CVPR #10

Tempered Self-Similarity Alignment for Physically Plausible Video Generation

Anonymous CVPR submission

Paper ID 10

Abstract

Despite remarkable advances in video generative models, they still struggle to generate physically realistic videos, frequently exhibiting appearance drift, implausible motion, and temporal inconsistencies. In this work, we address this limitation by transferring relational knowledge encoded in spatio-temporal self-similarity (STSS) from visual foundation models into video generative models. STSS represents pairwise similarities among features across space and time, revealing the relational structure of how objects interact with other entities throughout a video, effectively capturing real-world dynamics, including object motion and semantic transformations. To transfer this relational knowledge, we propose Tempered Self-similarity Alignment (TSA) loss, which transforms STSS into probabilistic correspondence distributions and trains the video generative model to align its correspondence distributions with those of the visual foundation model on dynamically changing regions. Evaluated on VideoPhy and VideoPhy2 benchmarks, our method demonstrates substantial improvements in physical plausibility across diverse interaction scenarios, validating the effectiveness of transferring relational knowledge for physically realistic video generation.

1. Introduction

Recent advances in video diffusion models [5, 6, 15, 22, 43, 52, 53] have enabled the generation of high-fidelity and high-resolution videos from text prompts. However, despite these remarkable achievements, current video generation models still struggle to produce physically plausible videos, frequently exhibiting appearance drift, unrealistic deformations, and inconsistent motion dynamics [2, 3, 32, 34]. These failures in physical realism limit the applicability of video generation models in domains requiring accurate motion simulation, such as robotics training, virtual environment construction, and video prediction.

Several approaches have been proposed to improve physical realism in video generation. One line of work leverages physics simulators to synthesize videos with physically accurate dynamics [26–29, 33, 49, 50, 55]. While effective, such simulation-based methods rely on heavy simulators and are typically restricted to specific physical domains, making them difficult to scale to diverse real-world scenarios. Another direction utilizes multimodal large language models (MLLMs) to reason about physical properties and guide the generation process toward improved physical consistency [27, 29, 45, 51]. However, these approaches often depend on iterative refinement processes that significantly increase generation latency and their effectiveness is inherently bounded by the physical reasoning capability of the MLLMs. A third line of work exploits auxiliary motion signals such as optical flow [4, 7, 14, 33]. Although optical flow provides useful motion cues, it primarily captures short-term displacements between adjacent frames, making it difficult to provide supervision for long-range motion dynamics or structural changes over time.

In this work, we propose to leverage spatio-temporal

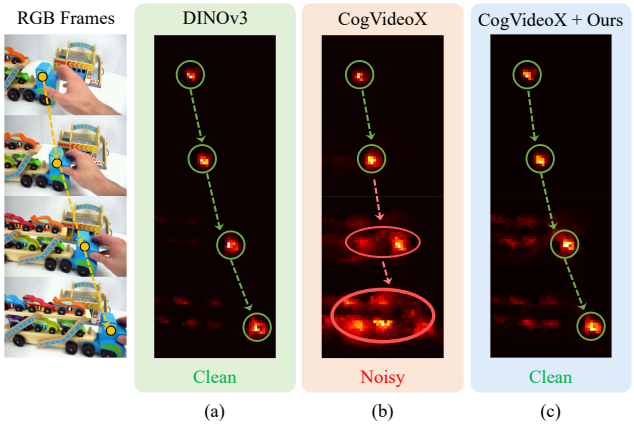


Figure 1. **Aligning spatio-temporal correspondence improves realistic dynamics in video generation.** Given the same video input, we compare spatio-temporal correspondence probability maps obtained from (a) a visual foundation model and (b) a video generative model. The foundation model captures clear correspondences, whereas the generative model produces noisy and inaccurate ones. (c) We here transfer the accurate spatio-temporal correspondences from the foundation model to the generative model, guiding the model toward more realistic motion dynamics.

self-similarity (STSS) [24, 39, 40] as a richer supervisory signal for physically plausible video generation. STSS, *i.e.*, pairwise similarities among visual features across space and time, suppresses photometric variations such as color and texture while emphasizing relational structures among visual entities in videos. Such relational patterns naturally encode how objects move, deform, and interact with surrounding entities across the entire video, capturing physically meaningful dynamics.

To effectively inject the relational knowledge encoded in STSS into video generation models, we introduce *Tempered Self-Similarity Alignment (TSA)* loss, which aligns STSS captured by pre-trained visual foundation models with that of video generative models. However, directly aligning raw STSS tensors often yields overly smooth correspondence distributions, providing ambiguous supervision for precise motion learning. Instead, we temper the STSS by applying temperature-scaled normalization to obtain sharpened correspondence probability distributions, which we then align across the two models. These distributions represent object motion as *a soft trajectory across the entire video*. Unlike optical flow, which represents displacement as a one-hot correspondence between adjacent frames only , this formulation captures long-range dynamics and inherent uncertainties in structural transformations. Furthermore, we introduce background masking that restricts this alignment exclusively to dynamic regions, enabling the model to focus on physically grounded motion dynamics while avoiding unnecessary overhead on static regions. We evaluate our method on VideoPhy [2] and VideoPhy2 [3] and demonstrate that TSA loss effectively improves the physical realism of generated videos.

Our main contributions are summarized as follows:

- We propose Tempered Self-Similarity Alignment (TSA) loss, which effectively transfers relational knowledge in STSS from visual foundation models into video diffusion models for enhancing realistic dynamics.
- We introduce a tempering mechanism that transforms STSS into sharpened correspondence probability distributions, enabling more precise motion supervision than direct STSS alignment.
- We demonstrate that our method significantly improves physical plausibility on VideoPhy and VideoPhy2 benchmarks, achieving state-of-the-art performance in generating physically realistic videos.

2. Related Work

Physically Plausible Video Generation. Despite recent advances in video diffusion models [5, 6, 15, 22, 43, 52, 53], they still struggle to produce physically plausible videos, often exhibiting identity drift, unrealistic deformations, or inconsistent dynamics [2, 3, 32, 34]. To address these limitations, recent studies have explored various strategies to enhance physical realism in video generation. One line of research integrates external physics simulators, such as MPM [9, 37, 41], with generative models to synthesize physically grounded videos [26–29, 33, 49, 50, 55]. Another direction leverages LLMs for physical property analysis and reasoning about physical phenomena, guiding video generation toward improved physical consistency [27, 29, 45, 51]. Additionally, several methods utilize optical flow as conditioning or guidance signals to generate more realistic motion [7, 14, 33]. Recently, several methods leverage representation alignment [54] to transfer physical knowledge from visual foundation models into video diffusion models [13, 47, 56]. Building on this direction, we propose an alignment loss that transfers spatio-temporal correspondence probabilities from visual foundation models into video generative models.

