PHYSFORCE: PHYSICS-AWARE FORCE CONTROL FOR INTERACTIVE VIDEO WORLD MODEL

Anonymous authors
Paper under double-blind review

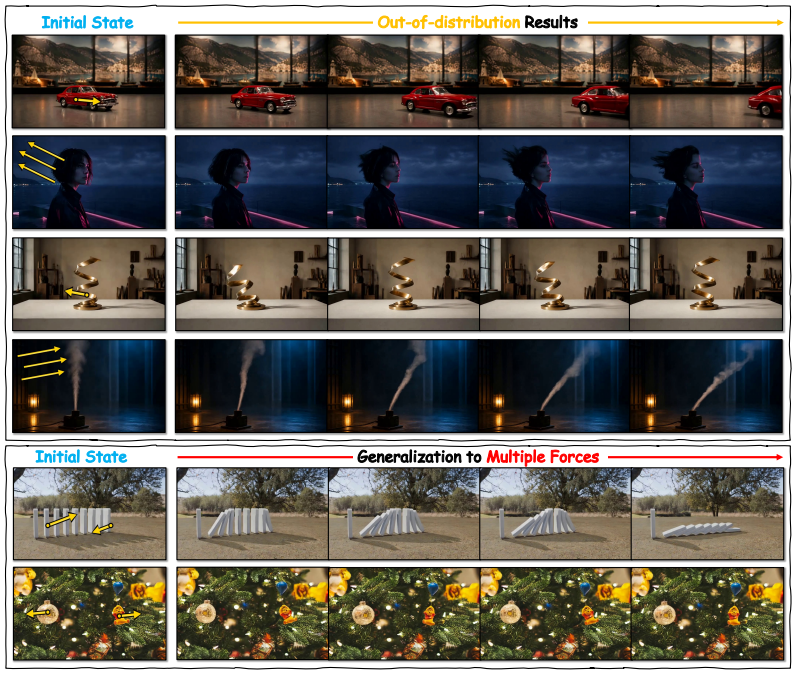

Figure 1: **Showcase of PhysForce under point and wind forces.** PhysForce generates physically consistent videos across diverse scenes and shows the potential to extend from single force to multiple force control. A single arrow denotes a point force, while three arrows denote wind. Arrow length and orientation indicate the force magnitude and direction respectively.

ABSTRACT

Interactive video world models aim to predict how visual environments evolve in response to user interventions, including physically consistent state changes induced by applied forces. Existing methods often represent forces as motion trajectories, which may introduce inappropriate motion priors, causing the model to over-rely on encoded displacement patterns. More fundamentally, the model's internal representations do not reliably localize force-induced state changes, making it difficult to associate applied forces with their resulting physical responses. Also, limited force-annotated data make it difficult to produce physically consistent and temporally stable predictions in unseen scenarios. To address these limitations, we propose PhysForce, a two-stage framework that integrates supervised fine-tuning and reinforcement learning through three complementary components. During supervised fine-tuning, (1) the Latent Force Intervention Encoder maps force location, direction, and magnitude into the video latent space without exposing future motion; and (2) the Physics Attribution Network associates force interventions with localized changes in the latent state, strengthening the learning of physical responses to applied forces. During reinforcement learning, (3) Physics-Aware Reinforcement Learning uses self-supervised rewards on out-of-distribution videos to improve physical consistency and generalization. Extensive experiments show that PhysForce improves force responsiveness, physical consistency, and generalization, demonstrating its effectiveness for interactive video world models. Our anonymous project page is available at `https://physforceproject.github.io/`.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

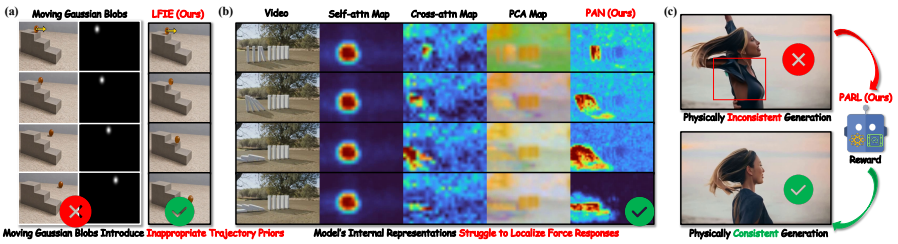

Figure 2: (a) Existing methods encode forces as moving Gaussian blobs, introducing inappropriate trajectory priors that may produce physically implausible responses. (b) Existing methods' internal representations such as self-attention, cross-attention, and PCA features fail to reliably localize force-induced state changes, whereas PAN provides more accurate response localization. (c) PARL further improves the physical consistency of generated videos.

