LaMo: Self-Supervised Latent Motion Priors for Physical Realism in Video Generation

Bo Jiang[1,*], **Depu Meng**[1], **Yihan Hu**[1], **Yichen Xie**[1,2,*], **Tianshuo Xu**[1,*], **Wei Zhan**[1,2,†]

[1] Applied Intuition [2] University of California, Berkeley

https://lamo-ai.github.io/

Abstract

Modern video generators produce visually compelling clips but still struggle with physical and motion consistency, limiting their use as reliable world simulators. Existing remedies often rely on external simulators, teacher models, or curated physics-focused data. We explore a complementary self-supervised direction: extracting motion cues from the unlabeled videos already used to train video diffusion models. We propose LaMo, which formulates a latent motion prior over frame-to-frame latent changes conditioned on the current latent and prompt. This prior is exposed through two lightweight readouts: a macro motion drift used during training as a Motion Drift Loss, and a learned micro motion field used during sampling as Motion Prior Guidance. Both components are plug-and-play with existing video diffusion backbones, requiring no architectural or I/O changes. On VideoPhy, VideoPhy2, and PisaBench, LaMo improves CogVideoX backbones and outperforms recent physics-aware baselines that use external supervision. On VBench, it preserves overall generation quality while improving motion-related dimensions. These results suggest that unlabeled video contains useful motion supervision for improving physical fidelity in modern video diffusion models.

1 Introduction

Video generation has advanced rapidly in recent years, with modern diffusion- and transformer-based models producing high-resolution, temporally coherent, and visually compelling videos [Wiedemer et al., 2025, Wan et al., 2025, Gao et al., 2025]. This progress has motivated a broader view of video generators as potential world simulators for spatial reasoning, embodied planning, and synthetic data generation [Agarwal et al., 2025, Bar et al., 2025, Hafner et al., 2025]. However, visual realism alone is insufficient for this goal: a useful world simulator must also capture how the physical world evolves over time, making physical and motion fidelity a central challenge for modern video generation.

A persistent limitation of current video generators is that they often produce motion that appears plausible locally but violates basic physical regularities, such as gravity, contact, material interaction, or deformation consistency [Bansal et al., 2024, 2025, Wu et al., 2026]. This suggests a mismatch between the standard training objective and the desired notion of physical realism. The diffusion denoising loss supervises per-token reconstruction, but it does not explicitly target the frame-to-frame changes that define motion and physical evolution. As a result, temporal coherence can improve with scale, yet physically consistent dynamics remain difficult to guarantee.

Recent methods address this issue by introducing additional sources of physical knowledge. Some approaches couple video generation with simulators or explicit dynamics models [Liu et al., 2026, Foo et al., 2026, Yuan et al., 2026], while others rely on physics-aware rewards, curated physics datasets, or distillation from auxiliary foundation models [Tan et al., 2026, Zhang et al., 2025b,

* Work done during internship at Applied Intuition. † Corresponding author: wei.zhan@applied.co.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

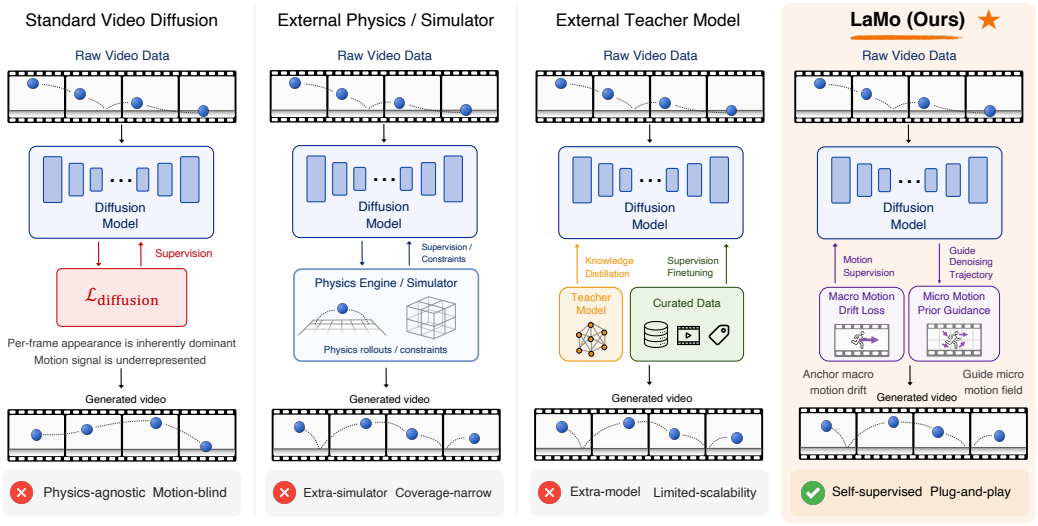


Figure 1: Existing approaches to physical realism rely on hand-crafted physics simulators (Col. 2) or on foundation-model teachers and curated physics-annotated corpora (Col. 3), while the vanilla diffusion objective is dominated by per-frame appearance (Col. 1). LaMo (Col. 4) instead extracts motion supervision from the same raw video data through a Motion Drift Loss during training and Motion Prior Guidance during inference.

Wang et al., 2025b]. These methods have shown promising improvements, but they also introduce external dependencies, domain-specific assumptions, or additional supervision. In contrast, large-scale unlabeled video already contains rich evidence of how objects move, interact, and deform. This motivates a simpler question: *can the motion signal already present in ordinary video data be extracted and reused to improve physical fidelity in video diffusion models?*

We propose LaMo (Latent Motion Prior), a self-supervised framework that models physical and motion regularities through frame-to-frame latent changes. Given clean video latents z, we define the τ-step latent difference $\Delta_\tau z := z^{(i+\tau)} - z^{(i)}$ and formulate a latent motion prior $p(\Delta_\tau z \mid z, c)$, where c denotes the conditioning already used by the backbone. This prior requires no physical annotations, simulators, or teacher models: every pair of nearby frames in an unlabeled video provides a training signal.

To make this prior usable in existing diffusion pipelines, LaMo exposes it through two lightweight readouts. First, a macro motion drift captures the channel-wise rate of latent change and is used during training as a Motion Drift Loss. Second, a learned micro motion field predicts spatially resolved latent changes and is used during sampling as Motion Prior Guidance. Both components are designed to preserve the original backbone architecture and input/output interface.

Experiments on VideoPhy [Bansal et al., 2024], VideoPhy2 [Bansal et al., 2025], and VBench [Huang et al., 2024] show that LaMo improves physical and motion fidelity while maintaining general generation quality. PisaBench [Li et al., 2025] adds a reference-based object-motion evaluation. Compared with recent physics-aware methods that use external supervision, LaMo achieves competitive or stronger results using only self-supervised signals from unlabeled video. These results suggest that latent motion differences provide a useful and minimally invasive source of supervision for improving physical realism in modern video diffusion models.

Our contributions are summarized as follows:

- **Self-supervised latent motion prior.** We formulate physical and motion regularities as a latent motion prior over frame-to-frame latent changes, which can be estimated directly from unlabeled video without simulators, physical annotations, or teacher models.
- **Lightweight training and sampling readouts.** We introduce a macro Motion Drift Loss for training and a micro Motion Prior Guidance for sampling, both of which preserve the backbone architecture and input/output interface.

- **Improved physical fidelity with preserved general video quality.** On VideoPhy and VideoPhy2, LaMo improves CogVideoX backbones and is competitive with recent externally supervised physics-aware baselines; on VBench, it maintains overall video generation quality while improving motion-related metrics. PisaBench provides complementary object-motion evidence.

2 Related Work

Physical and Motion Fidelity in Video Generation. Recent diffusion- and transformer-based video generators have advanced substantially in text alignment, visual quality, and temporal coherence [Yang et al., 2024, Kong et al., 2024, Wan et al., 2025, Gao et al., 2025]. Yet high visual fidelity does not guarantee physically faithful dynamics: physics-oriented evaluations show frequent violations of gravity, contact, material interaction, deformation, and conservation-related commonsense [Bansal et al., 2024, 2025]. Motion-focused analyses similarly show that clips improving appearance need not improve temporal dynamics [Wu et al., 2026]. These findings make physical and motion fidelity a central bottleneck for reliable world simulation, which requires plausible temporal evolution rather than only plausible frames.