Representation Alignment. Representation alignment (REPA) [54] is a regularization method that aligns the representation of the generative models with that of visual foundation models. This approach is initially proposed in image generation [25, 54], effectively stabilizing and accelerating diffusion model training. Several methods [13, 47, 56] extend this to video generation. Among them, VideoREPA [56] proposes to distill spatio-temporal self-similarity (STSS) from video foundation models to improve physical realism in generated videos. However, it directly aligns raw STSS tensors uniformly across all spatial regions, limiting precise motion supervision. In contrast, we convert STSS into probabilistic correspondence distributions, where temperature scaling enables flexible control over motion supervision granularity, and restrict this alignment to dynamically active regions for more effective motion-focused alignment.

Spatio-Temporal Self-Similarity (STSS). Self-similarity [39, 40], defined as pairwise similarities among visual features, effectively reveals underlying relational patterns in visual data, *e.g.*, structural layouts or object motions, while suppressing appearance variations. In the image domain, self-similarity has been employed as a relational descriptor for template matching [39, 40], capturing view-invariant geometric patterns [16, 17], and establishing semantic correspondences [18, 21, 42]. In videos, spatio-temporal self-similarity (STSS) has been leveraged to capture temporal dynamics such as motion. Early works [16, 17] introduced temporal self-similarity descriptors for action recognition under viewpoint changes. More recent methods compute spatial cross-similarities between adjacent frames to extract short-term motion cues [23, 44], while others directly learn to transform STSS representations into bi-directional motion features [24, 48]. Some methods [19, 20] explicitly incorporate STSS into self-attention mechanisms for motion-centric video representation learning. While these approaches primarily

CVPR #10

utilize STSS for video understanding tasks such as action recognition, we extend its application to video generation. Specifically, we propose to align STSS from visual foundation models with video diffusion models, transferring their knowledge of realistic motion dynamics to enable more physically plausible video generation.

3. Preliminaries

3.1. Video Diffusion Models

Recent text-to-video diffusion models [5, 15, 22, 43, 52] follow the Latent Diffusion Model (LDM) framework [38], where the generation process is performed in the latent space of a pretrained Variational Autoencoder (VAE). Given a video $\mathbf{V} \in \mathbb{R}^{F' \times H' \times W' \times 3}$, the VAE encoder compresses it into a latent representation $\mathbf{Z}^{(0)} \in \mathbb{R}^{F \times H \times W \times C}$ with reduced spatial and temporal resolutions. A forward diffusion process corrupts $\mathbf{Z}^{(0)}$ by gradually injecting Gaussian noise over timesteps $t \in \{1, \dots, T\}$, producing noisy latents $\mathbf{Z}^{(t)} = \alpha_t \mathbf{Z}^{(0)} + \sigma_t \epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$ and α_t, σ_t are schedule-dependent coefficients. A denoising network, typically a spatio-temporal transformer such as MM-DiT [11] in CogVideoX [52], is trained to reverse this process by estimating the noise added at each timestep. The training objective minimizes the mean squared error between the true and predicted noise as,

$$\mathcal{L}_{\text{diffusion}} = \mathbb{E}_{t, \mathbf{Z}^{(0)}, \epsilon} \left[\left\| \epsilon - \epsilon_\theta(\mathbf{Z}^{(t)}, t, c) \right\|_2^2 \right], \quad (1)$$

where c denotes text prompt. At inference, the model starts from sampling pure noise $\mathbf{Z}^{(T)} \sim \mathcal{N}(0, I)$ and iteratively refines it through the learned reverse process to obtain $\hat{\mathbf{Z}}^{(0)}$. The VAE decoder then transforms $\hat{\mathbf{Z}}^{(0)}$ into the final video. Applied to this LDM-based video diffusion model, we aim to enhance the physical plausibility of generated videos.

3.2. Representation Alignment

While diffusion-based generative models produce high-quality samples, their internal representations tend to be less expressive and discriminative compared to those of large-scale visual foundation models trained (*e.g.*, DINOv2 [36]). To bridge this representational gap, representation alignment (REPA) [54] has been proposed as a regularization method that transfers rich semantic knowledge from the pretrained foundation models to the diffusion models.

REPA is initially introduced for image diffusion models [25, 54]. Let the denoising network f_θ take a noisy image $\mathbf{Z}^{(t)}$ as input, conditioned on context $\mathbf{c}$ and timestep t. The network produces intermediate activations at its l-th layer $\mathbf{H} = f_\theta(\mathbf{Z}^{(t)}, t, \mathbf{c}, l) \in \mathbb{R}^{H \times W \times C_\text{d}}$, where C_d is the channel dimension of the denosing network. A pretrained visual encoder g_ϕ that maps a clean image $\mathbf{X}$ to visual features $\mathbf{E} = g_\phi(\mathbf{X}) \in \mathbb{R}^{H \times W \times C_\text{e}}$, with feature dimension C_e. To align the representations of the diffusion model with those of the visual foundation model, REPA introduces a lightweight projector h_ψ that maps $\mathbf{H}$ to the encoder's feature space. The alignment objective encourages feature-wise consistency as:

$$\mathcal{L}_{\text{REPA}} = -\mathbb{E}_{\mathbf{X}, \mathbf{Z}^{(t)}, t} \left[\frac{1}{HW} \sum_{i=1, j=1}^{H, W} \text{sim}\left(\mathbf{E}_{i,j}, h_\psi(\mathbf{H}_{i,j})\right) \right], \quad (2)$$

where $\text{sim}(\cdot, \cdot)$ computes cosine similarity between corresponding patches. This regularizer is combined with the standard diffusion loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{diffusion}} + \lambda \mathcal{L}_{\text{REPA}}, \quad (3)$$

where λ balances generation fidelity and feature alignment. Empirically, this approach accelerates training convergence and enhances perceptual quality by leveraging semantic priors from the pretrained encoder. While REPA aligns feature representations directly, our method aligns probabilistic distributions of spatio-temporal correspondences derived from STSS, enabling more effective transfer of physically grounded motion dynamics.