1 INTRODUCTION

Interactive video world models aim to simulate how visual environments evolve in response to user interventions Bruce et al. (2024); Chen et al. (2026); Huang et al. (2025). For force-conditioned interaction, the model must predict future states from an initial state and a force defined by its application point, direction, and magnitude, while ensuring that the resulting dynamics are physically plausible and consistent with the applied force.

Existing methods Gillman et al. (2026a;b); Wang et al. (2026a) introduce force conditioning to model physical interventions, yet face three limitations. First, trajectory-based force representations may introduce inappropriate motion priors, causing models to over-rely on encoded displacement patterns rather than infer force induced results, as shown in Fig. 2(a). Second, internal representations such as attention maps and PCA features do not reliably localize force-induced state changes, as shown in Fig. 2(b). This limits their ability to provide explicit supervision for physical response regions. Third, limited annotated force data make it difficult to produce physically consistent and temporally stable predictions in unseen scenarios, as shown in Fig. 2(c).

To address these limitations, we propose PhysForce, a two-stage framework that integrates supervised fine-tuning and reinforcement learning through three complementary components. During supervised fine-tuning (SFT), we introduce **Latent Force Intervention Encoder (LFIE)** to eliminate future-motion leakage from moving Gaussian blobs. In Fig. 2(a), the leaked trajectory drives the ball along a straight path, overriding its physically plausible motion down the stairs. LFIE directly maps force location, direction, and magnitude into a learnable latent condition without exposing temporal trajectories or future displacement, requiring the model to infer the resulting state changes.

Second, inspired by how children acquire intuitive physics through interaction and observation Piaget & Cook (1952); Gopnik et al. (2004), we introduce the **Physics Attribution Network (PAN)** to connect force interventions with observed state changes. Self-attention and cross-attention capture feature relationships but fail to reliably trace physical responses, while PCA remains entangled with appearance. PAN instead aligns force-conditioned latent responses with temporal differences to localize where and how strongly state changes occur, as shown in Fig. 2(b). This supervision makes the relation between intervention and response explicit, encouraging the model to learn how forces drive scene evolution rather than reproduce generic motion patterns.

During reinforcement learning (RL), we introduce **Physics-Aware Reinforcement Learning (PARL)** to improve physical consistency and generalization to unseen scenes, as shown in Fig. 2(c). PARL derives an attribution-guided physics reward from the learned dynamics prior of a self-supervised V-JEPA encoder and predictor Mur-Labadia et al. (2026); Assran et al. (2025), reinforcing physically coherent state evolution within PAN-localized response regions. Since the reward can be hacked by generating static videos or visual artifacts, we introduce a Dynamic Gate that evaluates motion, temporal stability, and visual quality to filter such degenerate rollouts.

Extensive evaluations on in-distribution (ID) and out-of-distribution (OOD) benchmarks show that PhysForce improves force controllability, physical consistency, and generalization while preserving visual quality and temporal coherence. These results establish PhysForce as an effective framework for physics-aware interaction and state prediction in video world models.

Overall, our contributions are summarized as follows:

- We introduce the **Latent Force Intervention Encoder**, which encodes force location, direction, and magnitude in the video latent space without trajectory leakage, allowing the model to predict state changes caused by the applied force.
- We propose the **Physics Attribution Network**, which learns a force conditioned response map that localizes state changes associated with the intervention.
- We develop **Physics-Aware Reinforcement Learning** to improve physical consistency and generalization using attribution guided self-supervised rewards from videos, with a Dynamic Gate mitigating reward hacking.
- PhysForce achieves state-of-the-art force controllability, physical consistency, and generalization while preserving video quality and temporal coherence.

