SIFT: Self-Imagination Fine-Tuning for Physically Plausible Motion in Video Diffusion Models

Ruoyu Wang[1], Jialun Liu[2⋆†], Huayang Huang[1], Haibin Huang[2], Jiepeng Wang[2], Chi Zhang[2], Xuelong Li[2], and Yu Wu[3†]

[1] School of Computer Science, Wuhan University
[2] Institute of Artificial Intelligence, China Telecom (TeleAI)
[3] School of Artificial Intelligence, Wuhan University

Abstract. Recent advances in video diffusion models have greatly improved visual fidelity, yet their generated motions often violate physical plausibility. We observe a common kinematic failure, "motion entanglement", the unintended coupling of independent motion sources, such as camera movement and object motion. We identify that this issue stems from data bias and the reconstruction-based training design of diffusion models. Training on noisy videos that still retain coarse motion cues inadvertently encourages the model to replicate existing motion without an incentive to learn how to model kinematically-grounded motions. To address this, we propose a Self-Imagination Fine-Tuning (SIFT) paradigm, which enables the model to learn from its own generated videos rather than directly reconstructing real ones, breaking the reconstruction shortcut. We further employ motion-aware discriminative supervision and a progressive hard-case replay strategy to stabilize and accelerate learning. By leveraging freely-generated text prompts, our method can densely cover a broad motion space, including rare or finely-disentangled scenarios that would be costly to collect as video data. Extensive experiments demonstrate that our approach substantially improves the physical realism, motion disentanglement, and controllability of generated videos.

Keywords: Text-to-Video Generation · Motion Entanglement · Video Diffusion Models · Self-Imagination Fine-Tuning

1 Introduction

While Video Diffusion Models (VDMs) [21,27,35,37,54,66,71] have achieved remarkable visual fidelity and semantic consistency, the physical plausibility of their generated motions remains a fundamental challenge. Existing efforts [3,25,33,51] largely center on *dynamics*, such as gravity, collisions, and fluid mechanics. In this work, we study a complementary yet highly visible failure from a *kinematics* perspective: *Motion Entanglement*. This phenomenon refers to the model's

⋆ Project leader
† Corresponding authors

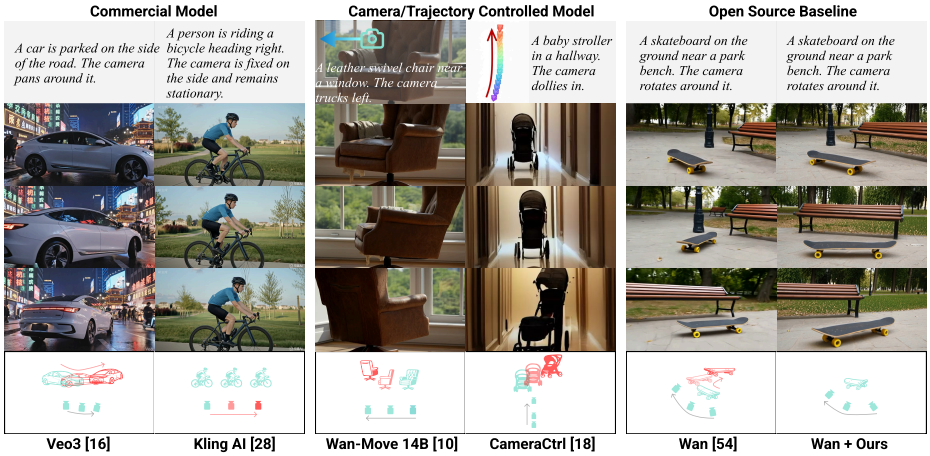

Fig. 1: Illustration of Motion Entanglement. The first row shows the input conditions, and the last row visualizes the generated trajectories where red indicates physically implausible relative motion. Note that camera-control methods take the camera trajectory and the first frame as extra conditions, while the others are purely text-to-video.

inability to independently control and disentangle motions originating from distinct sources, such as camera movement versus object motion. Fundamentally, this reflects a failure to preserve independent reference frames and accurately model relative motion. As illustrated in Fig. 1, this issue is pervasive across commercial models (Veo3 [16], Kling AI [28]), open-source models (Wan [54]), and even models with explicit camera (CameraCtrl [18]) or trajectory (Wan-Move [10]) controls. For example, when the camera is supposed to orbit around a stationary object, the generated video frequently shows the object drifting as well. Conversely, if the object moves but the camera should remain fixed, the camera often unintentionally tracks the object. Although individual frames look realistic, their kinematically flawed relative trajectories reveal a profound lack of structural motion understanding in existing VDMs. Beyond perceptual realism, such kinematic entanglement severely undermines the utility of VDMs as world simulators for downstream applications like autonomous driving, where strict decoupling of ego-motion and environmental dynamics is essential.

Why does motion entanglement arise? We attribute this phenomenon to two main factors. First, *a pronounced data-induced bias exists*: real-world videos are largely composed of scenes where camera and object motions co-occur, and large-scale datasets [2, 9, 24, 41, 56, 58] lack explicit annotations that distinguish their contributions. Consequently, models learn spurious statistical correlations, treating these independent kinematic variables as inherently linked. Second, *the training paradigm of VDMs inherently limits learning of kinematic reasoning.* VDMs are trained to reconstruct a clean video from a noisy counterpart using pixel-level objectives like MSE loss. However, even a heavily noised video still preserves substantial residual structural and temporal cues [22, 39, 57, 61].

This creates a "reconstruction shortcut": the model learns to replicate preexisting motion patterns inherited from the noisy input rather than inferring kinematically correct dynamics from the textual prompt from scratch. This renders the models inherently insensitive to motion semantics. Meanwhile, the pixel-level reconstruction objective further biases learning toward appearance fidelity. Each pixel's temporal displacement can stem from both camera-induced scene movement and object-intrinsic motion, but the loss only enforces accurate RGB matching. Thus, the model learns to replicate the overall visual distribution rather than disentangling the underlying kinematic reference frames.

Why traditional fine-tuning fails? A seemingly straightforward solution would be to perform supervised fine-tune (SFT) on additional motion-decoupled video data. However, such data is extremely scarce, and curating a dataset to cover the full diversity of real-world motion scenarios would require prohibitive data engineering efforts. More fundamentally, SFT inherits the same flaws as pretraining: the denoising shortcut and pixel-level reconstruction bias remain intact. This demonstrates that the motion is being *inherited* from the input data, not *inferred* from the semantics of the prompt, rendering standard SFT ineffective for teaching true motion reasoning.