Physics-Aware Video Generation. Existing attempts mostly inject additional physical or motion knowledge into the generator. One line grounds generation in explicit priors, using simulator-produced trajectories, reconstructed physical properties, physical simulation, dynamical equations, or learnable Newtonian dynamics as conditioning or generation signals [Wang et al., 2025a, Zhang et al., 2025a, Liu et al., 2026, Foo et al., 2026, Yuan et al., 2026]. Motion-aware objectives or inference-time guidance further compensate for appearance-oriented training objectives that underemphasize motion coherence [Chefer et al., 2025]. These methods provide controllability for targeted phenomena, but reliance on solvers, material parameters, 3D reconstruction, scene-specific states, or specialized motion representations limits scalability to unconstrained videos, especially when such signals must be estimated before generation.

A second line supplies physical knowledge through post-training objectives, curated data, or auxiliary foundation models. Reward- and preference-based methods enforce Newtonian constraints, collision rules, force-conditioned goals, VLM-derived preferences, or 3D consistency through RL-style fine-tuning [Le et al., 2025, Zhang et al., 2026, Gillman et al., 2026, Cai et al., 2025, Wang et al., 2026a]. WISA decomposes physical principles into textual, qualitative, and quantitative descriptors and injects them through Mixture-of-Physical-Experts attention trained on curated physics data [Wang et al., 2025b]. VideoREPA distills token-level physical relations from video foundation models into text-to-video diffusion transformers [Zhang et al., 2025b]. Other efforts align generation with VLM/VFM-derived physical cues, latent physical dynamics, or local physics annotations [Wang et al., 2026b, Shen et al., 2026, Satish et al., 2026, Pathak et al., 2026]. These methods are effective but still depend on external teachers, annotations, handcrafted rewards, simulator rollouts, or architectural components. In all cases, the physical signal is imported from an auxiliary source; we instead extract supervision from ordinary unlabeled videos already consumed by the generator.

Self-Supervised Video Dynamics. Self-supervised video learning provides evidence that raw video contains rich dynamics supervision. Masked reconstruction methods such as VideoMAE learn transferable video representations from unlabeled clips, but primarily target recognition and representation transfer [Tong et al., 2022, Wang et al., 2023]. Latent predictive models offer a more direct connection to dynamics: V-JEPA, V-JEPA 2, and LeWorldModel predict masked spatio-temporal regions or compact latent states while encoding physical structure [Bardes et al., 2024, Assran et al., 2025, Maes et al., 2026]. Probing and counterfactual world-modeling studies further show that latent predictors can acquire intuitive physics and dynamics-relevant structures such as keypoints, flow, and segmentations [Garrido et al., 2025, Venkatesh et al., 2024]. Interactive world models such as Genie and VideoWorld demonstrate that unlabeled videos can support control, planning, and transferable task knowledge [Bruce et al., 2024, Ren et al., 2025, 2026]. However, these works mainly target representation learning, planning, or interactive rollout, so their dynamics knowledge is not directly consumed by high-fidelity text-to-video diffusion models; LaMo converts raw-video dynamics into direct training and sampling signals.

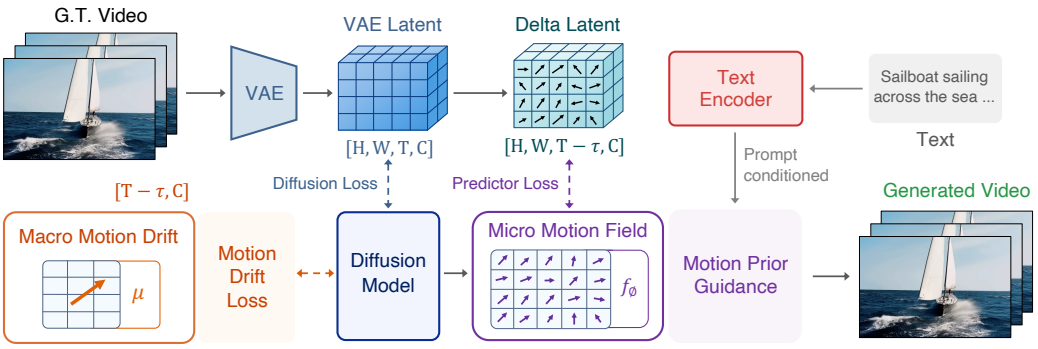


Figure 2: **LaMo Overview.** Frame-to-frame differences of clean VAE latents yield a self-supervised latent motion prior, exposed through two readouts: a *macro motion drift* used as a training-time *Motion Drift Loss*, and a lightweight learned *micro motion field* used at sampling as *Motion Prior Guidance*. Both share the backbone's prompt conditioning and leave its architecture unchanged.

3 Method

We organize LaMo around the method overview's data flow. First, ğ3.1 encodes raw videos into clean video latents and defines the latent motion prior over frame-to-frame latent changes. Second, ğ3.2 extracts the prior's stable macro component and uses it as a training-time *Motion Drift Loss*. Third, ğ3.3 learns a spatially resolved micro motion field and applies it as sampling-time *Motion Prior Guidance*. Finally, ğ3.4 summarizes the full training and inference pipeline, including the added cost.

3.1 Problem Setup and Latent Motion Encoding

Notation. Let x be a video clip and $z_{1:T} = \mathcal{E}(x) \in \mathbb{R}^{T\times C\times H\times W}$ its latent sequence encoded by a pretrained 3D causal VAE $\mathcal{E}$; T counts latent frames after temporal compression. We use $i \in \{1,\dots,T\}$ for the *frame index* and $t \in \{1,\dots,S\}$ for the *diffusion timestep*; $\bar{\alpha}_t \in (0,1]$ is the cumulative noise schedule and $\sigma_t := \sqrt{1-\bar{\alpha}_t}$. We write c for any auxiliary conditioning the backbone already accepts. The diffusion forward process produces $z_t^{(i)} = \sqrt{\bar{\alpha}_t}\, z_0^{(i)} + \sigma_t\, \epsilon^{(i)}$, and the backbone is trained with the standard per-token denoising loss $\mathcal{L}_{\text{denoise}} = \mathbb{E}_{t,z_0,\epsilon}\big[\|\epsilon_\theta(z_t,t,c) - \epsilon\|_2^2\big]$. We write $\hat{x}_0(\hat{\epsilon}, z_t, t)$ for the data-side projection of the model state, recoverable in closed form under either ϵ- or v-prediction.

The latent change object. Physical and motion realism is a property of *how* latents change between frames, not of *what* each latent contains. We therefore take as the carrier of dynamics the τ-step latent change on clean latents,

$$\Delta_\tau z^{(i)} := z_0^{(i+\tau)} - z_0^{(i)} \in \mathbb{R}^{C\times H\times W}, \qquad i \in \{1,\dots,T-\tau\}. \tag{1}$$

The choice $\tau \in \mathbb{N}$ sets the temporal scale: smaller values capture shorter-range changes, while larger values emphasize lower-frequency drift.

Latent motion prior. The standard denoising objective remains the main supervision for appearance and prompt-conditioned video synthesis, but it does not explicitly supervise the clean latent increment $\Delta_\tau z$ that carries frame-to-frame motion. We therefore collect the motion signal available in unlabeled videos into a conditional prior,

$$p(\Delta_\tau z \mid z, c), \tag{2}$$

where the conditioning variables are already present in the diffusion pipeline. This prior is self-supervised: every pair of nearby frames in an ordinary training clip provides a sample of $\Delta_\tau z$, requiring no simulator state, physics annotation, or external teacher.

Macro and micro readouts. We use the first moment of the prior,

$$m(z,c) := \mathbb{E}[\Delta_\tau z \mid z, c] \in \mathbb{R}^{C\times H\times W}, \tag{3}$$

as the operational target of LaMo. It naturally separates into a spatially averaged macro component and a spatially resolved micro component. The macro drift is:

$$\mu(z, c) := \mathbb{E}_{H,W}[m(z, c)] \in \mathbb{R}^C, \tag{4}$$

broadcastable over $H \times W$, while the residual field describes where and how this drift is allocated across the scene. This split gives two complementary readouts:

- **Macro readout (motion drift).** The channel-wise drift $\mu(z, c)$ admits a parameter-free closed-form estimator on any minibatch of clean clips, $\hat{\mu} = \frac{1}{HW} \sum_{h,w} \Delta_\tau z$. No predictor is fit; the observed statistic supplies a self-supervised target for the prior's first moment at this granularity.
- **Micro readout (motion field).** The full first moment is high-dimensional, so we model it by a small predictor $f_\phi(z, c) \approx m(z, c)$ trained on raw latent pairs.