4. Methods

Spatio-temporal self-similarity (STSS) effectively captures temporal dynamics such as motion by describing the relational structure across space and time in videos. By Transferring such STSS structures from visual foundation models into video diffusion models, we can guide the generative process toward producing physically realistic dynamics. To this end, we propose *Tempered Self-similarity Alignment (TSA)*, which converts the STSS tensor of each query token into probabilistic distributions of spatio-temporal correspondences of the query across frames, and aligns these correspondence distributions from visual foundation models with those from the video diffusion model. We further restrict this alignment to dynamic regions, focusing the supervision on motion-salient areas where physically meaningful dynamics occur.

4.1. Tempered Self-Similarity Alignment

Given an input video $\mathbf{V} \in \mathbb{R}^{F' \times H' \times W' \times 3}$, we first extract spatio-temporal features $\mathbf{E}^{\text{VFM}}$ using a pretrained visual foundation model g_ϕ. We then obtain intermediate features from the l-th layer of the denoising network f_θ and pass them through a lightweight projector h_ψ that maps the diffusion features into the pretrained encoder's feature space with matched spatio-temporal resolution, yielding $\mathbf{E}^{\text{VDM}}$:

$$\mathbf{E}^{\text{VFM}} = g_\phi(\mathbf{V}) \in \mathbb{R}^{F \times H \times W \times C} \quad (4)$$

$$\mathbf{E}^{\text{VDM}} = h_\psi(f_\theta(\mathbf{Z}^{(t)}, t, c, l)) \in \mathbb{R}^{F \times H \times W \times C}. \quad (5)$$

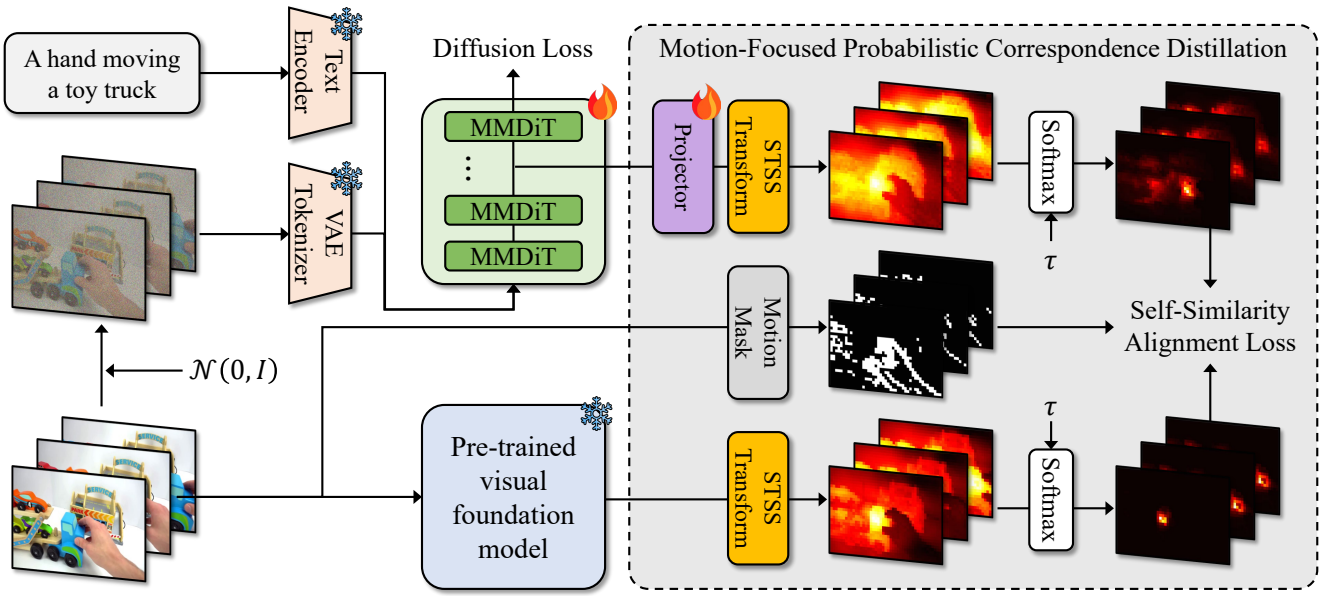


Figure 2. **Framework overview.** Our method aligns the noisy spatio-temporal correspondences of a video diffusion model with the accurate correspondences of a visual foundation model, guiding the generative model toward more realistic motion dynamics. By restricting this alignment to dynamic regions, it encourages motion-focused alignment for physically plausible video generation.

We apply L2 normalization along the channel dimension and flatten the spatio-temporal dimensions to obtain $\bar{\boldsymbol{E}}^{\text{VFM}}, \bar{\boldsymbol{E}}^{\text{VDM}} \in \mathbb{R}^{N \times C}$, where $N = FHW$. We then compute pairwise cosine similarity matrices that encode STSS relations:

$$\boldsymbol{R}^{\text{VFM}} = \bar{\boldsymbol{E}}^{\text{VFM}} (\bar{\boldsymbol{E}}^{\text{VFM}})^{\top} \in \mathbb{R}^{N \times N}, \tag{6}$$

$$\boldsymbol{R}^{\text{VDM}} = \bar{\boldsymbol{E}}^{\text{VDM}} (\bar{\boldsymbol{E}}^{\text{VDM}})^{\top} \in \mathbb{R}^{N \times N}, \tag{7}$$

where each row $\boldsymbol{R}_i$ encodes the similarities between token i and all other tokens across space and time.