To overcome these limitations, we introduce a novel paradigm shift from reconstruction-based training to imagination-based learning. Our Self-Imagination Fine-Tuning (SIFT) framework aims to enable the model to learn to understand, correct, and model motion from its own generated videos, rather than merely copying motion patterns from real videos that are inherently physically correct. To achieve this, we discard real video inputs and force the model to generate videos from randomly initialized noise, guided solely by textual prompts. This completely removes the input reconstruction shortcut, compelling the model to derive kinematic relationships strictly from semantic intent. Additionally, we replace the pixel-level reconstruction objective with motion-aware discriminative supervision to refine its motion generation toward kinematically consistent dynamics. To further stabilize training and improve learning efficiency, we introduce a progressive hard-case replay strategy that gradually exposes the model to increasingly difficult self-imagined samples. Crucially, the text prompts can be freely generated by large language models to densely cover diverse and even rare motion scenarios without requiring paired or motion-decoupled video data. Experiments demonstrate that our method significantly improves the physical plausibility and disentanglement of generated motion.

In summary, our contributions are as follows:

- We identify and analyze the Motion Entanglement in video diffusion models, revealing how their current training pipeline limits learning kinematically-grounded motion.
- We propose a novel self-imagination training paradigm that removes the reconstruction shortcut and enables learning of physically consistent, disentangled motion from textual descriptions.

– We conduct extensive experiments demonstrating that our framework significantly improves motion realism and relative dynamics, without requiring motion-decoupled video-text data.

2 Related Work

Text-to-Video Generation. Text-to-video (T2V) generation has undergone rapid development in recent years, with diffusion models emerging as the predominant backbone. Early video diffusion models [4, 7, 8, 17, 19, 22, 26] typically "inflate" 2D generative models into the video domain by directly adding temporal attention layers to the U-Net architecture of pre-trained text-to-image models (T2I) [42, 45–47] to achieve inter-frame coherence. Despite their simplicity and efficiency, these methods often suffer from limited temporal modeling capacity, leading to issues like frame flicker, motion jitter, or inconsistent object evolution over time. A major shift came with the advent of Diffusion Transformers (DiTs) [43], which replaced the U-Net's backbone with a full transformer architecture to enable 3D full attention. Benefiting from this architectural innovation and advances in large-scale training, recent DiT-based T2V models [21, 27, 35, 37, 38, 54, 66, 71] have achieved remarkable breakthroughs in visual-fidelity, generating high-resolution videos with detailed textures, accurate semantic alignment to text prompts, and improved temporal smoothness. Nevertheless, current T2V methods often struggle with physically consistent motion dynamics. This limitation highlights that visual fidelity and semantic alignment alone are insufficient for generating truly realistic and controllable videos; a deeper, physically-grounded motion understanding is required.

Physics-Grounded Video Generation. Although video generation models have made significant progress in visual quality, a fundamental challenge is becoming increasingly prominent: the generated dynamic content often lacks physical rationality [3, 25, 40, 70]. The research community is trying to solve this problem from multiple directions, though most existing efforts primarily focus on *dynamics*. (1) Simulation-driven generation [33, 51, 62, 69] integrates explicit physical simulation (rigid-body, elastodynamics, material modeling) with generative video rendering. For instance, PhysGen [33] simulates the rigid-body motion and interactions of each instance based on Newton's Laws and physical constraints. (2) LLM-guided or semantic-planning approaches [29, 36, 63, 65, 68] leverage large language models or motion planning modules to script object trajectories or physical interactions, which are then used to condition the video generation model. While both directions mark important progress toward physical realism, they typically inject external physical knowledge or simulate dynamics beforehand, treating the generative video model merely as a renderer that translates pre-computed trajectories into pixels. Furthermore, they largely overlook the fundamental *kinematic* failures inherent in the models' underlying reference frames. In contrast, our work shifts the focus to kinematics, endowing VDMs with internalized physical reasoning to resolve motion entanglement. We train

models to imagine and synthesize motion directly from semantic intent, without relying on external physics engines or pre-simulated trajectories.

3 Method

3.1 Preliminaries

Text-to-video (T2V) generation aims to synthesize high-quality videos that are semantically aligned with user prompts. Most state-of-the-art approaches build on diffusion models, which generate videos by iteratively denoising a latent representation conditioned on a text prompt p. During training, noise ϵ is added to the real video x_0 sampled from the training dataset to produce a noisy version x_t. For instance, in standard discrete diffusion models (e.g., DDPMs [20,48,49]), this process is formulated as:

$$x_t = \sqrt{\alpha_t}\, x_0 + \sqrt{1-\alpha_t}\, \epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \tag{1}$$

where α_t is cumulative noise schedule. Alternatively, Flow Matching-based models [13, 32, 34] instead define a linear interpolation between the data and noise:

$$x_t = (1-t)\, x_0 + t\, \epsilon, \quad t \in [0, 1]. \tag{2}$$

The diffusion model f_θ denoises x_t by predicting ϵ, x_0, or a velocity field v, but in practice these variants typically share a pixel-level MSE training objective:

$$\mathcal{L}_{\text{MSE}} = \mathbb{E}_{x_0, \epsilon, t}\Big[\|f_\theta(x_t, t, p) - y_t\|^2\Big], \tag{3}$$

where y_t denotes the corresponding target (noise, video, vector field). As we will demonstrate, this objective encourages accurate low-level signal reconstruction, but it does not directly enforce reasoning about the underlying motion structure.

3.2 Motivation

Recent studies [5, 22, 39, 57, 61] have demonstrated that noisy inputs in diffusion models retain substantial residual information from the original data. This residual information acts as a "reconstruction shortcut" during training, allowing the model to recover the clean video by replicating preexisting motion rather than by reasoning from the conditioning signals (e.g., textual prompts). This effect is particularly pronounced in pre-trained diffusion models, which already possess strong low-level reconstruction abilities.