Although both readouts come from the same prior, their granularity and estimation stability make them better suited to different stages of the diffusion pipeline. The macro motion drift is parameter-free and stable at minibatch scale, making it a natural auxiliary target during backbone fine-tuning. At inference, however, it would provide only a spatially uniform correction, i.e., a single vector in $\mathbb{R}^C$ broadcast over $H \times W$, and therefore cannot indicate where motion should occur. Conversely, the learned micro field captures spatially resolved motion, but using it as a training-time target would introduce an additional source of co-evolving supervision whose errors may interact with backbone optimization. We therefore use the macro readout as a training-time loss (ğ3.2) and the micro readout as an inference-time guidance term (ğ3.3), where the predictor is frozen and only steers the sampling trajectory. This stage-matched design is supported empirically in ğ4.3: replacing the channel-mean drift target with dense per-pixel motion supervision degrades both physical-fidelity axes relative to our method (Table 6).

3.2 Macro Readout: Latent Motion Drift Loss

We now derive the training-time **Motion Drift Loss** from the macro drift μ.

Anchor on $\hat{x}_0$. We anchor the loss on $\hat{x}_0$, the data-side projection of the model state and the very target that $\mathcal{L}_{\text{denoise}}$ optimizes, so that the auxiliary term composes additively without rescaling between timesteps; an analogous constraint on $\hat{\epsilon}$ would change units between steps, while one placed on z_t would not depend on θ at all. The same anchor carries over to inference (ğ3.3) with the differentiation variable switched from θ to $\hat{\epsilon}$, a deliberate symmetry between the two stages. We write $\hat{\mu}^{(i)} := \mathbb{E}_{H,W}[\hat{x}_0^{(i+\tau)} - \hat{x}_0^{(i)}]$ and $\mu^{\star,(i)} := \mathbb{E}_{H,W}[z_0^{(i+\tau)} - z_0^{(i)}]$ for the predicted and ground-truth drifts at frame index i.

Lag choice. The lag τ controls the temporal bandwidth of the supervision. We use τ=2 by default: in the two-point VideoPhy ablation, it outperforms τ=1 (Table 6). This is consistent with a possible contribution from short-range VAE jitter, but the ablation does not isolate that mechanism or compare latent with pixel-space targets.

Scale-normalized L2. We seek a distance that decouples *magnitude* from *direction* error without introducing a free trade-off hyperparameter. The scale-normalized form:

$$\mathcal{L}_{\text{drift}} \;=\; \mathbb{E}_{b,i}\left[\frac{\|\hat{\mu}^{(i)} - \mu^{\star,(i)}\|_2^2}{\text{sg}\big[\|\mu^{\star,(i)}\|_2^2 + \varepsilon\big]}\right] \;\approx\; (r-1)^2 \;+\; 2\,r\,(1 - \cos\vartheta), \tag{5}$$

with $r := \|\hat{\mu}\|/\|\mu^\star\|$ and $\vartheta := \angle(\hat{\mu}, \mu^\star)$, reduces to the right-hand decomposition when ε is negligible. The small constant ε prevents quiet clips from dominating the gradient when $\|\mu^\star\|$ is close to zero, while the stop-gradient $\text{sg}[\cdot]$ keeps the denominator as a per-clip rescaling factor rather than a learnable target.

Final training objective. At high noise levels $\hat{x}_0$ is dominated by $\hat{\epsilon}$ and is uninformative about the data drift, so we damp the loss with a per-batch averaged schedule weight $w(\sigma_t) := \mathbb{E}_b[\bar{\alpha}_{t,b}]$ that vanishes at maximal noise. The final training objective is:

$$\mathcal{L}_{\text{train}} \;=\; \mathcal{L}_{\text{denoise}} \;+\; \lambda_{\text{drift}} \cdot w(\sigma_t) \cdot \mathcal{L}_{\text{drift}}, \tag{6}$$

which is parameter-free apart from the scalar λ_{drift} and the numerical stabilizer ε. The schedule weight suppresses the auxiliary loss in the high-noise regime where $\hat{x}_0$ is unreliable, while the normalized denominator prevents low-motion clips from producing unstable gradients. By construction, $\mathcal{L}_{\text{drift}}$ constrains only the macro rate of latent change; the spatial allocation of motion is handled by the inference-time micro readout.

3.3 Micro Readout: Motion Prior Guidance

The micro readout f_ϕ is consumed at sampling time as a CFG-style [Ho and Salimans, 2022] gradient on the noise prediction. We first describe the predictor itself, then justify why this signal is best applied in noise-prediction space rather than directly on the latent.

Predictor and training. $f_\phi : \mathbb{R}^{C\times H\times W} \times \mathbb{R}^{D_c} \to \mathbb{R}^{C\times H\times W}$ is a small CNN (∼10M parameters, against ∼2–5B for the backbones we evaluate): an input projection, N conditioned residual blocks each terminating in an SE channel-recalibration unit [Hu et al., 2018], and an output projection, with the conditioning c (defined in ğ3.1) injected via FiLM [Perez et al., 2018]. SE recalibrates channels using pooled spatial context, while FiLM injects prompt-conditioned feature modulation; these are practical design choices, not theoretically necessary components. The output projection, the FiLM heads, and a learnable null-prompt parameter are zero-initialized, so f_ϕ launches as the safe constant-zero predictor and the inference-time gradient is well-defined throughout early training.

f_ϕ is trained on clean ground truth latent pairs from the same dataset the backbone is trained on with an MSE-plus-cosine loss:

$$\mathcal{L}_\phi \;=\; \text{MSE}\big(f_\phi(z,c),\ \Delta_\tau z\big) \;+\; \alpha\left(1-\cos\angle\big(f_\phi(z,c),\ \Delta_\tau z\big)\right), \qquad \alpha = 0.5, \tag{7}$$

that disentangles direction from magnitude at the spatial-field granularity. Two practicalities matter. *(i) Diffusion-aligned input augmentation.* At inference f_ϕ is fed $\hat{x}_0$, which carries timestep-dependent residual noise rather than a clean latent; with probability $p_{\text{aug}} = 0.5$ we therefore replace the input by $z + \sigma_{t'}\epsilon$ for t' sampled uniformly from the scheduler timestep grid, matching the sampling-time input distribution without modifying the diffusion forward process. *(ii) Classifier-free training.* With probability $p_{\text{drop}} = 0.2$ the prompt is replaced by the learnable null embedding, so f_ϕ jointly learns conditional and unconditional behaviour and remains well-defined when no prompt is supplied at sampling time.

Why guidance on $\hat{\epsilon}$ rather than on z_t. We apply motion guidance to the noise prediction $\hat{\epsilon}$ rather than directly editing the latent z_t: editing z_t desynchronizes the sampler trajectory from the noise estimate used by the next update, whereas editing $\hat{\epsilon}$ keeps the prior aligned with the diffusion update law. Appendix E provides the full rationale and ablation evidence.

Guidance loss and inference rule. At sampling step s, after the standard CFG mix has produced $\hat{\epsilon}_{\text{CFG}}$, we form the per-pair motion-consistency loss:

$$\mathcal{L}_{\text{guide}} \;=\; \frac{1}{T-\tau}\sum_{i=1}^{T-\tau}\Big\|\hat{x}_0^{(i)} + f_\phi\big(\hat{x}_0^{(i)}, c\big) \;-\; \hat{x}_0^{(i+\tau)}\Big\|_2^2, \tag{8}$$

which asks $\hat{x}_0^{(i+\tau)}$ to be exactly one τ-step motion increment away from $\hat{x}_0^{(i)}$ as predicted by f_ϕ. We then apply guidance only after the first $\lfloor(1-\rho)\,S\rfloor$ sampling steps have elapsed (default $\rho = 0.85$, indexing s from 0 at the high-noise end): early steps are dominated by appearance assembly where $\hat{x}_0$ is too noisy for f_ϕ to be reliable, and gating guidance to the lower-noise window lets motion supervision act precisely when it can. The full inference rule is:

$$\hat{\epsilon}_{\text{guided}} \;=\; \hat{\epsilon}_{\text{CFG}} \;-\; \lambda_{\text{guide}} \cdot \mathbf{1}\big\{s \geq \lfloor(1-\rho)\,S\rfloor\big\} \cdot \nabla_{\hat{\epsilon}}\mathcal{L}_{\text{guide}}\big(\hat{x}_0(\hat{\epsilon}, z_t, t),\, c\big), \tag{9}$$

applied after the CFG mix so that the gradient is taken with respect to the prompt-consistent noise estimate. Eq. 9 adds a single forward-and-backward through the small CNN per sampling step; f_ϕ is frozen and no architectural change to the scheduler or backbone is required. We refer to this update as **Motion Prior Guidance**, denoted $\mathcal{G}_{\text{motion}}$.