To obtain frame-wise correspondences, we reshape $\boldsymbol{R} \in \mathbb{R}^{N \times N}$ to $\mathbb{R}^{N \times F \times HW}$ and apply a temperature-scaled softmax along the spatial dimension:

$$\mathbf{P}^{\text{VFM}}_{i,f} = \text{softmax}\left(\frac{\boldsymbol{R}^{\text{VFM}}_{i,f}}{\tau}\right), \mathbf{P}^{\text{VDM}}_{i,f} = \text{softmax}\left(\frac{\boldsymbol{R}^{\text{VDM}}_{i,f}}{\tau}\right), \tag{8}$$

where $\mathbf{P}_{i,f} \in \mathbb{R}^{HW}$ represents the probability distribution over spatial positions in the f-th frame that query token i corresponds to. The TSA loss aligns these correspondence distributions between the visual foundation model and the video diffusion model by minimizing the KL divergence:

$$\mathcal{L}_{\text{TSA}} = \frac{1}{NF} \sum_{i=1}^{N} \sum_{f=1}^{F} D_{\text{KL}} \left(\mathbf{P}^{\text{VFM}}_{i,f} \,\|\, \mathbf{P}^{\text{VDM}}_{i,f} \right). \tag{9}$$

TSA offers two key advantages. First, it provides stronger signals for misaligned regions while attenuating gradients for well-aligned correspondences. The gradient of the TSA loss with respect to $\boldsymbol{R}^{\text{VDM}}_{i,f}$ is computed as,

$$\frac{\partial \mathcal{L}_{\text{TSA}}}{\partial \boldsymbol{R}^{\text{VDM}}_{i,f}} = \frac{1}{\tau} \left(\mathbf{P}^{\text{VDM}}_{i,f} - \mathbf{P}^{\text{VFM}}_{i,f} \right), \tag{10}$$

which is proportional to the probability mismatch between the two distributions, unlike L1-based objectives [56] that apply uniform gradients regardless of alignment quality. Second, the temperature τ enables fine-grained motion supervision through correspondence sharpening. Without temperature scaling, directly aligning raw STSS tensors [56] yields overly smooth distributions that fail to provide precise motion supervision for each query token (Fig. 3, 2nd row). By lowering τ to sharpen the correspondence distributions, TSA provides more precise spatial supervision per query, facilitating fine-grained motion learning while preserving object part-level structural information encoded in STSS (Fig. 3, 4th row).

4.2. Motion-focused Self-Similarity Alignment

While TSA effectively aligns STSS across the entire video, most regions in videos are static and lack meaningful temporal dynamics. Aligning correspondences in such static areas introduces unnecessary computational overhead and may hinder the model from focusing on dynamic regions where physically grounded motion cues exist. To address this inefficiency and improve motion modeling, we introduce *Masked TSA (M-TSA)*, which reinforces STSS alignment on such dynamic regions.

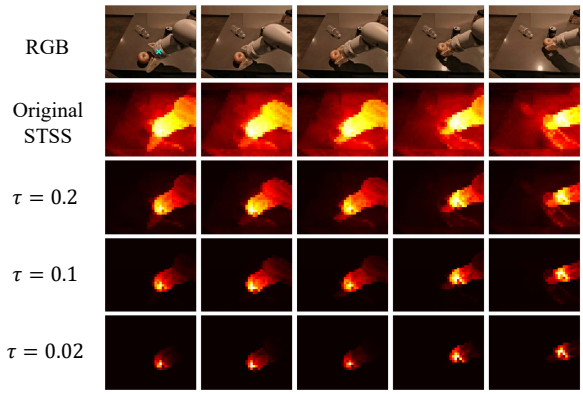


Figure 3. **Controlling correspondence granularity by temperature.** As temperature gets lower, correspondence changes from object-level, to part-level, to point-level.

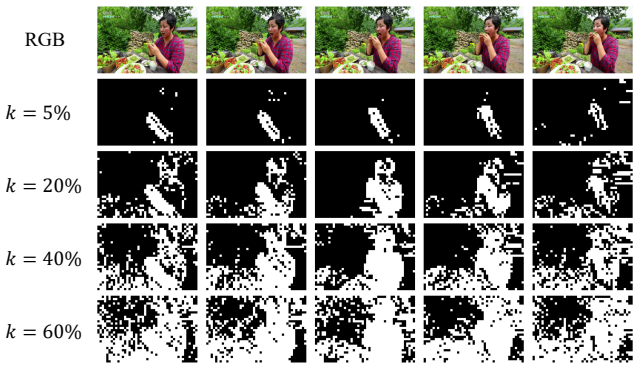


Figure 4. **Motion-saliency masks at different thresholds k.** Increasing k highlights a larger proportion of the most dynamic regions in the video.

Given an input video $\mathbf{V} \in \mathbb{R}^{F' \times H' \times W' \times 3}$, we first divide the video into non-overlapping tubulets of size $(P_\mathrm{t}, P_\mathrm{h}, P_\mathrm{w})$, resulting in patch tokens $\mathbf{Y} \in \mathbb{R}^{F \times H \times W \times 3 P_\mathrm{t} P_\mathrm{h} P_\mathrm{w}}$. We compute the temporal difference across frames and calculate L1 norm for each patch token as,

$$\Delta_{t,h,w} = \begin{cases} \left\| \mathbf{Y}_{t,h,w} - \mathbf{Y}_{t-1,h,w} \right\|_1, & t \in \{2, \dots, F\} \\ \Delta_{2,h,w}, & t = 1 \end{cases} \tag{11}$$

We identify dynamic regions by selecting the top-k (e.g., $k = 20\%$) tokens with the largest temporal differences for each frame, constructing a motion-saliency mask $\mathcal{M}$:

$$\mathcal{M} = \left\{ I(t,h,w) \;\middle|\; \Delta_{t,h,w} \geq r_t,\ \forall t \in \{1, \cdots, F\} \right\}, \tag{12}$$

where $I(\cdot)$ denotes the index of the patch token in the flattened dimension ($N = F \times H \times W$), and $r_t = \text{top-}k\left(\{\Delta_{t,h,w}\}_{h=1,w=1}^{H,W}\right)$ denotes the threshold that retains the highest $k\%$ of temporal differences at the t-th frame. The M-TSA loss then performs STSS alignment exclusively to these motion-salient tokens:

$$\mathcal{L}_{\text{M-TSA}} = \frac{1}{|\mathcal{M}|F} \sum_{i \in \mathcal{M}} \sum_{f=1}^{F} D_{\mathrm{KL}} \left(\mathbf{P}_{i,f}^{\mathrm{VFM}} \,\|\, \mathbf{P}_{i,f}^{\mathrm{VDM}} \right). \tag{13}$$

This selective alignment reduces computational and memory overhead by excluding static regions and enables the model to focus on capturing physically meaningful motions. We combine the M-TSA loss with the diffusion loss for finetuning video diffusion models as,

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{diffusion}} + \lambda \mathcal{L}_{\text{M-TSA}}. \tag{14}$$

5. Experiments

5.1. Implementation Details

We adopt CogVideoX-2B [52] and VideoMAEv2-B [46] as the base text-to-video diffusion model and visual foundation model, respectively. We employ a lightweight projector composed of a 3-layer MLP followed by a 3D convolutional layer, which adjusts the spatio-temporal resolution of the diffusion model's hidden representations to match that of the foundation model. We set the loss weight λ, temperature τ, and masking ratio k to 0.5, 0.1, and 20, respectively. We subsample 64K videos from OpenVid [35] for finetuning [56]. We fully finetune the model using AdamW [30] with a learning rate of 2×10^{-6} and batch size of 32 for 4000 iterations on eight NVIDIA RTX 6000 Ada GPUs.