To verify this hypothesis, we conduct a simple diagnostic experiment inspired by Videojam [6]. Specifically, we evaluate a pretrained Wan2.1-T2V-1.3B [54] model under four distinct input settings to assess how sensitive its reconstruction behavior is to different conditioning factors:

- (i) *Normal*: the original video with its correct prompt;
- (ii) *Prompt Mismatched*: the original video with an unrelated prompt;
- (iii) *Frame Shuffled*: a temporally shuffled video with the correct prompt;

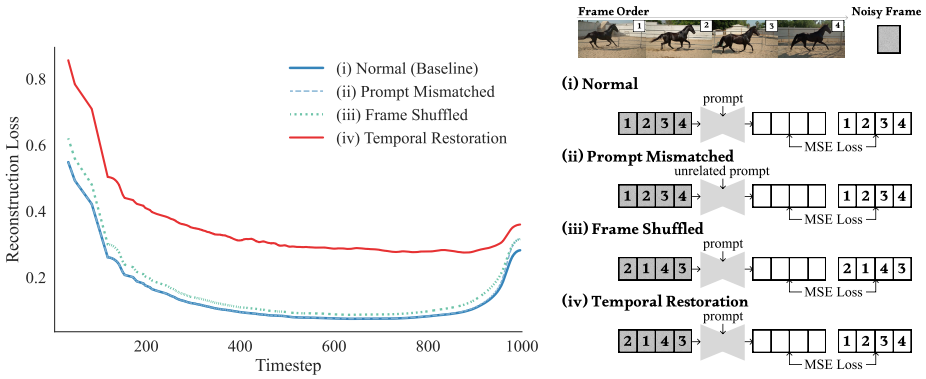

Fig. 2: MSE loss curves of four input settings. This experiment analyzes pretrained diffusion model behavior in the training setting (one-step prediction). Because the loss is dominated by shortcut cues in the noisy input, it is largely insensitive to prompt and frame-order perturbations. This explains why standard SFT is ineffective for learning physically plausible motion generation and motivating our Self-Imagination paradigm.

– (iv) *Temporal Restoration*: a shuffled video with the correct prompt, with reconstruction loss computed against the original temporally coherent video.

In each case, we add random noise levels $t \in [0, 1]$ (corresponding to 0-1000 diffusion timesteps) to 1,000 video-text pairs and compute the reconstruction loss (velocity prediction for Wan) at each timestep. As shown in Fig. 2, the first three settings yield almost identical loss curves, indicating that the model relies predominantly on the residual information in the noisy input rather than the text prompt or meaningful temporal order. In contrast, setting (iv) yields significantly higher losses, suggesting that the model has difficulty addressing the temporal dynamics when the prompt strongly contradicts the residual motion cues inherent in the noisy input.

These results reveal that the model behaves largely as a pixel-level refiner rather than a physically grounded generator. Such reliance on reconstruction shortcuts fundamentally limits the effectiveness of conventional supervised fine-tuning (SFT), which may require massive data yet still fails to improve motion understanding. This motivates us to break the reconstruction shortcut and force the model to learn motion modeling from prompts, which we achieve through our proposed self-imagination fine-tuning.

3.3 Self-Imagination Fine-tuning

This section presents an overview of our method. The pipeline is shown in Fig. 3, and pseudocode is provided in Algorithm 1. For clarity, the VAE encoder/decoder and classifier-free guidance (CFG) are omitted.

Compared to standard reconstruction-based training of diffusion models, we discard real video inputs and replace them with pure Gaussian noise. Therefore,

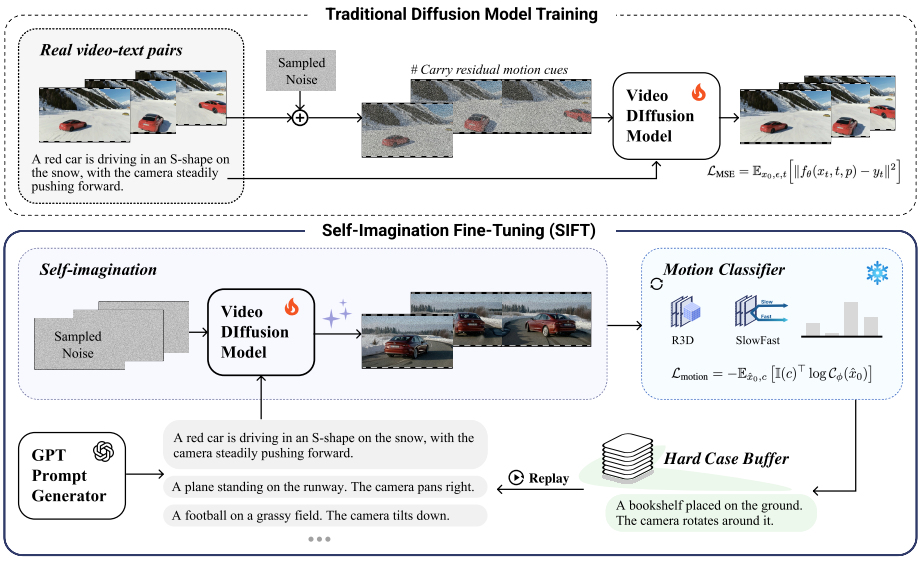

Fig. 3: Pipeline comparison between traditional diffusion model training (top) and our proposed self-imagination fine-tuning (bottom).

the model needs to imagine the video and construct motion solely from textual prompts that specify scene content and motion relations. These text prompts, which can be freely produced by large language models (e.g., GPT [23]), provide an unlimited source of training scenarios spanning diverse and even rare combinations of relative motions that are difficult to obtain from real videos. To encourage the imagined motions to be physically plausible, we replace the primary reconstruction objective in the self-imagination branch with motion-aware discriminative supervision, while retaining a lightweight MSE term on real video-text pairs to preserve visual quality. Through this process, the model learns to infer and synthesize coherent, physically plausible motion dynamics directly from semantic intent.

In practice, since diffusion models establish global motion structure primarily during early denoising stages [31, 60], we restrict training to high-noise regimes and perform only a few denoising steps (e.g., 3 steps) from pure noise. This allows us to efficiently probe the model's motion imagination where it matters most, significantly reducing computational cost compared to full-sequence generation.

3.4 Decoupled Motion Feedback

To provide motion-specific supervision rather than low-level appearance guidance, we train motion classifiers $\mathcal{C}_\phi$ to categorize each video into four motion types: *camera-only motion*, *object-only motion*, *both in motion*, and *both static*. Since the motion type for each self-generated video is known from its original

Algorithm 1 Self-Imagination Fine-Tuning

Require: Pretrained VDM f_θ, motion classifiers $\mathcal{C}_\phi$ (R3D & SlowFast), prompt generator, real batch $\mathcal{D}_{\text{real}}$, Hard Case Buffer $\mathcal{B}$