Table 1: VideoPhy results. SA measures semantic adherence and PC measures physical commonsense. Bold and underline denote the best and second-best values within each comparison block.

Methods	Extra Supervision	Solid-Solid		Solid-Fluid		Fluid-Fluid		Overall	
		SA	PC	SA	PC	SA	PC	SA	PC
CogVideoX-2B	-	49.6	13.3	71.2	28.1	60.0	50.9	60.5	25.6
DreamWorld-1.3B	VGGT, DINOv2	<u>54.5</u>	**24.5**	48.6	25.4	60.1	32.7	52.9	26.2
MoAlign-2B (paper)	VideoMAE	24.7	31.7	66.9	40.7	67.3	56.4	49.3	39.4
MoAlign-2B (reimpl.)	VideoMAE	54.6	<u>18.3</u>	73.5	<u>31.9</u>	<u>66.2</u>	<u>55.7</u>	<u>64.5</u>	<u>30.1</u>
VideoREPA-2B	VideoMAEv2	52.4	18.2	**77.4**	**32.2**	60.0	52.7	64.2	29.7
LaMo-2B (Ours)	Self-supervised	**58.7**	16.8	<u>74.7</u>	**32.2**	**69.1**	**67.3**	**67.2**	**31.4**
CogVideoX-5B	-	**62.9**	19.6	76.0	33.6	72.7	61.8	70.0	32.3
PhyT2V-5B	o1-preview	-	-	-	-	-	-	61	37
WISA-5B	Qwen2VL	-	-	-	-	-	-	67	38
PHANTOM-5B	V-JEPA2	-	-	-	-	-	-	47.5	37.9
MoAlign-5B (reimpl.)	VideoMAE	<u>62.5</u>	26.3	79.6	38.4	78.0	<u>76.2</u>	<u>72.2</u>	39.4
VideoREPA-5B	VideoMAEv2	58.0	**28.0**	**82.9**	<u>39.0</u>	**80.0**	74.5	72.1	<u>40.1</u>
LaMo-5B (Ours)	Self-supervised	**62.9**	<u>26.6</u>	<u>80.8</u>	**41.1**	<u>78.2</u>	**78.2**	**73.0**	**41.0**

Table 2: Quantitative results on the VideoPhy2 benchmark.

Method	SA	PC
CogVideoX-2B	21.0	68.0
PHANTOM-5B	27.8	71.7
MoAlign-2B (paper)	28.8	75.0
MoAlign-2B (reimpl.)	<u>24.6</u>	<u>73.1</u>
VideoREPA-2B	21.0	72.5
LaMo-2B (Ours)	**25.4**	**75.4**

Table 3: Quantitative results on the PisaBench benchmark, aggregated over 421 videos.

Method	L2 ↓	CD ↓	IoU ↑
CogVideoX-5B-I2V	<u>0.1180</u>	<u>0.3017</u>	<u>0.1481</u>
Cosmos-Predict2.5	0.1373	0.3871	0.1467
Cosmos3-Nano	0.1250	0.3437	0.1361
LTX-Video	0.1181	0.3071	0.1017
OpenSoraV2	0.1293	0.3310	0.0846
LaMo-5B (Ours)	**0.1113**	**0.2888**	**0.1590**

3.4 Full Pipeline and Cost

We instantiate LaMo on top of the open-source CogVideoX [Yang et al., 2024] family. At training time, the diffusion backbone is fine-tuned with $\mathcal{L}_{\text{train}}$ (Eq. 6) against a macro drift target read directly from each minibatch, while the micro motion-field predictor f_ϕ is trained on clean latent pairs from the same video data via Eq. 7. Since f_ϕ depends only on clean latents and prompt conditioning, not on the current backbone state, the two objectives can be optimized separately or in parallel.

At inference time, the backbone architecture, input/output interface, and backbone parameter count are unchanged. The only added component is the frozen $\sim$10M-parameter predictor f_ϕ, which contributes one lightweight forward-and-backward pass during the active guidance window of Eq. 9. Thus LaMo surfaces motion supervision from the same raw videos used for diffusion fine-tuning, without simulators, physical annotations, or foundation-model teachers, while keeping the pretrained video diffusion backbone itself intact.

4 Experiments

4.1 Experimental Setup

We evaluate on VideoPhy [Bansal et al., 2024], VideoPhy2 [Bansal et al., 2025], and VBench [Huang et al., 2024]. VideoPhy/VideoPhy2 report Semantic Adherence (SA) and Physical Commonsense (PC), while VBench measures general video quality and motion-related dimensions. PisaBench [Li et al., 2025] provides reference-based object-motion metrics. We separately report an image-conditioned Physics-IQ evaluation, using its own matched control and protocol. We fine-tune CogVideoX [Yang et al., 2024] backbones on OpenVid [Nan et al., 2024] following VideoREPA's protocol [Zhang et al., 2025b], without physics annotations, simulator rollouts, or external teacher signals. Full training and inference details are provided in Appendix B.

Table 4: Results on VBench. LaMo improves all reported dimensions over CogVideoX, indicating that the motion prior preserves video quality while strengthening motion and spatial consistency.

Method	Motion Smooth.	Multi. Obj.	Obj. Class	Overall Consist.	Scene	Spatial Relation.	Temp. Flicker.	Quality Score	Semantic Score	Total Score
CogVideoX	97.6	50.4	78.7	25.0	40.3	52.3	97.3	80.5	68.7	78.2
LaMo (Ours)	**98.2**	**51.6**	**82.0**	**25.7**	**42.6**	**62.2**	**98.4**	**81.9**	**70.7**	**79.6**

Table 5: Key component ablation on VideoPhy.

Component	SA	PC
Baseline	64.8	35.2
+ $\mathcal{L}_{\text{drift}}$ only	<u>71.8</u>	<u>39.0</u>
+ $\mathcal{G}_{\text{motion}}$ only	68.9	38.4
LaMo (full)	**73.0**	**41.0**

Table 6: Design-choice ablation on VideoPhy.

Variant	SA	PC
Dense motion loss	64.2	34.6
Raw L2 motion loss	67.3	36.1
Adj-frame lag (τ=1)	<u>68.0</u>	37.5
Direct latent edit (z_t)	62.8	31.6
No predictor aug.	66.4	<u>37.6</u>
LaMo (Ours)	**73.0**	**41.0**

4.2 Main Results

VideoPhy. Table 1 evaluates whether LaMo improves physical fidelity without sacrificing semantic adherence. Relative to CogVideoX-2B, LaMo-2B improves the overall score from $60.5/25.6$ to $67.2/31.4$ SA/PC, a gain of $+6.7$ SA and $+5.8$ PC. It also improves over the reproduced 2B physics-aware baselines on the overall SA/PC pair, while using only self-supervised video signals. At the larger scale, LaMo-5B improves CogVideoX-5B from $70.0/32.3$ to $73.0/41.0$, and is slightly higher than VideoREPA-5B on both overall axes ($72.1/40.1$). The largest PC gains appear in the *Fluid–Fluid* category, where LaMo reaches 67.3 PC at 2B and 78.2 PC at 5B. This category involves less rigid, harder-to-script dynamics, so the improvement is consistent with our goal of extracting motion regularities from ordinary video rather than relying on hand-specified physics supervision.