2 Related Works

Video Diffusion Transformer. Diffusion models have achieved remarkable success in generating high-quality videos from text Blattmann et al. (2023); Singer et al. (2023); Khachatryan et al. (2023); Ma et al. (2024). Leveraging the scalability of the Diffusion Transformer (DiT) architecture Peebles & Xie (2023), recent models Yang et al. (2024); Wan et al. (2025) have substantially improved both visual fidelity and motion quality. The rapid proliferation of open-source models has further accelerated progress in this field, enabling increasingly realistic video creation. Building upon these advances, controllable video generation (Ma et al., 2025; Xu et al., 2026; Zhang et al., 2025) enables generated videos to adhere to specific control conditions, thereby laying a solid foundation for action-conditioned video world models.

Physics-aware Video Generation. Despite the impressive progress of existing video generation models in producing cinematic and visually appealing videos, their ability to faithfully follow real-world physical consistency remains limited. Physics-aware video generation aims to enhance the capability of foundation video generation models to perceive, represent, and generate videos that conform to physical laws. Existing methods can be broadly categorized into several lines. Some approaches Zhang et al. (2026a); Gillman et al. (2026a;b); Liu et al. (2024); Zhang et al. (2024) introduce implicit physical modeling by converting physical correctness into computable losses or rewards. Other works Zhang et al. (2026b); Wang et al. (2026d); Yuan et al. (2026a) align physical representations with model features through external networks or physics priors. Another line of research Wang et al. (2026a); Liu et al. (2026b); Foo et al. (2026) reconstructs particle fields and combines them with physics solvers, using predicted future particle trajectories to guide video generation. In addition, Yuan et al. (2025) explicitly models dynamics and supervises model learning with Newtonian laws of motion.

Interactive World Model. An interactive world model aims to predict future state transitions based on user-specified initial conditions. Recent works (Sun et al., 2025; Yuan et al., 2026b; Ye et al., 2025; He et al., 2025; Xiang et al., 2025; Shin et al., 2026; Gu et al., 2026; Li et al., 2025) further incorporate control signals, such as camera parameters and trajectories, enabling video models to interactively explore and predict real-world dynamics. However, these control signals typically contain little explicit physical information. In contrast, another line of work (Gillman et al., 2026a;b; Wang et al., 2026b; Narayanan et al., 2026; Li et al., 2026) introduces physical quantities, such as forces, as initial triggering conditions, with the goal of improving models' understanding of physical interactions and eliciting physically consistent responses to such conditions.

3 Method

In this section, we present PhysForce, a two-stage framework for physics-aware force control in interactive video world models. The SFT stage learns latent force conditioning and intervention-response attribution, while the reinforcement learning stage improves physical consistency in unseen scenarios. The overall pipeline is shown in Fig. 3. Sec. 3.2.1 presents the **Latent Force Intervention Encoder (LFIE)**, which maps outcome-agnostic force conditions into the video latent space. Sec. 3.2.2 introduces the **Physics Attribution Network (PAN)**, which associates force interventions with localized latent-state changes. Finally, Sec. 3.3 describes **Physics-aware Reinforcement Learning (PARL)**, which combines attribution-guided self-supervised rewards with a Dynamic Gate to improve unseen-state prediction and mitigate reward hacking.

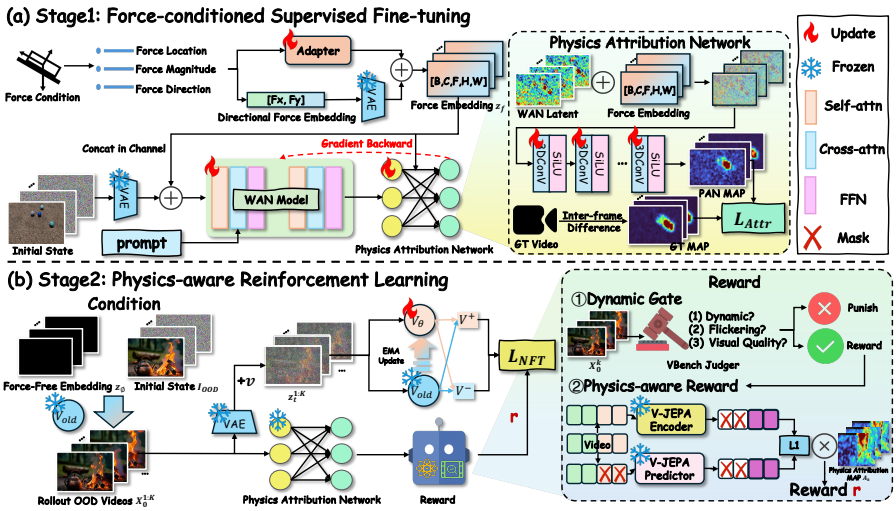

Figure 3: **Overview of PhysForce.** Stage 1 performs force-conditioned supervised fine-tuning with LFIE and PAN, while Stage 2 applies PARL with a self-supervised reward and Dynamic Gate.