Initialize $\mathcal{B} \leftarrow \emptyset$
for each training iteration s **do**
  $\mathcal{L}_{\text{motion}} \leftarrow 0$
  Generate a batch of prompts $\{(p_i, c_i)\}$
  for each (p_i, c_i) in the batch **do**
    $x_T \sim \mathcal{N}(0, I)$
    $\hat{x}_0 \leftarrow f_\theta^{\text{denoise}}(x_T, p_i)$
    $c_{\text{pred}} \leftarrow \mathcal{C}_\phi(\hat{x}_0)$
    $\mathcal{L}_{\text{motion}} \leftarrow \mathcal{L}_{\text{motion}} - \mathbb{I}(c_i)^\top \log c_{\text{pred}}$
    if $\arg\max(c_{\text{pred}}) \neq c_i$ **then**
      Add (p_i, c_i) to $\mathcal{B}$
    end if
  end for
  Sample a real batch and compute $\mathcal{L}_{\text{MSE}}$
  $\mathcal{L}_{\text{total}} \leftarrow \lambda \mathcal{L}_{\text{motion}} + \mathcal{L}_{\text{MSE}}$
  Update $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_{\text{total}}$
  Sample cases from $\mathcal{B}$ with probability $p_s = \min(1, s/S_{\text{warmup}})$ and replay them
end for

LLM-generated prompt, we obtain ground-truth labels automatically without manual annotation. To mitigate model bias and capture complementary inductive priors, we employ two heterogeneous classifiers. The first is a *3D-ResNet (R3D)* [53], which treats spatial and temporal dimensions symmetrically through unified 3D convolutions, making it adept at modeling short-term local motion coherence. The second is a *SlowFast* network [11], which explicitly factorizes temporal and spatial processing into dual pathways operating at different frame rates, one focusing on detailed spatial semantics and the other on high-frequency motion dynamics. This architectural dichotomy ensures our motion feedback encompasses both specialized and unified temporal understandings.

During training the classifiers, we explicitly simulate the noisy inference of SIFT: adding heavy noise to clean videos ($t \in [900, 1000)$) and running the same one-step denoising to obtain predicted $\hat{x}_0$. Classifiers are trained on these noisy, regenerated $\hat{x}_0$, which match the distribution they see when supervising SIFT and ensure robustness to noise and artifacts. On a held-out validation set constructed under the same noisy supervision distribution, R3D and SlowFast achieve 78.4% and 82.8% accuracy, respectively.

To further prevent overfitting to a single classifier, we employ an *alternating supervision strategy*. Rather than averaging their outputs, we alternate between the two classifiers across different training batches. This ensures the model benefits from both spatial and temporal inductive biases. The motion loss is then defined using cross-entropy between the predicted and true motion classes:

$$\ell(\hat{x}_i, c_i) = -\,\mathbb{I}(c_i)^\top \log \mathcal{C}_\phi(\hat{x}_i), \tag{4}$$

where $\mathbb{I}(c)$ is the one-hot vector of the ground-truth motion type.

3.5 Progressive Hard Case Replay

To further improve stability and maximize learning efficiency, we introduce a *progressive hard case replay strategy*. The key idea is to gradually expose the model to increasingly difficult motion cases, rather than overwhelming it with hard samples from the beginning. After each imagination step, videos that fail to produce correct motion categories are stored in a hard case buffer $\mathcal{B}$. During the early training phase, these hard samples are temporarily excluded from gradient updates to avoid unstable optimization caused by noisy supervision. As training progresses, hard cases are incorporated into the loss with a gradually increasing participation p_s:

$$p_s = \min\left(1, \frac{s}{S_{\text{warmup}}}\right), \tag{5}$$

where s is the current training iteration step and S_{warmup} is a predefined warm-up period. Formally, let b_s denote the current batch and $H \subseteq \mathcal{B}$ be the set of hard samples identified by motion classifiers. For each sample $i \in b_s$, we draw $u_i \sim \text{Uniform}[0, 1]$, and include it in the loss only if $u_i \leq p_s$. The batch-wise loss is defined as:

$$\mathcal{L}_{\text{motion}} = \sum_{i \in b_s} \left[\mathbf{1}_{\{i \notin H\}} + \mathbf{1}_{\{i \in H\}} \mathbf{1}_{\{u_i \leq p_s\}}\right] \ell(\hat{x}_i, c_i), \tag{6}$$

where $\ell(\hat{x}_i, c_i)$ is the motion classification loss for the generated video $\hat{x}_i$ with ground-truth motion label c_i.

Meanwhile, we periodically replay samples from the buffer to reinforce challenging motion patterns. This progressive replay strategy allows the model to first stabilize its understanding on borderline cases before tackling the complex motion dynamics, leading to more robust and stable convergence.

4 Experiment

4.1 Experimental Setup

Dataset. To train motion classifiers, we first construct a dataset comprising 4,000 videos, evenly distributed across four motion categories: camera-only motion, object-only motion, both in motion, and both static. The videos are manually filtered and collected from existing large-scale video datasets [12,14,30,41,64] and open-source websites [44]. Each video is annotated with two types of captions: (1) general content descriptions generated using Qwen-VL-2.5 [1], and (2) detailed camera movement descriptions generated by a specialized model [30] fine-tuned on Qwen-VL-2.5 with camera-specific data.

Benchmark. For testing, we focus on the two critical categories that require disentanglement: camera-only motion and object-only motion. Our full test set

Table 1: Quantitative comparison of our method with baselines across VLM and human evaluations. Metrics include Semantic Adherence (SA) and Physical Commonsense (PC). Our method achieves the highest scores in all categories.

Model Setting	VLM Score (↑)				Human Score (↑)			
	Camera-only		Object-only		Camera-only		Object-only	
	SA	PC	SA	PC	SA	PC	SA	PC
Wan [54]	3.58	4.06	3.80	3.60	3.00	2.79	2.91	2.75
+ SFT	3.94	4.73	4.32	4.30	3.26	3.53	2.91	2.43
+ SIFT (Ours)	**4.80**	**4.93**	**4.75**	**4.72**	**3.89**	**4.24**	**3.98**	**3.84**
CogVideoX [66]	3.95	3.10	3.75	2.81	3.51	3.20	2.73	2.39
+ SFT	4.68	4.32	4.25	3.69	3.50	3.10	3.18	2.86
+ VideoREPA [70]	4.53	4.05	4.00	3.50	3.34	4.05	2.20	2.60
+ SIFT (Ours)	**4.89**	**4.84**	**4.38**	**4.25**	**3.88**	**4.25**	**3.75**	**3.68**

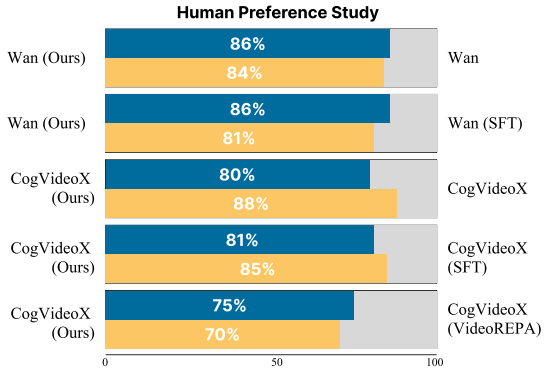

Fig. 4: Human preference study results, showing the winning rate of our method vs. baselines. Blue bars indicate Semantic Adherence (SA) and yellow bars indicate Physical Commonsense (PC).

comprises 100 test prompts, each following a precise two-clause format: {content prompt} {camera prompt}. The camera clause always begins with "The camera..." and encompasses a diverse set of 12 motion templates (e.g., pan, tilt, truck, orbit, etc.) and 5 static templates. For moving object scenarios, content prompts are sampled from the motion binding subset of T2V-CompBench [50]. For static object scenarios, content prompts are generated using Gemini-Pro-Vision-1.5 [52], covering a comprehensive range of entities including vehicles, common objects, and living beings.