VideoPhy2. On the more challenging VideoPhy2 benchmark (Table 2), which targets a broader and harder distribution of physics-laden prompts, LaMo-2B reaches 25.4 SA / 75.4 PC, improving over the CogVideoX-2B baseline by $+4.4$ SA / $+7.4$ PC and over the strongest prior 2B competitor, VideoREPA-2B ($21.0/72.5$), by $+4.4$ SA / $+2.9$ PC. Notably, LaMo-2B also surpasses PHANTOM-5B ($27.8/71.7$) on PC despite being less than half its size, while approaching its SA score. These results indicate that the gains observed on VideoPhy generalize to a markedly more demanding and diverse distribution of physical phenomena.

PisaBench. On the 421-video evaluation in Table 3, LaMo-5B achieves the lowest trajectory L2 and CD and the highest IoU among the evaluated methods. These results show its strong performance on PisaBench relative to existing video generation methods.

VBench. Table 4 checks whether the physics-oriented objective harms general generation quality. LaMo improves the CogVideoX Total Score from 78.2 to 79.6, with gains in both Quality ($80.5 \rightarrow 81.9$) and Semantic ($68.7 \rightarrow 70.7$). The largest motion-related change is Spatial Relationship ($52.3 \rightarrow 62.2$), while Motion Smoothness and Temporal Flickering also increase slightly. Since all reported VBench dimensions move upward in this comparison, the results suggest that the latent motion prior improves physical and spatial consistency without an observed trade-off in the measured general-quality metrics.

4.3 Ablation Study

Effect of each component. Table 5 ablates the two core components of LaMo on VideoPhy. Starting from the baseline at 64.8 SA / 35.2 PC, the training-time Motion Drift Loss alone improves to $71.8/39.0$ ($+7.0/+3.8$), showing that a macro latent-change target already improves physical fidelity without modifying sampling. The inference-time Motion Prior Guidance alone reaches $68.9/38.4$ ($+4.1/+3.2$), indicating that the frozen motion-field predictor also contributes useful spatial motion structure. Combining both readouts gives the best result, $73.0/41.0$, higher than either single com-

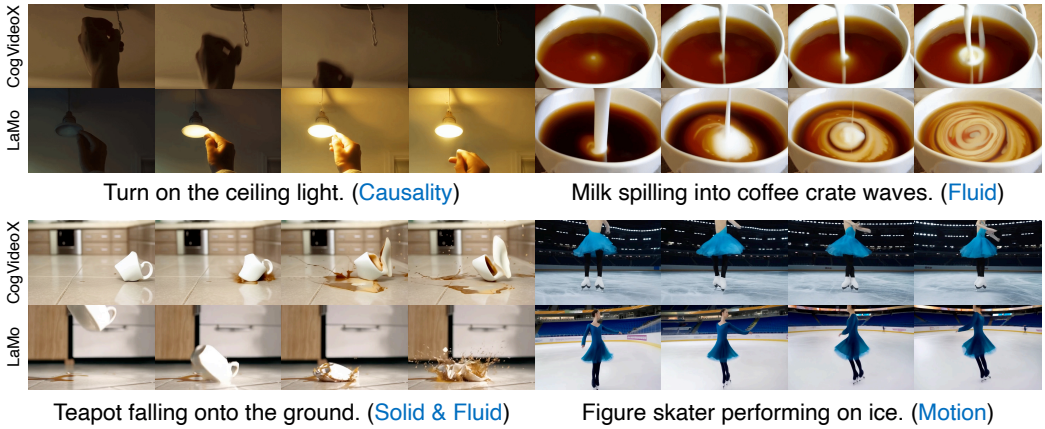


Figure 3: **Qualitative comparisons on physics-heavy prompts.** LaMo produces more coherent motion and more consistent relationships between actions and their effects.

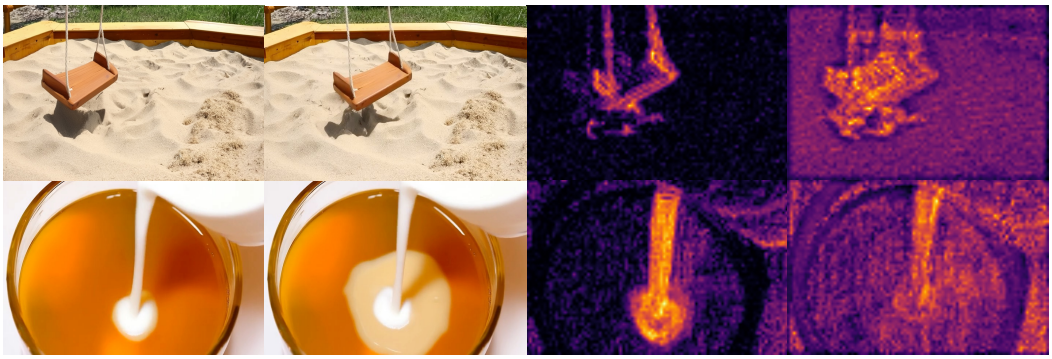


Figure 4: **Interpretability of LaMo's latent motion readouts.** For each clip, we show sampled frames, a Motion Drift Heatmap, and a Motion Field Heatmap. The drift heatmap visualizes the macro latent-change direction supervised by $\mathcal{L}_{\text{drift}}$; the field heatmap shows the response of f_ϕ used in $\mathcal{G}_{\text{motion}}$. Both focus on physically active regions. Appendix A describes the visualization.

ponent on both SA and PC; this supports our stage-matching argument that macro drift anchors the global latent-change rate during training, while micro guidance allocates that change spatially during sampling.

Validating individual design choices. Table 6 validates the main design choices by replacing each with the corresponding alternative. On the drift side, dense per-pixel supervision (64.2/34.6), raw L2 (67.3/36.1), and adjacent-frame lag τ=1 (68.0/37.5) all trail the full method (73.0/41.0), supporting the channel-mean target, scale-normalized loss, and τ=2 lag. On the guidance side, direct latent editing of z_t drops to 62.8/31.6, below the baseline, while removing predictor augmentation gives 66.4/37.6; these results support noise-space guidance and diffusion-aligned predictor training rather than post-hoc alternatives.

4.4 Qualitative Analysis

Qualitative comparisons. Figure 3 compares generated sequences across causal state changes, fluid dynamics, object-fluid impacts, and articulated human motion. CogVideoX produces plausible individual frames, but its temporal dynamics often fail to reflect the prompted events: switching on the light does not consistently change illumination, milk entering coffee shows limited waves or mixing, teapot contact and spilling are unstable, and the skater's legs become tangled across frames. In contrast, LaMo better links actions to their consequences, producing consistent illumination, coher-

ent waves and mixing, a localized splash and spill, and continuous skating motion with stable body structure.

Heatmap interpretability. Figure 4 shows that the macro motion drift and micro motion field heatmaps concentrate on active regions, such as the moving swing and milk stream, rather than spreading uniformly over the scene. This localization supports the qualitative observation that LaMo improves event-level dynamics rather than merely changing frame appearance. Appendix A details the heatmap computation.

5 Conclusions and Limitations

Conclusions. We presented LaMo, a self-supervised framework that extracts a latent motion prior from ordinary video and exposes it as a training-time Motion Drift Loss and sampling-time Motion Prior Guidance. Without simulators, teacher models, or physics annotations, LaMo improves physical fidelity on VideoPhy, VideoPhy2, and PisaBench while preserving general generation quality on VBench.

Limitations. LaMo encourages plausible latent motion but is not a constraint-satisfying simulator. Its gains are bounded by the coverage of the video data and by the granularity of VideoPhy-style model-judged metrics, which do not isolate contact, deformation, fluids, conservation, or long-horizon stability. Future work should pair latent motion priors with targeted physical measurements and integrate them more directly into large-scale pretraining.

Latent differences are compressed feature changes, not explicit displacement or trajectory fields. They can mix object motion with camera motion, illumination, or occlusion; VAE compression may hide fine-scale contact, and spatial averaging can attenuate localized or opposing motion. The spatially resolved micro readout mitigates the last issue but does not remove these limitations. LaMo is complementary to explicit correspondence cues such as optical flow. The reference-based evaluations cover specific motion settings and do not replace a controlled human study or establish long-horizon physical correctness.

We use latent differences to supervise the generator in its native state space, but have not established that they are cleaner than RGB differences through a controlled pixel-target comparison. VAE compression may suppress some variation while retaining appearance and camera changes.