3.1 Preliminary

Online Diffusion Reinforcement Learning. Existing approaches Liu et al. (2026a) formulate reverse denoising as a sequential decision-making process and optimize diffusion models using policy-gradient methods. In contrast, DiffusionNFT Zheng et al. (2025b) performs reinforcement learning directly on the forward process by contrasting generated samples based on their rewards.

Given a generated sample x_0, its condition c, and a normalized reward $r(X_0, c) \in [0, 1]$, DiffusionNFT constructs implicit positive and negative velocity predictors:

$$\begin{aligned} v_\theta^+(z_t, c, t) &= (1-\beta)v_{\text{old}}(z_t, c, t) + \beta v_\theta(z_t, c, t), \\ v_\theta^-(z_t, c, t) &= (1+\beta)v_{\text{old}}(z_t, c, t) - \beta v_\theta(z_t, c, t), \end{aligned} \tag{1}$$

where v_{old} denotes the data-collection policy and β controls the reinforcement strength. The model is then optimized using the reward-weighted flow matching objective

$$\mathcal{L}_{\text{NFT}} = \mathbb{E}_{X_0,c,t}\left[r(X_0,c)\left\|v_\theta^+ - v_t\right\|_2^2 + \left(1 - r(X_0,c)\right)\left\|v_\theta^- - v_t\right\|_2^2\right]. \tag{2}$$

where the arguments (z_t, c, t) of v_θ^+ and v_θ^- are omitted for brevity. This objective promotes high-reward samples while leveraging low-reward samples as negative feedback.

3.2 Stage1: Force-conditioned supervised fine-tuning

3.2.1 Latent Force Intervention Encoder

Trajectory-based force representations specify a sequence of positions and therefore expose the expected motion. To avoid this trajectory leakage, we introduce the LFIE, which represents only the applied force without prescribing its outcome. Given an application point $\mathbf{p}$, a normalized magnitude s, and direction $(\cos\theta, \sin\theta)$ in the image coordinate, we construct a four-channel force field:

$$\mathbf{F}(\mathbf{p}) = [M(\mathbf{p}), W(\mathbf{p})\cos\theta, W(\mathbf{p})\sin\theta, W(\mathbf{p})s], \tag{3}$$

where M is a truncated Gaussian centered at $\mathbf{p}_0$, and W denotes its binary support. The support has a fixed spatial extent, ensuring that force magnitude is represented by s rather than by the size of the force region. For global forces such as wind, we set $M(\mathbf{p}) = W(\mathbf{p}) = 1$. The same force field is repeated along the temporal dimension while its center remains fixed. Consequently, the condition contains no future position, displacement, or object trajectory.

LFIE projects the force condition into the video latent space through two parallel branches. The first branch constructs spatially localized directional components:

$$F_x(\mathbf{p}) = M(\mathbf{p}) \cdot s \cdot \cos\theta, \qquad F_y(\mathbf{p}) = M(\mathbf{p}) \cdot s \cdot \sin\theta, \tag{4}$$

A fixed mapping ϕ converts these components into the native control format of the pretrained model. The frozen video VAE then produces the Directional Force Embedding:

$$\mathbf{z}_{\mathrm{DFE}} = \mathcal{E}_{VAE}\left(\phi(F_x, F_y)\right), \tag{5}$$

This branch contains no trainable parameters and preserves compatibility with the pretrained control pathway. In parallel, a trainable adapter $\mathcal{A}_\psi$ maps the complete factorized force field directly onto the video latent grid. The two branches are combined by elementwise addition:

$$\mathbf{z}_f = \mathbf{z}_{\mathrm{DFE}} + \mathcal{A}_\psi(\mathbf{F}). \tag{6}$$

The output projection of $\mathcal{A}_\psi$ is initialized to zero, so LFIE initially follows the frozen directional embedding and progressively learns complementary representations of force location, direction, and magnitude. Finally, the resulting force embedding is concatenated with the visual conditioning latent along the channel dimension. Since neither branch receives the future motion, the model must infer the resulting state evolution from the interaction between the applied force and the current scene.