Evaluation Metrics. As the first work to investigate motion entanglement in videos, we design evaluation metrics based on VideoPhy [3] and PhysCtrl [55] with necessary adaptations. We employ both a vision-language model (VLM) and human evaluations to assess each video on 1-5 scale across two dimensions: (1)

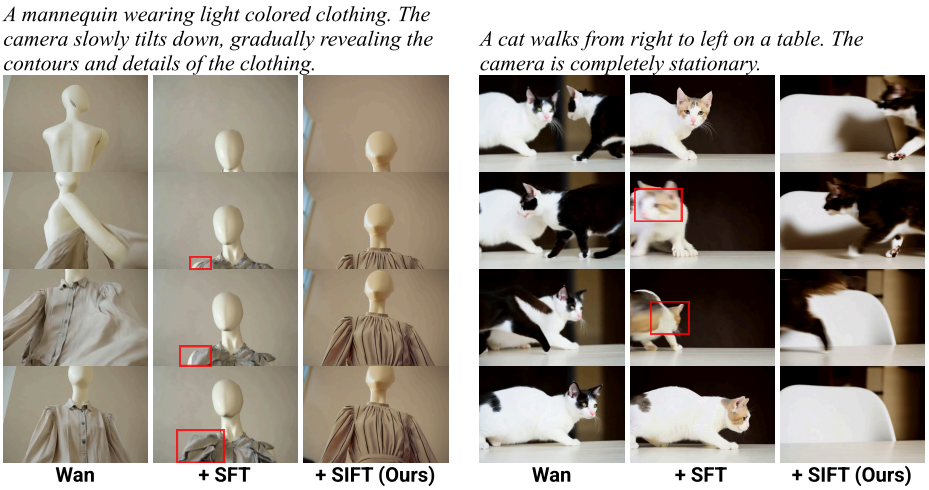

Fig. 5: Qualitative comparison on the Wan backbone. SIFT better preserves prompt-specified relative motion and temporal consistency than other baselines.

Semantic Adherence (SA): measures how well the video content aligns with the text prompt, particularly whether the generated motions match the prompt; (2) *Physical Commonsense (PC)*: measures whether the observed motion follows the physics laws in the real world. Additionally, we run a pairwise *human preference study*: for each prompt, we present two videos (e.g., our method versus a baseline) and ask evaluators which video better satisfies SA / PC.

Implementation Details. All experiments are conducted on 8 NVIDIA H100 GPUs. We adopt two pre-trained video diffusion models as backbones: Wan2.1-T2V-1.3B [54] and CogVideoX [66]. Both models use an identical learning rate of $5e^{-6}$ and are optimized for 1,000 steps with a batch size of 1. For evaluation, we use InternVideo2.5 [59] as the VLM evaluator. During the self-imagination phase, video generation proceeds through timesteps $t = 1000, 980, 960$. The Progressive Hard Case Replay warm-up period T_{warmup} is set to 500 iterations, with hard case replay performed every 10 batches. The motion-aware discriminative loss weight is set to 0.01. Alternating supervision between motion classifiers is applied every 25 batches.

4.2 Comparison with Existing Baselines

Baselines. We conduct extensive comparative experiments on two open source video diffusion models: Wan2.1 [54] and CogVideoX [66]. The compared methods include: (1) *Original*: the base pre-trained models without any modification; (2) *SFT*: supervised fine-tuning on our human-collected dataset of 4,000 motion-decoupled videos; (3) *VideoREPA [70]*: a method that applies REPA [67] to video diffusion model, distilling general physical understanding from video foundation models by aligning token-level relations; and (4) our proposed *SIFT*.

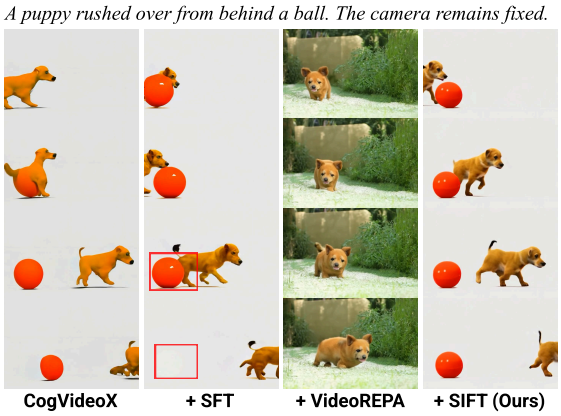

Fig. 6: Qualitative comparison on the CogVideoX backbone. SIFT produces more coherent object interactions and more physically plausible relative motion than others.

Quantitative Evaluation. As summarized in Tab. 1, both VLM-based and human evaluations across two backbone models demonstrate that our method achieves the best performance in both Semantic Adherence (SA) and Physical Commonsense (PC). The SFT baseline brings only limited gains and even underperforms the original model in some human evaluations (e.g., Wan + SFT obtains a PC score of 2.43, versus 2.75 for vanilla Wan). This supports our claim that conventional SFT helps the model imitate motion distributions present in the data but fails to cultivate genuine physical reasoning. The model tends to memorize statistical correlations of training data but cannot generalize to novel scenarios, often generating kinematically inconsistent motions that human raters can easily spot. VideoREPA [70] enhances temporal stability by distilling general physical knowledge via feature alignment but lacks explicit supervision for motion disentanglement. As a result, improvements in PC metrics are modest, with smoother motion visuals yet frequent errors in relative dynamics (e.g., camera versus object motion). In contrast, our method demonstrates significant improvement across all metrics. By breaking the reconstruction shortcut and learning from self-imagined videos, SIFT compels the model to infer motion directly from semantic prompts, rather than replicating residual motion cues. The motion-aware discriminative supervision and progressive hard-case replay further guide the model to correct mistakes and stabilize learning, leading to robust disentanglement of independent motion sources. Notably, human preference studies in Fig. 4 confirm that SIFT is strongly favored over all baselines, showing that our method enhances both the realism and controllability of generated motion without sacrificing semantic fidelity.