Broader impacts. More physically realistic generated videos could support simulation and synthetic-data creation. However, generated videos may still contain physically implausible motion or interactions, limiting their reliability in these applications. Evaluation and release should account for misleading uses of generated content as well as failures in physical fidelity.

References

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.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

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.

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.

Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. In *CVPR*, 2025.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*, 2024.

Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *ICML*, 2024.

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

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.

Lin Geng Foo, Mark He Huang, Alexandros Lattas, Stylianos Moschoglou, Thabo Beeler, and Christian Theobalt. Physical simulator in-the-loop video generation. In *CVPR*, 2026.

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

Quentin Garrido, Nicolas Ballas, Mahmoud Assran, Adrien Bardes, Laurent Najman, Michael Rabbat, Emmanuel Dupoux, and Yann LeCun. Intuitive physics understanding emerges from self-supervised pretraining on natural videos. *arXiv preprint arXiv:2502.11831*, 2025.

Nate Gillman, Yinghua Zhou, Zitian Tang, Evan Luo, Arjan Chakravarthy, Daksh Aggarwal, Michael Freeman, Charles Herrmann, and Chen Sun. Goal force: Teaching video models to accomplish physics-conditioned goals. *arXiv preprint arXiv:2601.05848*, 2026.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, pages 1–7, 2025.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, 2018.

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 *CVPR*, 2024.

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.

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

Chenyu Li, Oscar Michel, Xichen Pan, Sainan Liu, Mike Roberts, and Saining Xie. PISA experiments: Exploring physics post-training for video diffusion models by watching stuff drop. In *ICML*, 2025.

Wei Liu, Ziyu Chen, Zizhang Li, Yue Wang, Hong-Xing Yu, and Jiajun Wu. Realwonder: Real-time physical action-conditioned video generation. *arXiv preprint arXiv:2603.05449*, 2026.

Lucas Maes, Quentin Le Lidec, Damien Scieur, Yann LeCun, and Randall Balestriero. Leworldmodel: Stable end-to-end joint-embedding predictive architecture from pixels. *arXiv preprint arXiv:2603.19312*, 2026.

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles? In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 948–958, 2026.

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.

Saurabh Pathak, Elahe Arani, Mykola Pechenizkiy, and Bahram Zonooz. Physvid: Physics aware local conditioning for generative video models. In *CVPR*, 2026.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *AAAI*, 2018.

Zhongwei Ren, Yunchao Wei, Xun Guo, Yao Zhao, Bingyi Kang, Jiashi Feng, and Xiaojie Jin. Videoworld: Exploring knowledge learning from unlabeled videos. In *CVPR*, 2025.

Zhongwei Ren, Yunchao Wei, Xiao Yu, Guixun Luo, Yao Zhao, Bingyi Kang, Jiashi Feng, and Xiaojie Jin. Videoworld 2: Learning transferable knowledge from real-world videos. *arXiv preprint arXiv:2602.10102*, 2026.

Siddarth Nilol Kundur Satish, Devesh Jaiswal, Hongyu Chen, and Abhishek Bakshi. Physvideogenerator: Towards physically aware video generation via latent physics guidance. *arXiv preprint arXiv:2601.03665*, 2026.

Ying Shen, Jerry Xiong, Tianjiao Yu, and Ismini Lourentzou. Phantom: Physics-infused video generation via joint modeling of visual and latent physical dynamics. In *CVPR*, 2026.

Boming Tan, Xiangdong Zhang, Ning Liao, Yuqing Zhang, Shaofeng Zhang, Xue Yang, Qi Fan, and Yanyong Zhang. Dreamworld: Unified world modeling in video generation. *arXiv preprint arXiv:2603.00466*, 2026.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *NeurIPS*, 2022.

Rahul Venkatesh, Honglin Chen, Kevin Feigelis, Daniel M Bear, Khaled Jedoui, Klemen Kotar, Felix Binder, Wanhee Lee, Sherry Liu, Kevin A Smith, et al. Understanding physical dynamics with counterfactual world modeling. In *ECCV*, 2024.

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.

Chen Wang, Chuhao Chen, Yiming Huang, Zhiyang Dou, Yuan Liu, Jiatao Gu, and Lingjie Liu. Physctrl: Generative physics for controllable and physics-grounded video generation. In *NeurIPS*, 2025a.

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*, 2025b.

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 *CVPR*, 2023.

Weijie Wang, Xiaoxuan He, Youping Gu, Yifan Yang, Zeyu Zhang, Yefei He, Yanbo Ding, Xirui Hu, Donny Y. Chen, Zhiyuan He, Yuqing Yang, and Bohan Zhuang. World-r1: Reinforcing 3d constraints for text-to-video generation. *arXiv preprint arXiv:2604.24764*, 2026a.

Zijun Wang, Panwen Hu, Jing Wang, Terry Jingchen Zhang, Yuhao Cheng, Long Chen, Yiqiang Yan, Zutao Jiang, Hanhui Li, and Xiaodan Liang. Prophy: Progressive physical alignment for dynamic world simulation. In *CVPR*, 2026b.

Thaddäus Wiedemer, Yuxuan Li, Paul Vicol, Shixiang Shane Gu, Nick Matarese, Kevin Swersky, Been Kim, Priyank Jaini, and Robert Geirhos. Video models are zero-shot learners and reasoners. *arXiv preprint arXiv:2509.20328*, 2025.

Xindi Wu, Despoina Paschalidou, Jun Gao, Antonio Torralba, Laura Leal-Taixé, Olga Russakovsky, Sanja Fidler, and Jonathan Lorraine. Motion attribution for video generation. *arXiv preprint arXiv:2601.08828*, 2026.

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.

Yu Yuan, Xijun Wang, Tharindu Wickremasinghe, Zeeshan Nadir, Bole Ma, and Stanley H Chan. Newtongen: Physics-consistent and controllable text-to-video generation via neural newtonian dynamics. In *ICLR*, 2026.

Haoze Zhang, Tianyu Huang, Zichen Wan, Xiaowei Jin, Hongzhi Zhang, Hui Li, and Wangmeng Zuo. Physchoreo: Physics-controllable video generation with part-aware semantic grounding. *arXiv preprint arXiv:2511.20562*, 2025a.

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

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. In *NeurIPS*, 2025b.

A Heatmap Computation for Figure 4

We compute the two heatmaps in Figure 4 on the same generated latent trajectory so that the macro and micro readouts are visualized at the same physical event. Let $z \in \mathbb{R}^{1\times T\times C\times H\times W}$ denote the clean video latent produced for a prompt c, and let $\tau = 2$ be the same temporal lag used in training. For each valid latent frame index t, we compute

$$\Delta_t = z^{(t+\tau)} - z^{(t)}, \qquad b_t = \frac{1}{HW}\sum_{h,w} \Delta_t(:,h,w) \in \mathbb{R}^C, \tag{10}$$

where b_t is the empirical macro drift vector for that frame pair. We choose the visualization index

$$t^\star = \arg\max_t \|b_t\|_2, \tag{11}$$

which selects the pair with the largest spatially averaged latent change. The RGB frame shown in the first column is the decoded video frame corresponding to $t^\star$ under the VAE temporal compression.

Motion Drift Heatmap. The Motion Drift Heatmap visualizes how strongly each spatial cell's latent change aligns with the macro drift direction. For the selected frame pair, we compute

$$R_{\text{drift}}(h,w) = \frac{|\langle \Delta_{t^\star}(:,h,w),\, b_{t^\star}\rangle|}{\|b_{t^\star}\|_2 + \varepsilon}. \tag{12}$$

This quantity is parameter-free: it is the per-spatial-cell projection of the observed latent difference onto the channel-mean drift direction supervised by $\mathcal{L}_{\text{drift}}$. Taking the absolute value visualizes the magnitude of aligned latent change regardless of the sign of the projection.