5.2. Benchmarks

VideoPhy [2] is a benchmark designed to evaluate the physical plausibility of video generation models. It consists of 344 prompts involving interactions between various material types in the physical world, categorized into solid-solid, solid-fluid, and fluid-fluid interactions. The benchmark measures two metrics: *Semantic Adherence (SA)*, which assesses whether the generated video faithfully follows the text prompt, and *Physical Commonsense (PC)*, which evaluates whether the generated video exhibits physically realistic dynamics. We use the VideoConPhysics auto-evaluator to compute scores in the range $[0, 1]$ and report the proportion of samples with scores ≥ 0.5. For evaluation, we use the upsampled captions provided by VideoREPA [56].

VideoPhy2 [3] extends the evaluation to human-object interactions, in contrast to the material-centric focus of VideoPhy. It comprises 590 text prompts and measures SA and PC on an integer scale from 1 to 5 Additionally, the Joint score is reported as the proportion of samples where both SA ≥ 4 and PC ≥ 4. For automated evaluation, we employ the VideoPhy2-AutoEval model and utilize the upsampled prompts officially released by the benchmark [3].

VBench. [12] is a comprehensive benchmark suite for evaluating video generative models. It consists of 946 text prompts and evaluates two complementary aspects: video quality, which measures the perceptual quality of

Table 1. Quantitative results on VideoPhy benchmark. We report Semantic Adherence (SA) and Physical Commonsense (PC) scores as the percentage of videos achieving a score ≥ 0.5 across different interaction categories. $*$ finetuned on OpenVid subset.

Method	Overall		Solid-Solid		Solid-Fluid		Fluid-Fluid	
	SA	PC	SA	PC	SA	PC	SA	PC
Cosmos-Diffusion-7B [1]	57.0	18.0	-	-	-	-	-	-
DreamMachine [31]	57.5	21.8	55.1	21.7	59.6	23.3	58.2	18.2
VideoCrafter2 [8]	50.3	29.7	50.4	32.2	50.7	27.4	48.1	29.1
HunyuanVideo [22]	60.2	28.2	55.2	16.1	67.1	30.1	54.5	54.5
CogVideoX-2B [52]	59.9	25.0	44.8	15.4	71.9	22.6	67.3	56.4
CogVideoX-2B + PhyT2V [51]	42	29	-	-	-	-	-	-
CogVideoX-2B*	59.6	25.3	45.5	15.4	71.2	23.3	65.5	56.4
+$\mathcal{L}_{\text{REPA}}$ [54]	60.5	22.7	45.5	9.1	72.6	**27.4**	67.3	45.5
+$\mathcal{L}_{\text{TRD}}$ [56]	63.1	27.6	49.7	18.2	**74.0**	**27.4**	69.1	52.7
+$\mathcal{L}_{\text{TSA}}$ (ours)	62.8	29.1	51.8	19.6	70.6	26.0	**70.9**	61.8
+$\mathcal{L}_{\text{M-TSA}}$ (ours)	**64.5**	**30.8**	**53.9**	**21.0**	**74.0**	**27.4**	67.3	**65.5**

Table 2. Quantitative results on VideoPhy2. We report SA, PC, and Joint scores as the percentage of videos achieving scores ≥ 4.

Method	Joint	SA	PC
CogVideoX-2B	22.4	27.1	65.9
CogVideoX-2B*	22.9	26.8	68.0
+$\mathcal{L}_{\text{REPA}}$ [54]	23.3	27.5	66.8
+$\mathcal{L}_{\text{TRD}}$ [56]	23.1	27.0	68.4
+$\mathcal{L}_{\text{TSA}}$ (ours)	24.0	**28.2**	69.3
+$\mathcal{L}_{\text{M-TSA}}$ (ours)	**24.4**	**28.2**	**69.8**

the generated video independent to the text prompt, and video–condition consistency, which evaluates how faithfully the video aligns with the text prompt. Each aspect comprises 8 metrics, resulting in a total of 16 metrics. In addition to the per-metric scores, we report the aggregated Quality Score, Semantic Score, and the Total Score to summarize performance across these aspects. For evaluation, we use the upsampled prompts provided by NOVA [10].

5.3. Results

VideoPhy. We evaluate our method on VideoPhy and present the results in Tab. 1. Compared to the baseline CogVideoX-2B, applying TSA significantly improves the PC score, achieving 29.1% overall, which surpasses VideoREPA (27.6%). This improvement demonstrates the effectiveness of our tempered self-similarity alignment approach, which provides fine-grained motion supervision and stronger gradients for misaligned regions through KL divergence minimization. Adding background masking (M-TSA) further boosts performance, reaching an overall PC score of 30.8%, outperforming PhyT2V [51] (29.0%) that iteratively refines videos using physical knowledge from LLMs. The largest improvement is observed in the Solid-Solid category, where PC increases from 15.4% to 21.0% (36.4% relative improvement), with consistent gains also observed in Solid-Fluid (17.6%) and Fluid-Fluid (16.1%) interactions. These results validate that focusing self-similarity alignment on dynamic regions enables more effective learning of physically grounded motion dynamics.

VideoPhy2. We evaluate our method on VideoPhy2, a benchmark focused on human-object interactions, and summarize the results in Tab. 2. Consistent with the VideoPhy results, TSA outperforms VideoREPA, improving the Joint score from 23.1% to 24.0%. Motion masking further enhances performance, with M-TSA achieving a Joint score of 24.4%, demonstrating the consistent effectiveness of our approach across diverse human-object interaction scenarios.

VBench. To ensure that our method preserves overall video quality while improving physical plausibility, we evaluate on VBench and present the results in Tab. 3. Compared to the baseline, applying M-TSA preserves the overall video quality while yielding a slight improvement in the Total Score (81.0%→81.2%). This demonstrates that our method enhances physical plausibility without compromising the inherent video generation quality of the base model.