Qualitative Evaluation. The qualitative results between our method and baselines can be found in Figs. 5 and 6, and they further illustrate the advantages of SIFT. For instance, in the puppy-and-ball scene, SIFT generates natural relative motion and consistent object presence across frames. However, the original

CogVideoX exhibits unrealistic interactions where the puppy partially merges with the ball; CogVideoX + SFT produces temporal discontinuities with the ball abruptly disappearing; and CogVideoX + VideoREPA fails to generate the ball altogether. These artifacts demonstrate the limitations of existing approaches in modeling kinematically correct and coherent motion dynamics.

Camera-controlled Methods. Camera-controlled methods [10, 18] primarily target precise controllability through dense external conditions, such as manually designed trajectories and reference first frames, whereas SIFT aims to improve the backbone model's intrinsic motion prior under the standard text-to-video setting. For completeness, we report a separate comparison on a camera-only subset in Tab. 2, where we align the inputs as closely as possible: camera poses are specified using a small set of fixed directional trajectories, and required reference first frames are generated by Nano Banana [15].

Table 2: Comparison with camera-controlled methods.

Method	SA ↑	PC ↑
CameraCtrl	3.07	3.03
Wan-Move	3.73	3.70
SIFT	**4.10**	**4.30**

Generalization Beyond Basic Motion Disentanglement. Beyond the camera-only and object-only settings in our main benchmark, we further evaluate SIFT on three more challenging motion scenarios: multi-object motion, articulated motion, and long-horizon motion. Each setting contains 50 prompts. For the long-horizon setting, each prompt involves at least two sequential motion phases, and the generated videos contain twice as many frames as those in the standard benchmark. As shown in Tab. 3, SIFT consistently improves both semantic adherence and physical commonsense across all three settings. These results suggest that SIFT improves the model's understanding of motion, enabling it to achieve better performance in more complex and general scenarios.

Table 3: Generalization to more complex motion settings.

Setting	Method	SA ↑	PC ↑
Multi-object	Wan	3.88	3.92
	+ SIFT	**4.50**	**4.18**
Articulated	Wan	3.76	4.04
	+ SIFT	**4.02**	**4.36**
Long horizon	Wan	3.50	3.82
	+ SIFT	**3.88**	**4.31**

4.3 Ablation Study

To better understand the contribution of each component of our SIFT, we conducted a comprehensive ablation study, as shown in Tab. 4. All ablations are evaluated on a randomly sampled set of 40 prompts from the test benchmark.

Self-Imagination Generation. We ablate the self-imagination design by initializing generation from noisy real videos instead of pure noise, while keeping all other components unchanged. As shown in Tab. 4, this variant performs worse in both semantic adherence and physical plausibility. When conditioned on noised real videos, the model exploits residual motion cues, such as coarse

Table 4: Ablation study of SIFT components on the Wan using VLM evaluation. SA: semantic adherence; PC: physical commonsense. For a fair and accurate comparison, videos generated from the same prompt by all methods are jointly provided to the VLM for scoring. **So scores are relative within each group and should not be directly comparable to those in Tab. 1.**

Model Variant	Camera-only		Object-only	
	SA(↑)	PC(↑)	SA(↑)	PC(↑)
(a) Self-Imagination Generation				
w/ Reconstruction Shortcut	3.05	3.55	3.45	3.65
(b) Alternating Supervision				
w/ Single Classifier (R3D)	3.65	3.55	3.20	3.60
w/ Single Classifier (SlowFast)	3.35	3.25	3.50	3.85
(c) Progressive Hard Case Replay				
w/o Progressive Hard Case Replay	3.00	3.45	3.35	3.70
Ours (Full SIFT)	**3.85**	**4.15**	**3.50**	**4.00**

object trajectories and temporal coherence, to guide reconstruction, bypassing the need to infer motion from text. Moreover, because real videos are inherently kinematically plausible, the model tends to copy motion patterns present in the input rather than learning to infer kinematically plausible motion dynamics. As a result, it fails to develop a robust understanding of motion semantics from language, leading to poorer generalization and reduced motion disentanglement.

Alternating Supervision. We evaluate our alternating strategy by comparing it against using only a single motion classifier supervision (R3D or SlowFast). As shown in Tab. 4, both single-classifier variants perform worse than our full approach. As illustrated in Fig. 7, training with a single classifier causes its accuracy to saturate quickly, indicating overfitting to its own inductive bias. This can mislead the generator to reproduce artifacts favored by that classifier rather than generating physically plausible motion. In contrast, alternating between heterogeneous classifiers prevents overfitting to a fixed inductive prior and encourages the generator to satisfy complementary motion criteria, resulting in more stable and kinematically coherent motion generation.

Progressive Hard Case Replay. We ablate this component by removing both the hard case buffer and the progressive scheduling, i.e., exposing the model to hard cases from the beginning of training. As shown in Tab. 4, this variant leads to clear degradation in both physical plausibility and semantic adherence. The performance drop occurs because introducing challenging cases too early injects noisy gradients when the model's parameters are not yet sufficiently prepared. This disrupts motion learning and even harms overall visual quality, as evidenced by the more significant decline in semantic adherence (SA). In contrast, our progressive strategy allows the model to first master basic motion patterns before tackling more challenging cases, leading to more stable convergence and better

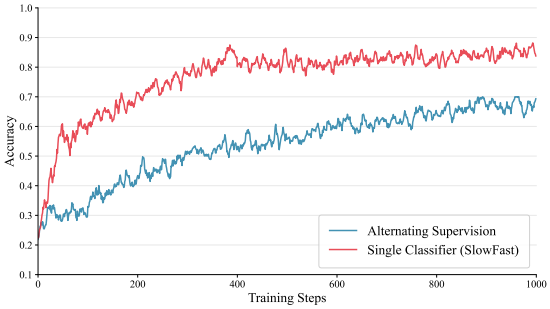

Fig. 7: Accuracy curves of motion classifiers during training of video generative model.

final performance. Moreover, the hard case replay improves training efficiency, enabling the model to achieve greater improvement within the same number of optimization steps.

5 Conclusion

In this work, we identify a key kinematic limitation in current video diffusion models, which we term "Motion Entanglement". We show that this issue arises from data-induced spurious correlations as well as the reconstruction shortcut and objective biases inherent in their training paradigm. These factors prevent models from learning physically grounded motion. To address this, we propose Self-Imagination Fine-Tuning (SIFT), a paradigm shift from reconstruction-based training to imagination-driven learning that initializes generation from random noise. Combined with motion-aware discriminative supervision and a progressive hard-case replay strategy, SIFT enables the model to infer, evaluate, and correct motion dynamics without relying on scarce motion-decoupled data. Extensive experiments on both Wan and CogVideoX demonstrate that SIFT significantly improves physical plausibility and motion disentanglement while preserving semantic fidelity, paving the way toward more realistic and controllable text-to-video generation.