Motion Field Heatmap. The Motion Field Heatmap visualizes the spatial response of the learned predictor f_ϕ. A direct norm $\|f_\phi(z^{(t^\star)}, c)(:,h,w)\|_2$ is often dense because a CNN predictor can produce a non-zero default response even in static background regions. We therefore subtract a no-motion baseline. Specifically, we form a temporally averaged static latent

$$z_{\text{static}} = \frac{1}{T}\sum_{t=1}^{T} z^{(t)}, \tag{13}$$

and compute

$$R_{\text{field}}(h,w) = \left\| f_\phi(z^{(t^\star)}, c)(:,h,w) - f_\phi(z_{\text{static}}, c)(:,h,w) \right\|_2. \tag{14}$$

This time-averaged baseline retains aggregate scene context and reduces the predictor's dense background response. The subtraction is used only for visualization, not for training or sampling.

B Experimental Setup Details

Benchmarks. VideoPhy [Bansal et al., 2024] contains 344 prompts spanning three interaction categories: *Solid–Solid*, *Solid–Fluid*, and *Fluid–Fluid*. VideoPhy and VideoPhy2 [Bansal et al., 2025] report Semantic Adherence (SA), measuring whether the generated content matches the prompt, and Physical Commonsense (PC), measuring whether the generated dynamics obey elementary physical regularities. PisaBench [Li et al., 2025] measures image-conditioned object motion using trajectory L2, Chamfer Distance, and IoU. We also evaluate image-conditioned continuation on Physics-IQ [Motamed et al., 2026]; these protocols differ from the text-to-video VideoPhy tests. VBench [Huang et al., 2024] evaluates general video generation quality along sixteen dimensions; in the main paper we report Quality, Semantic, and Total scores together with seven motion- and consistency-related sub-dimensions: Motion Smoothness, Temporal Flickering, Multiple Objects, Object Class, Overall Consistency, Scene, and Spatial Relationship.

Training data. Following VideoREPA [Zhang et al., 2025b], we fine-tune all backbones on a clean subset of OpenVid [Nan et al., 2024]. LaMo-2B is trained on 32k clips and LaMo-5B on 64k clips, both filtered to 480p and 49 frames per clip. No physics annotations, simulator rollouts, or external teacher signals are used.

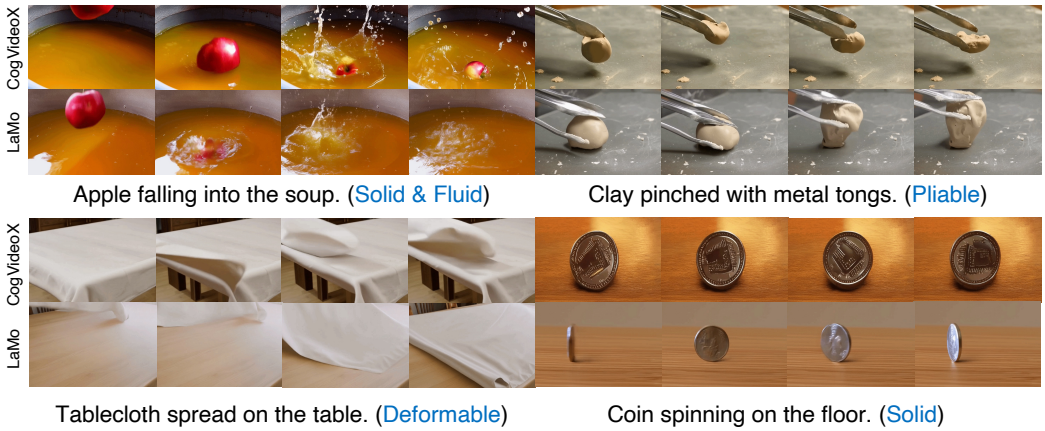


Figure 5: **Additional qualitative results on physics-heavy prompts.** The sampled frames compare liquid impact, cloth spreading, contact deformation, and rotational continuity.

Implementation details. We instantiate LaMo on CogVideoX-2B and CogVideoX-5B [Yang et al., 2024]. The 2B model is fully fine-tuned, while the 5B model uses LoRA [Hu et al., 2022] with rank 128 and alpha 64. All experiments use 16 A100 GPUs, global batch size 32, and 2k optimizer steps. We set the latent-change lag to τ=2 and the Motion Drift Loss weight to λ_{drift}=0.4. The motion-field predictor f_ϕ has $\sim$10M parameters and is trained with diffusion-aligned input augmentation probability p_{aug}=0.5 and classifier-free dropout probability p_{drop}=0.2. At inference, Motion Prior Guidance uses λ_{guide}=25.0 and active-window ratio ρ=0.85.

C Additional Qualitative Comparisons

To further substantiate the qualitative observations in the main paper, Figure 5 shows four additional prompts that stress different physical regimes: solid–fluid interaction, deformable cloth, pliable material, and rigid-body angular motion. For each prompt, we generate videos with the CogVideoX baseline and with LaMo under the same prompt, seed, and sampler configuration, and visualize representative frames sampled at matched temporal positions. Across these examples, LaMo shows more coherent event progression in the sampled frames.

Apple falling into soup (solid–fluid). The prompt describes an apple dropping into a bowl of soup. Both sampled sequences show an impact, while LaMo depicts more pronounced local liquid displacement and a deeper entry in the shown frames. These images illustrate the impact sequence; they do not establish the apple's eventual buoyant state.

Tablecloth spread on the table (deformable). The prompt describes a tablecloth being spread onto a table. The baseline frames show the cloth mostly laid out from the outset, while LaMo shows fabric above the table followed by spreading onto the surface, better matching the event order in the prompt.

Clay pinched with metal tongs (pliable). The prompt depicts tongs pinching a clay block. In the sampled frames, the baseline shows limited local deformation; LaMo shows a clearer indentation near contact and a change in shape. This supports a qualitative comparison of contact deformation, but does not establish the material's plastic response.

Coin spinning on the floor (solid). The prompt depicts a coin spinning on a flat surface. Across the sampled frames, the baseline changes the coin's orientation irregularly, while LaMo shows a more continuous progression of orientation and tilt. These frames support a qualitative comparison of rotational continuity, not a measured claim about angular velocity or friction.

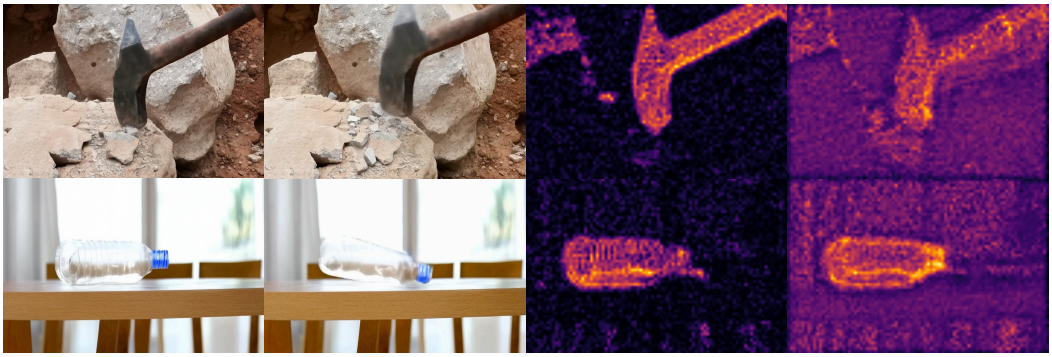


Figure 6: **Additional interpretability visualizations of LaMo's latent motion readouts.** Each row shows a sampled RGB frame, Motion Drift Heatmap, and Motion Field Heatmap (Appendix A). Top: *hammer striking a stone*. Bottom: *water bottle rolling on a table*. Both heatmaps emphasize regions associated with the depicted motion.

Summary. Across these four prompts, the sampled LaMo frames show more coherent impact, cloth spreading, contact deformation, and rotation. These visual examples complement the quantitative results; sampled frames alone cannot establish complete physical trajectories.

D Additional Interpretability Visualizations

We further visualize LaMo's two latent motion readouts on a hammer-strike clip and a rolling-bottle clip in Figure 6. Following Appendix A, we select the latent frame pair with the strongest macro drift and show the corresponding decoded frame, the projected latent-change heatmap, and the predictor-response heatmap. The time-averaged baseline reduces dense background response only in this visualization.

Hammer striking a stone. The sampled frame depicts a hammer striking a stone. Both heatmaps respond near the hammer head and contact region, with a more localized field response. This links the readouts to the visible action without measuring impact forces.

Water bottle rolling on a table. In the rolling-bottle clip, both heatmaps respond on the bottle rather than the background. The field heatmap resolves more of its outline, consistent with a spatially resolved readout; the images do not quantify rolling speed.