5.4. Ablation Studies

Temperature. In Fig. 5a, we evaluate the effect of temperature by gradually decreasing τ from 1.0 to lower values. We observe a clear performance improvement when decreasing τ from 1.0 to 0.1, indicating that a sharper correspondence distribution encourages the model to capture more fine-grained temporal relations, which facilitates precise motion modeling. However, lowering the temperature beyond this point leads to performance degradation. We conjecture that excessively sharp distributions suppress broader structural cues—such as object part-level geometry—that are necessary for coherent motion understanding. These results suggest that an appropriate balance between fine-grained motion cues and structural context is critical for effective self-similarity alignment.

Background Masking. In Fig. 5b, we analyze the effect of the masking ratio k, which determines the proportion of

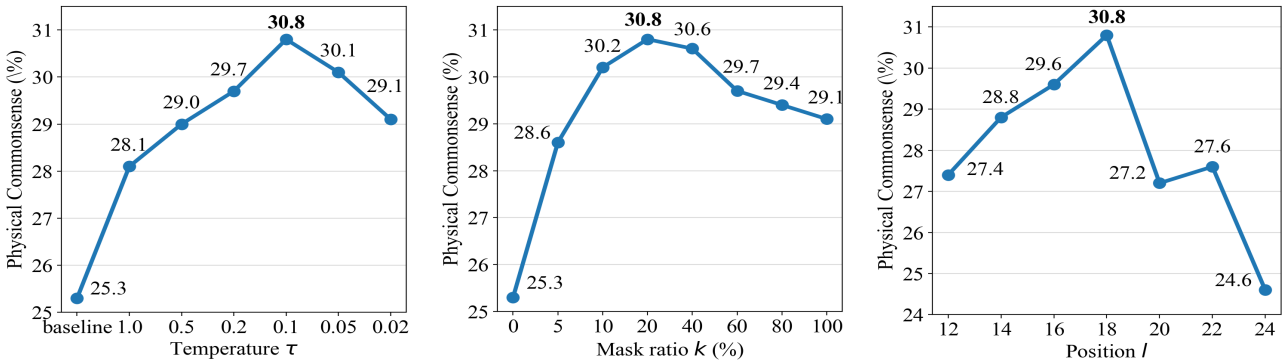
Table 3. **Quantitative results on VBench**. $*$ finetuned on OpenVid subset.

(a) Total scores and video quality metrics.

Method	Total Score	Quality Score	Subject Consist.	Background Consist.	Temporal Flickering	Motion Smoothness	Dynamic Degree	Aesthetic Quality	Imaging Quality	Object Class
CogVideoX-2B*	81.0	81.4	93.4	94.2	**98.0**	98.3	**56.9**	**63.4**	63.2	**91.3**
+$\mathcal{L}_{\text{M-TSA}}$ (ours)	**81.2**	**81.5**	**93.5**	**94.5**	**98.0**	**98.4**	56.1	63.0	**63.5**	90.2

(b) Video-condition consistency metrics.

Method	Semantic Score	Multiple Objects	Human Action	Color	Spatial Relationship	Scene	Appearance Style	Temporal Style	Overall Consistency
CogVideoX-2B*	79.4	71.6	98.2	**85.0**	75.5	52.2	24.4	**25.2**	27.1
+$\mathcal{L}_{\text{M-TSA}}$ (ours)	**79.7**	**73.0**	**98.8**	84.4	**75.9**	**53.4**	**24.5**	**25.2**	**27.2**



(a) Temperature τ. (b) Masking ratio k. (c) Position l.

Figure 5. Ablation studies on VideoPhy. The PC scores are reported.

dynamic regions selected for alignment. Remarkably, utilizing only the top 5% of dynamic tokens yields performance superior to the baseline, with the optimal performance achieved at approximately $k = 20\%$. As the ratio increases beyond this point, we observe a gradual decline in performance. This result validates our hypothesis that aligning static regions introduces redundancy and computational noise. By restricting the alignment to motion-salient regions, our method effectively prevents the model from being biased toward static backgrounds, allowing it to concentrate its learning capacity on generating physically grounded motion dynamics.

Alignment Position. In Fig. 5c, we evaluate the effectiveness of feature alignment across various depths of the denoising network. As a result, we identify the 18-th transformer block of CogVideoX as the optimal layer for injecting the correspondence priors and set as default for our main experiments.

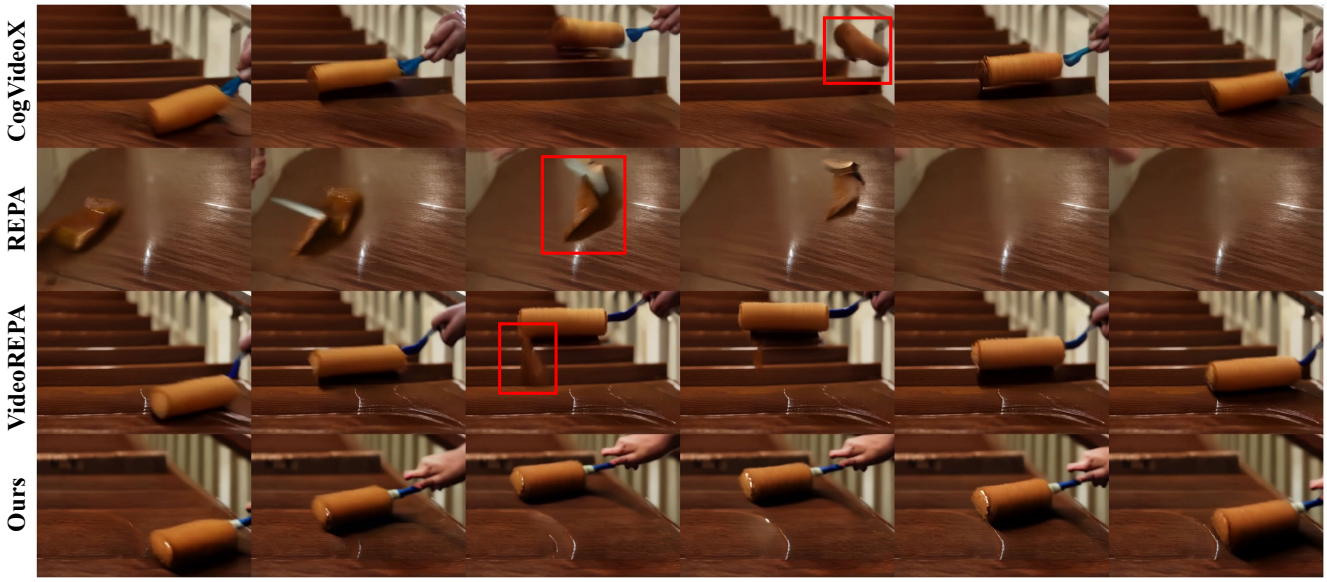
Different Architectures. To demonstrate the architectural generality of our approach, we here apply our loss to NOVA [10], an auto-regressive video generation model. In Tab. 4, our method consistently improves PC scores across all material types, despite structural differences between bi-directional diffusion and causal auto-regressive models. This validates that our method is model-agnostic, broadly applicable to a wide range of video generative frameworks.