Limitations and Future Work. Although the four-way motion taxonomy provides effective supervision for disentangling camera and object motion, it remains coarse. In particular, it does not explicitly model motion direction, magnitude, individual trajectories, or temporal transitions between multiple motion states. This limitation becomes more evident in involving multiple moving entities and complex motion composition scenarios, where a single categorical label may not fully describe the underlying motion structure. In addition, the quality of the feedback depends on the robustness of the motion classifiers, which can be uncertain for ambiguous or severely corrupted self-generated videos. A promising direction is to introduce richer factorized motion representations, such as object-level trajectories, optical flow, or motion fields, and to design more fine-grained feedback for complex interactions and longer temporal horizons.

References

1. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv (2025)
2. Bain, M., Nagrani, A., Varol, G., Zisserman, A.: Frozen in time: A joint video and image encoder for end-to-end retrieval. In: ICCV. pp. 1728–1738 (2021)
3. Bansal, H., Lin, Z., Xie, T., Zong, Z., Yarom, M., Bitton, Y., Jiang, C., Sun, Y., Chang, K.W., Grover, A.: Videophy: Evaluating physical commonsense for video generation. arXiv (2024)
4. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv (2023)
5. Burgert, R., Xu, Y., Xian, W., Pilarski, O., Clausen, P., He, M., Ma, L., Deng, Y., Li, L., Mousavi, M., et al.: Go-with-the-flow: Motion-controllable video diffusion models using real-time warped noise. In: CVPR. pp. 13–23 (2025)
6. Chefer, H., Singer, U., Zohar, A., Kirstain, Y., Polyak, A., Taigman, Y., Wolf, L., Sheynin, S.: Videojam: Joint appearance-motion representations for enhanced motion generation in video models. arXiv (2025)
7. Chen, H., Xia, M., He, Y., Zhang, Y., Cun, X., Yang, S., Xing, J., Liu, Y., Chen, Q., Wang, X., et al.: Videocrafter1: Open diffusion models for high-quality video generation. arXiv (2023)
8. Chen, H., Zhang, Y., Cun, X., Xia, M., Wang, X., Weng, C., Shan, Y.: Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In: CVPR. pp. 7310–7320 (2024)
9. Chen, T.S., Siarohin, A., Menapace, W., Deyneka, E., Chao, H.w., Jeon, B.E., Fang, Y., Lee, H.Y., Ren, J., Yang, M.H., et al.: Panda-70m: Captioning 70m videos with multiple cross-modality teachers
10. Chu, R., He, Y., Chen, Z., Zhang, S., Xu, X., Xia, B., Wang, D., Yi, H., Liu, X., Zhao, H., et al.: Wan-move: Motion-controllable video generation via latent trajectory guidance. arXiv preprint arXiv:2512.08765 (2025)
11. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: ICCV. pp. 6202–6211 (2019)
12. Fu, X., Liu, X., Wang, X., Peng, S., Xia, M., Shi, X., Yuan, Z., Wan, P., Zhang, D., Lin, D.: 3dtrajmaster: Mastering 3d trajectory for multi-entity motion in video generation. arXiv (2024)
13. Geng, Z., Deng, M., Bai, X., Kolter, J.Z., He, K.: Mean flows for one-step generative modeling. arXiv (2025)
14. Gillman, N., Herrmann, C., Freeman, M., Aggarwal, D., Luo, E., Sun, D., Sun, C.: Force prompting: Video generation models can learn and generalize physics-based control signals. arXiv (2025)
15. Google DeepMind: Gemini image — nano banana. `https://deepmind.google/models/gemini-image/` (2025), official model page, accessed 2026-03-03
16. Google DeepMind: Veo 3. `https://deepmind.google/models/veo/` (2025)
17. Guo, Y., Yang, C., Rao, A., Liang, Z., Wang, Y., Qiao, Y., Agrawala, M., Lin, D., Dai, B.: Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. arXiv (2023)
18. He, H., Xu, Y., Guo, Y., Wetzstein, G., Dai, B., Li, H., Yang, C.: Cameractrl: Enabling camera control for text-to-video generation. arXiv preprint arXiv:2404.02101 (2024)

19. Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D.P., Poole, B., Norouzi, M., Fleet, D.J., et al.: Imagen video: High definition video generation with diffusion models. arXiv (2022)
20. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. NeurIPS (2020)
21. Hong, W., Ding, M., Zheng, W., Liu, X., Tang, J.: Cogvideo: Large-scale pretraining for text-to-video generation via transformers. arXiv (2022)
22. Hou, C., Chen, Z.: Training-free camera control for video generation. arXiv (2024)
23. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv (2024)
24. Ju, X., Gao, Y., Zhang, Z., Yuan, Z., Wang, X., Zeng, A., Xiong, Y., Xu, Q., Shan, Y.: Miradata: A large-scale video dataset with long durations and structured captions. NeurIPS **37**, 48955–48970 (2024)
25. Kang, B., Yue, Y., Lu, R., Lin, Z., Zhao, Y., Wang, K., Huang, G., Feng, J.: How far is video generation from world model: A physical law perspective. arXiv (2024)
26. Khachatryan, L., Movsisyan, A., Tadevosyan, V., Henschel, R., Wang, Z., Navasardyan, S., Shi, H.: Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In: ICCV. pp. 15954–15964 (2023)
27. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al.: Hunyuanvideo: A systematic framework for large video generative models. arXiv (2024)
28. Kuaishou: Kling ai. `https://app.klingai.com/global/` (2025)
29. Lian, L., Shi, B., Yala, A., Darrell, T., Li, B.: Llm-grounded video diffusion models. arXiv (2023)
30. Lin, Z., Cen, S., Jiang, D., Karhade, J., Wang, H., Mitra, C., Ling, T., Huang, Y., Liu, S., Chen, M., et al.: Towards understanding camera motions in any video. arXiv (2025)
31. Ling, P., Bu, J., Zhang, P., Dong, X., Zang, Y., Wu, T., Chen, H., Wang, J., Jin, Y.: Motionclone: Training-free motion cloning for controllable video generation. arXiv (2024)
32. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv (2022)
33. Liu, S., Ren, Z., Gupta, S., Wang, S.: Physgen: Rigid-body physics-grounded image-to-video generation. In: ECCV. pp. 360–378. Springer (2024)
34. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. arXiv (2022)
35. Liu, Y., Zhang, K., Li, Y., Yan, Z., Gao, C., Chen, R., Yuan, Z., Huang, Y., Sun, H., Gao, J., et al.: Sora: A review on background, technology, limitations, and opportunities of large vision models. arXiv (2024)
36. Lv, J., Huang, Y., Yan, M., Huang, J., Liu, J., Liu, Y., Wen, Y., Chen, X., Chen, S.: Gpt4motion: Scripting physical motions in text-to-video generation via blender-oriented gpt planning. In: CVPR. pp. 1430–1440 (2024)
37. Ma, G., Huang, H., Yan, K., Chen, L., Duan, N., Yin, S., Wan, C., Ming, R., Song, X., Chen, X., et al.: Step-video-t2v technical report: The practice, challenges, and future of video foundation model. arXiv (2025)
38. Ma, X., Wang, Y., Jia, G., Chen, X., Liu, Z., Li, Y.F., Chen, C., Qiao, Y.: Latte: Latent diffusion transformer for video generation. arXiv (2024)
39. Mao, J., Wang, X., Aizawa, K.: Guided image synthesis via initial image editing in diffusion model. In: ACM Multimedia. pp. 5321–5329 (2023)