Discussion. In both examples, the visualized responses are strongest near the moving objects and contact regions rather than the background. These heatmaps help interpret the two readouts, but do not by themselves establish physical correctness.

E Additional Details on Guidance Space

A natural alternative to Eq. 9 is to apply the motion-consistency gradient directly to the latent z_t, in the spirit of latent-space score editing. We avoid this choice because the diffusion sampler defines its latent update through $\hat{\epsilon}$ and the noise schedule. An external edit on z_t therefore acts at the wrong scale for the current step, desynchronizing the latent trajectory chosen by the scheduler from the noise estimate expected by the next step. This mismatch can compound across sampling and visibly degrade sample quality.

A correction specified on $\hat{x}_0$ can be mapped algebraically to the scheduler's native prediction; this affine conversion does not itself require an iterative inversion or imply accumulated inversion error. We apply the motion-consistency gradient directly to the native prediction so that the scheduler interface and timestep scaling are explicit. This follows the same principle as classifier-free guidance, which corrects the model prediction rather than editing the sampled trajectory. This keeps the

Table 7: Matched evaluation on Wan2.2-TI2V-5B.

Method	VideoPhy		VideoPhy2		
	SA ↑	PC ↑	SA ↑	PC ↑	Joint ↑
Wan2.2-5B baseline	65.70	**30.52**	25.76	68.47	23.05
Wan2.2-5B + LaMo	**65.99**	29.65	**26.61**	**70.17**	**23.22**

motion prior aligned with the diffusion update law and gives a built-in time-dependent attenuation through $\hat{x}_0$: by the chain rule, the guidance prefactor scales as $\sigma_t/\sqrt{\bar{\alpha}_t}$ for ϵ-prediction or σ_t for v-prediction, so the correction naturally shrinks as $\bar{\alpha}_t \to 1$.

The ablation in Table 6 supports this design choice. The direct-z_t variant drops to 62.8 SA / 31.6 PC, below the unaugmented backbone, consistent with the risk of scheduler mismatch, although this ablation does not isolate that mechanism.

F Physics-IQ Evaluation

Physics-IQ requires continuation from a fixed initial frame. We therefore trained matched image-to-video CogVideoX-5B control and LaMo adaptations on the same OpenVid data, with identical adapter initialization, optimizer, and 8-GPU, $2{,}000$-step budget. Only the LaMo arm adds the motion-field predictor and motion objectives. This shorter adaptation budget differs from the main CogVideoX-5B recipe.

We evaluate all 38 scenarios of the Original Physics-IQ Solid Mechanics category from three fixed camera views (114 conditions). The matched LaMo adaptation scores 38.56 versus 36.31 for its CogVideoX-5B control, a 2.25-point improvement on this category.

G Wan2.2 Flow-Matching Transfer

To probe transfer beyond CogVideoX, we ported LaMo to Wan2.2-TI2V-5B, which has a different 48-channel causal spatiotemporal VAE and a flow-matching sampler. We re-encoded OpenVid with Wan's VAE, recomputed latent-motion targets, trained a 48-channel motion-field predictor, and fine-tuned matched baseline and LaMo models using the same data, optimizer, seed, $2{,}000$-step budget, and LoRA rank/alpha of $128/64$. The target is VAE-specific; plug-and-play refers to retaining the target backbone architecture and I/O, not zero-shot transfer of a previously trained predictor.

For the linear-flow convention in this implementation, $x_t = (1-\sigma_t)x_0 + \sigma_t\epsilon$, $u = \epsilon - x_0$, and the model-side clean-latent estimate is $\hat{x}_0 = x_t - \sigma_t\hat{u}$. The drift loss remains on $\hat{x}_0$; at sampling, guidance acts on the native predicted field $\hat{u}$ through $\nabla_{\hat{u}}\mathcal{L}_{\text{guide}}(\hat{x}_0, c)$. Since $\partial\hat{x}_0/\partial\hat{u} = -\sigma_t I$, the corresponding correction to $\hat{x}_0$ is proportional to $-\sigma_t^2\nabla_{\hat{x}_0}\mathcal{L}_{\text{guide}}$ and attenuates as $\sigma_t \to 0$. This affine conversion introduces no iterative inversion.

Table 7 shows small improvements on VideoPhy SA and all three VideoPhy2 scores, alongside a decline in VideoPhy PC. It provides initial evidence of transfer to another VAE and denoising parameterization, rather than a universal improvement guarantee.

Wan2.2 still uses a causal spatiotemporal VAE, so this experiment does not establish transfer to a framewise 2D VAE. Because τ is measured in latent frames, a different VAE temporal stride changes the physical duration of the target difference and may require retuning τ.

H Inference Cost and Sensitivity

Inference cost. We measured end-to-end generation on one A100-80GB with bfloat16, 720×480 output, 49 frames, and 50 denoising steps, excluding model loading. At strength 25, guidance runs for the final 43 steps. Full LaMo adds 4.656% runtime and 0.248% peak memory relative to the public CogVideoX-2B run. On the matched predictor path, turning on guidance adds 1.311% runtime and no additional backbone forward pass. The 43-step window uses the default $\rho = 0.85$.

Table 8: Inference metrics on one A100-80GB. Mean $\pm$ SD over eight matched trials.

Condition	E2E time (s/video) $\downarrow$	Peak mem. (GB) $\downarrow$	Runtime increase (%) $\downarrow$	Memory increase (%) $\downarrow$
Public CogVideoX-2B	95.308 ± 1.407	8.929	0	0
LaMo adapter only	99.780 ± 0.825	8.929	$+4.708$	0
LaMo predictor, guidance off	98.443 ± 0.917	8.951	$+3.312$	$+0.248$
Full LaMo	99.725 ± 0.914	8.951	$+4.656$	$+0.248$

The adapter-only and guidance-off rows use different inference call paths, so their times are not a component timing comparison.

Table 9: Guidance strength and active-window sensitivity on VideoPhy2.

λ_{guide}	ρ	SA $\uparrow$	PC $\uparrow$
15	0.80	25.25	74.41
20	0.80	24.75	74.92
25	0.80	23.56	74.07
25	0.85	**25.42**	**75.42**

Guidance sensitivity. Table 9 reports the available VideoPhy2 sweep. Across the tested settings, SA ranges from 23.56 to 25.42 and PC from 74.07 to 75.42. The default $(\lambda_{\text{guide}}, \rho) = (25, 0.85)$ has the highest scores among these tested settings; this limited sweep does not establish a global optimum.

Table 10: Joint SE and FiLM ablation in the micro predictor on VideoPhy.

Micro predictor	SA $\uparrow$	PC $\uparrow$
Vanilla CNN	62.79	28.49
CNN + SE + FiLM	**63.37**	**30.23**

Micro-predictor architecture. The matched 8-GPU, 2,000-step VideoPhy ablation in Table 10 compares a vanilla CNN with a predictor containing both SE and FiLM. The joint addition improves SA by 0.58 and PC by 1.74 points; this experiment does not isolate the modules individually.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes, the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Yes, the paper discuss the limitations of this work in Section 5.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: This paper does not present formal theoretical results requiring proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Yes, the paper fully disclose all necessary details to reproduce the experimental results, including implementation details, hyperparameters, and experimental settings.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Yes, all datasets used in this paper are publicly available. We will release the source code, model weights, training and evaluation scripts to facilitate future research.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Yes, the paper specifies all training and test details, including data splits, hyperparameters, optimizer choices, and how they were selected.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Yes, the paper includes appropriate statistical significance information to ensure the reliability of the experimental results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Yes, the paper provides sufficient information about the computational resources used to reproduce the experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Yes, the research conducted in this paper fully adheres to the NeurIPS Code of Ethics, ensuring ethical standards in data use, experimentation, and reporting.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Yes, the paper discusses the potential societal impacts of our work.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [No]

    Justification: LaMo inherits potential misuse risks from video generation models, as discussed in Section 5 (Broader Impacts). However, no specific safeguards have been implemented in this work.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Yes, all assets used in this paper have been properly credited and are in compliance with their respective licenses and terms of use.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Yes, all new assets introduced in the paper are well documented, and the documentation is provided alongside the assets to facilitate future use.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.