Table 4. **Effect with NOVA [10] on VideoPhy**. $*$ finetuned on OpenVid subset.

method	overall		solid-solid		solid-fluid		fluid-fluid	
	SA	PC	SA	PC	SA	PC	SA	PC
NOVA-0.6B [10]	42.7	21.2	27.3	12.6	54.8	23.3	50.9	38.2
NOVA-0.6B*	44.5	20.1	31.5	12.6	55.5	21.2	49.1	36.4
+$\mathcal{L}_{\text{TRD}}$ [56]	**45.9**	22.7	30.8	12.6	**56.2**	23.3	**58.2**	47.3
+$\mathcal{L}_{\text{TSA}}$ (ours)	45.1	29.4	31.5	18.9	53.4	30.1	**58.2**	54.5
+$\mathcal{L}_{\text{M-TSA}}$ (ours)	45.1	**30.5**	**32.2**	**19.6**	53.4	**30.8**	56.4	**58.2**

5.5. Qualitative Results

In Fig. 6, we provide qualitative comparisons of videos generated by CogVideoX [52], REPA [54], VideoREPA [56], and our method. In Fig. 6a, existing methods exhibit appearance drift of the paint roller or generate unrealistic rolling motion, whereas our method preserves the object's appearance throughout the video while depicting realistic motion dynamics. In Fig. 6b, other methods generate physically implausible water splashes that violate causality, *e.g.*, splashes appearing in front of the jetski before it passes, while our method produces temporally coherent and physically plausible splash patterns. These results validate that our approach effectively transfers realistic motion dynamics



(a) Prompt: A paint roller is used to apply a coat of brown paint ...

(b) Prompt: Two jetskis race side-by-side, their wakes colliding ...

Figure 6. **Qualitative results.** Red rectangles highlight regions with physically implausible or temporally inconsistent motion. Our method generates videos with physically realistic dynamics.

embedded in STSS to video generative models. Please refer to the supplementary material for the full generated videos.

6. Conclusion

We have introduced Tempered Self-Similarity Alignment (TSA), a novel framework that transfers spatio-temporal correspondences in visual foundation models into video diffusion models for physically plausible video generation. By converting spatio-temporal self-similarity into probabilistic correspondence distributions and aligning such distributions exclusively on dynamically changing regions, our method enables effective alignment of physically grounded dynamics. Through extensive ablation studies, we found that controlling correspondence granularity plays a critical role in motion learning and that filtering out static regions via motion masking substantially improves performance. Our experiments on VideoPhy and VideoPhy2 demonstrate that TSA significantly improves the physical realism of generated videos in diverse real-world scenarios.

References

[1] Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025. 6

[2] 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. *arXiv preprint arXiv:2406.03520*, 2024. 1, 2, 5

[3] Hritik Bansal, Clark Peng, Yonatan Bitton, Roman Goldenberg, Aditya Grover, and Kai-Wei Chang. Videophy-2: A challenging action-centric physical commonsense evaluation in video generation. *arXiv preprint arXiv:2503.06800*, 2025. 1, 2, 5

[4] Aritra Bhowmik, Denis Korzhenkov, Cees GM Snoek, Amirhossein Habibian, and Mohsen Ghafoorian. Moalign: Motion-centric representation alignment for video diffusion models. *arXiv preprint arXiv:2510.19022*, 2025. 1

[5] 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. 1, 2, 3

[6] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. *OpenAI Blog*, 1(8):1, 2024. 1, 2

[7] Hila Chefer, Uriel Singer, Amit Zohar, Yuval Kirstain, Adam Polyak, Yaniv Taigman, Lior Wolf, and Shelly Sheynin. Videojam: Joint appearance-motion representations for enhanced motion generation in video models. *arXiv preprint arXiv:2502.02492*, 2025. 1, 2

[8] 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 *CVPR*, pages 7310–7320, 2024. 6

[9] Gilles Daviet and Florence Bertails-Descoubes. A semi-implicit material point method for the continuum simulation of granular materials. *ACM Transactions on Graphics (TOG)*, 35(4):1–13, 2016. 2

[10] Haoge Deng, Ting Pan, Haiwen Diao, Zhengxiong Luo, Yufeng Cui, Huchuan Lu, Shiguang Shan, Yonggang Qi, and Xinlong Wang. Autoregressive video generation without vector quantization. *arXiv preprint arXiv:2412.14169*, 2024. 6, 7

[11] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. 3

[12] 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024. 5

[13] 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. 2

[14] Hyeonho Jeong, Chun-Hao P Huang, Jong Chul Ye, Niloy J Mitra, and Duygu Ceylan. Track4gen: Teaching video diffusion models to track points improves video generation. In *CVPR*, pages 7276–7287, 2025. 1, 2

[15] Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. *arXiv preprint arXiv:2410.05954*, 2024. 1, 2, 3

[16] Imran N Junejo, Emilie Dexter, Ivan Laptev, and Patrick PÚrez. Cross-view action recognition from temporal self-similarities. In *ECCV*, 2008. 2

[17] Imran N Junejo, Emilie Dexter, Ivan Laptev, and Patrick Perez. View-independent action recognition from temporal self-similarities. *IEEE TPAMI*, 2010. 2

[18] Dahyun Kang, Heeseung Kwon, Juhong Min, and Minsu Cho. Relational embedding for few-shot classification. In *ICCV*, pages 8822–8833, 2021. 2

[19] Manjin Kim, Heeseung Kwon, Chunyu Wang, Suha Kwak, and Minsu Cho. Relational self-attention: What's missing in attention for video understanding. *NeurIPS*, 34:8046–8059, 2021. 2