40. Meng, F., Liao, J., Tan, X., Shao, W., Lu, Q., Zhang, K., Cheng, Y., Li, D., Qiao, Y., Luo, P.: Towards world simulator: Crafting physical commonsense-based benchmark for video generation. arXiv (2024)
41. Nan, K., Xie, R., Zhou, P., Fan, T., Yang, Z., Chen, Z., Li, X., Yang, J., Tai, Y.: Openvid-1m: A large-scale high-quality dataset for text-to-video generation. arXiv (2024)
42. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv (2021)
43. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: ICCV. pp. 4195–4205 (2023)
44. Pexels: Pexels – free stock photos, royalty free images & videos. `https://www.pexels.com/` (2025), accessed: 2025-11-13
45. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv (2022)
46. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
47. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding (2022)
48. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: ICLR (2021)
49. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv (2020)
50. Sun, K., Huang, K., Liu, X., Wu, Y., Xu, Z., Li, Z., Liu, X.: T2v-compbench: A comprehensive benchmark for compositional text-to-video generation. In: CVPR. pp. 8406–8416 (2025)
51. Tan, X., Jiang, Y., Li, X., Zong, Z., Xie, T., Yang, Y., Jiang, C.: Physmotion: Physics-grounded dynamics from a single image. arXiv (2024)
52. Team, G., Georgiev, P., Lei, V.I., Burnell, R., Bai, L., Gulati, A., Tanzer, G., Vincent, D., Pan, Z., Wang, S., et al.: Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv (2024)
53. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: CVPR. pp. 6450–6459 (2018)
54. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv (2025)
55. Wang, C., Chen, C., Huang, Y., Dou, Z., Liu, Y., Gu, J., Liu, L.: Physctrl: Generative physics for controllable and physics-grounded video generation. arXiv (2025)
56. Wang, Q., Shi, Y., Ou, J., Chen, R., Lin, K., Wang, J., Jiang, B., Yang, H., Zheng, M., Tao, X., et al.: Koala-36m: A large-scale video dataset improving consistency between fine-grained conditions and video content. In: CVPR. pp. 8428–8437 (2025)
57. Wang, R., Huang, H., Zhu, Y., Russakovsky, O., Wu, Y.: The silent assistant: Noisequery as implicit guidance for goal-driven image generation. In: ICCV. pp. 17618–17628 (2025)
58. Wang, Y., He, Y., Li, Y., Li, K., Yu, J., Ma, X., Li, X., Chen, G., Chen, X., Wang, Y., et al.: Internvid: A large-scale video-text dataset for multimodal understanding and generation. arXiv (2023)

59. Wang, Y., Li, X., Yan, Z., He, Y., Yu, J., Zeng, X., Wang, C., Ma, C., Huang, H., Gao, J., et al.: Internvideo2. 5: Empowering video mllms with long and rich context modeling. arXiv (2025)
60. Wu, T., Zhang, Y., Wang, X., Zhou, X., Zheng, G., Qi, Z., Shan, Y., Li, X.: Customcrafter: Customized video generation with preserving motion and concept composition abilities. In: AAAI. vol. 39, pp. 8469–8477 (2025)
61. Wu, T., Si, C., Jiang, Y., Huang, Z., Liu, Z.: Freeinit: Bridging initialization gap in video diffusion models. In: ECCV. pp. 378–394. Springer (2024)
62. Xie, T., Zhao, Y., Jiang, Y., Jiang, C.: Physanimator: Physics-guided generative cartoon animation. In: CVPR. pp. 10793–10804 (2025)
63. Xue, Q., Yin, X., Yang, B., Gao, W.: Phyt2v: Llm-guided iterative self-refinement for physics-grounded text-to-video generation. In: CVPR. pp. 18826–18836 (2025)
64. Xue, Z., Zhang, J., Hu, T., He, H., Chen, Y., Cai, Y., Wang, Y., Wang, C., Liu, Y., Li, X., et al.: Ultravideo: High-quality uhd video dataset with comprehensive captions. arXiv (2025)
65. Yang, X., Li, B., Zhang, Y., Yin, Z., Bai, L., Ma, L., Wang, Z., Cai, J., Wong, T.T., Lu, H., et al.: Towards physically plausible video generation via vlm planning. arXiv **2** (2025)
66. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv (2024)
67. Yu, S., Kwak, S., Jang, H., Jeong, J., Huang, J., Shin, J., Xie, S.: Representation alignment for generation: Training diffusion transformers is easier than you think. arXiv (2024)
68. Zhang, K., Xiao, C., Mei, Y., Xu, J., Patel, V.M.: Think before you diffuse: Llms-guided physics-aware video generation. arXiv (2025)
69. Zhang, T., Yu, H.X., Wu, R., Feng, B.Y., Zheng, C., Snavely, N., Wu, J., Freeman, W.T.: Physdreamer: Physics-based interaction with 3d objects via video generation. In: ECCV. pp. 388–406. Springer (2024)
70. Zhang, X., Liao, J., Zhang, S., Meng, F., Wan, X., Yan, J., Cheng, Y.: Videorepa: Learning physics for video generation through relational alignment with foundation models. arXiv (2025)
71. Zheng, Z., Peng, X., Yang, T., Shen, C., Li, S., Liu, H., Zhou, Y., Li, T., You, Y.: Open-sora: Democratizing efficient video production for all. arXiv (2024)