinyao Liao, Deyu Zhou, Zixin Yin, Xili Dai, Daxin Jiang, and Gang Yu. Universe-1: Unified audio-video generation via stitching of experts. *arXiv preprint arXiv:2509.06155*, 2025.

[341] Ziqiao Peng, Jiwen Liu, Haoxian Zhang, Xiaoqiang Liu, Songlin Tang, Pengfei Wan, Di Zhang, Hongyan Liu, and Jun He. Omnisync: Towards universal lip synchronization via diffusion transformers. *arXiv preprint arXiv:2505.21448*, 2025.

[342] Sijing Wu, Yunhao Li, Huiyu Duan, Yanwei Jiang, Yucheng Zhu, and Guangtao Zhai. Hveval: Towards unified evaluation of human-centric video generation and understanding. In *Proceedings of the 33rd ACM International Conference on Multimedia (ACM MM)*, 2025.

[343] 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 *Proceedings of the Thirty-First International Conference on Neural Information Processing Systems (NeurIPS)*, 2017.

[344] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 2004.

[345] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

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

[347] Jie Wu, Yu Gao, Zilyu Ye, Ming Li, Liang Li, Hanzhong Guo, Jie Liu, Zeyue Xue, Xiaoxia Hou, Wei Liu, et al. Rewarddance: Reward scaling in visual generation. *arXiv preprint arXiv:2509.08826*, 2025.

[348] Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

[349] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[350] Jianxiong Gao, Zhaoxi Chen, Xian Liu, Jianfeng Feng, Chenyang Si, Yanwei Fu, Yu Qiao, and Ziwei Liu. Longvie: Multimodal-guided controllable ultra-long video generation. *arXiv preprint arXiv:2508.03694*, 2025.