[20] Manjin Kim, Paul Hongsuck Seo, Cordelia Schmid, and Minsu Cho. Learning correlation structures for vision transformers. In *CVPR*, pages 18941–18951, 2024. 2

[21] Seungryong Kim, Dongbo Min, Bumsub Ham, Sangryul Jeon, Stephen Lin, and Kwanghoon Sohn. Fcss: Fully convolutional self-similarity for dense semantic correspondence. In *CVPR*, 2017. 2

[22] 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. 1, 2, 3, 6

[23] Heeseung Kwon, Manjin Kim, Suha Kwak, and Minsu Cho. Motionsqueeze: Neural motion feature learning for video understanding. *arXiv preprint arXiv:2007.09933*, 2020. 2

[24] Heeseung Kwon, Manjin Kim, Suha Kwak, and Minsu Cho. Learning self-similarity in space and time as generalized motion for action recognition. *arXiv preprint arXiv:2102.07092*, 2021. 2

[25] Xingjian Leng, Jaskirat Singh, Yunzhong Hou, Zhenchang Xing, Saining Xie, and Liang Zheng. Repa-e: Unlocking vae for end-to-end tuning with latent diffusion transformers. *arXiv preprint arXiv:2504.10483*, 2025. 2, 3

[26] Jiajing Lin, Zhenzhong Wang, Yongjie Hou, Yuzhou Tang, and Min Jiang. Phy124: Fast physics-driven 4d content generation from a single image. *arXiv preprint arXiv:2409.07179*, 2024. 1, 2

[27] Jiajing Lin, Zhenzhong Wang, Shu Jiang, Yongjie Hou, and Min Jiang. Phys4dgen: A physics-driven framework for con-

trollable and efficient 4d content generation from a single image. *arXiv e-prints*, pages arXiv–2411, 2024. 1, 2

[28] Shaowei Liu, Zhongzheng Ren, Saurabh Gupta, and Shenlong Wang. Physgen: Rigid-body physics-grounded image-to-video generation. In *ECCV*, pages 360–378. Springer, 2024.

[29] Zhuoman Liu, Weicai Ye, Yan Luximon, Pengfei Wan, and Di Zhang. Unleashing the potential of multi-modal foundation models and video diffusion for 4d dynamic physical scene simulation. In *CVPR*, pages 11016–11025, 2025. 1, 2

[30] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[31] Luma AI. Dream machine — ai video generator, 2024. 6

[32] 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. *arXiv preprint arXiv:2410.05363*, 2024. 1, 2

[33] Antonio Montanaro, Luca Savant Aira, Emanuele Aiello, Diego Valsesia, and Enrico Magli. Motioncraft: Physics-based zero-shot video generation. *Advances in Neural Information Processing Systems*, 37:123155–123181, 2024. 1, 2

[34] Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models learn physical principles from watching videos?, 2025. *URL https://arxiv. org/abs/2501.09038*, 2025. 1, 2

[35] Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. *arXiv preprint arXiv:2407.02371*, 2024. 5

[36] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 3

[37] Daniel Ram, Theodore Gast, Chenfanfu Jiang, Craig Schroeder, Alexey Stomakhin, Joseph Teran, and Pirouz Kavehpour. A material point method for viscoelastic fluids, foams and sponges. In *Proceedings of the 14th ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pages 157–163, 2015. 2

[38] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 3

[39] Eli Shechtman and Michal Irani. Space-time behavior based correlation. In *CVPR*, pages 405–412. IEEE, 2005. 2

[40] Eli Shechtman and Michal Irani. Matching local self-similarities across images and videos. In *CVPR*, 2007. 2

[41] Alexey Stomakhin, Craig Schroeder, Lawrence Chai, Joseph Teran, and Andrew Selle. A material point method for snow simulation. *ACM Transactions on Graphics (TOG)*, 32(4): 1–10, 2013. 2

[42] Atousa Torabi and Guillaume-Alexandre Bilodeau. Local self-similarity-based registration of human rois in pairs of stereo thermal-visible videos. *Pattern Recognition*, 46(2): 578–589, 2013. 2

[43] 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. 1, 2, 3

[44] Heng Wang, Du Tran, Lorenzo Torresani, and Matt Feiszli. Video modeling with correlation networks. In *CVPR*, 2020. 2

[45] Jing Wang, Ao Ma, Ke Cao, Jun Zheng, Zhanjie Zhang, Jiasong Feng, Shanyuan Liu, Yuhang Ma, Bo Cheng, Dawei Leng, et al. Wisa: World simulator assistant for physics-aware text-to-video generation. *arXiv preprint arXiv:2503.08153*, 2025. 1, 2

[46] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14549–14560, 2023. 5

[47] 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. *arXiv preprint arXiv:2507.07982*, 2025. 2

[48] Wenhao Wu, Yuxin Song, Zhun Sun, Jingdong Wang, Chang Xu, and Wanli Ouyang. What can simple arithmetic operations do for temporal modeling? In *ICCV*, pages 13712–13722, 2023. 2

[49] Tianyi Xie, Zeshun Zong, Yuxing Qiu, Xuan Li, Yutao Feng, Yin Yang, and Chenfanfu Jiang. Physgaussian: Physics-integrated 3d gaussians for generative dynamics. In *CVPR*, pages 4389–4398, 2024. 1, 2

[50] Tianyi Xie, Yiwei Zhao, Ying Jiang, and Chenfanfu Jiang. Physanimator: Physics-guided generative cartoon animation. In *CVPR*, pages 10793–10804, 2025. 1, 2

[51] Qiyao Xue, Xiangyu Yin, Boyuan Yang, and Wei Gao. Phyt2v: Llm-guided iterative self-refinement for physics-grounded text-to-video generation. In *CVPR*, pages 18826–18836, 2025. 1, 2, 6

[52] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 1, 2, 3, 5, 6, 7

[53] Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast causal video generators. *arXiv e-prints*, pages arXiv–2412, 2024. 1, 2

[54] Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. *arXiv preprint arXiv:2410.06940*, 2024. 2, 3, 6, 7

[55] Tianyuan Zhang, Hong-Xing Yu, Rundi Wu, Brandon Y Feng, Changxi Zheng, Noah Snavely, Jiajun Wu, and William T Freeman. Physdreamer: Physics-based interaction with 3d objects via video generation. In *ECCV*, pages 388–406. Springer, 2024. 1, 2

[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. 2, 4, 5, 6, 7