3.2.2 Physics Attribution Network

Standard video-generation objectives supervise the generated video globally without explicitly associating force interventions with their induced state changes. Inspired by Piaget's sensorimotor theory, which relates physical knowledge to actions and their observed consequences Piaget & Cook (1952), we introduce the PAN to strengthen intervention-response learning.

At timestep t, let $\hat{\mathbf{z}}_t$ and $\mathbf{z}_t^*$ denote the predicted and target latent sequences under the same noise realization, and the force embedding $\mathbf{z}_f$ from Sec. 3.2.1. Their temporal responses are defined as:

$$\Delta\hat{\mathbf{z}}_{t,i} = \hat{\mathbf{z}}_{t,i+1} - \hat{\mathbf{z}}_{t,i}, \qquad \Delta\mathbf{z}^*_{t,i} = \mathbf{z}^*_{t,i+1} - \mathbf{z}^*_{t,i}. \tag{7}$$

Here, i indexes temporal positions in the latent sequence. Since the simulated videos are captured by static cameras and their state changes are driven by the applied force, these temporal differences provide dense cues for localizing state changes caused by the applied force. PAN encodes the predicted response and detached force embedding through separate branches. The force feature modulates the response through a zero-initialized residual gate to produce the attribution feature $\hat{\mathbf{a}}_i$. PAN aligns it with the observed temporal response:

$$\hat{\mathbf{a}}_i = \mathrm{PAN}\left(\Delta\hat{\mathbf{z}}_{i,t}, \mathrm{sg}(\mathbf{z}_f)\right), \qquad \mathcal{L}_{\mathrm{attr}} = \frac{1}{T-1}\sum_{i=1}^{T-1}\left\|\hat{\mathbf{a}}_i - \Delta\mathbf{z}^*_{i,t}\right\|_1, \tag{8}$$

Here, $\mathrm{sg}(\cdot)$ denotes gradient detachment. This alignment captures where and how strongly states change without region annotations. The SFT objective is:

$$\mathcal{L}_{\mathrm{SFT}} = \mathcal{L}_{\mathrm{FM}} + \lambda_{\mathrm{attr}}\mathcal{L}_{\mathrm{attr}}. \tag{9}$$

where $\mathcal{L}_{\mathrm{FM}}$ denotes the loss of flow matching. By conditioning temporal response alignment on $\mathbf{z}_f$, PAN encourages the generator to associate force interventions with their induced state changes rather than rely only on correlations in the training distribution. We further include 3,000 force free videos from WISA-80K Wang et al. (2026c) and drop the force embedding for 10% of simulated samples to improve OOD robustness and support Stage 2. Further details are provided in Appendix A.2.

3.3 Physics-aware Reinforcement Learning

Supervised force interaction data cover only limited physical dynamics. We introduce PARL to improve physical consistency and generalization using auxiliary OOD videos. PARL combines an attribution-guided world model reward with a Dynamic Gate that filters degenerate rollouts.

Given an OOD initial state $\mathbf{I}_{\mathrm{ood}}$ and a force-free embedding $\mathbf{z}_\varnothing$, the old policy generates K clean rollout latents $\mathbf{z}_0^{1:K}$. These latents are decoded into videos $\mathbf{X}^{1:K}$ for reward evaluation and forward noised into $\mathbf{z}_t^{1:K}$ for DiffusionNFT optimization.

Attribution-aligned physics reward. Following Yuan et al. (2026a), we divide each decoded video $\mathbf{X}^k$ into C context frames and T future frames. For each window ℓ, the frozen V-JEPA predictor P_ϕ estimates the future features from the context, while its encoder E_ψ extracts the corresponding target features from the complete window:

$$\hat{\mathbf{H}}^{k,\mathrm{fut}}_\ell = P_\phi\big(E_\psi(\mathbf{X}^k_{\ell-C+1:\ell}), \Delta_t\big), \qquad \mathbf{H}^{k,\mathrm{fut}}_\ell = \big[E_\psi(\mathbf{X}^k_{\ell-C+1:\ell+T})\big]_{\Delta_t}, \tag{10}$$

where Δ_t specifies the future-token positions. The predicted and target features are arranged in their original spatiotemporal token layouts for subsequent attribution weighting.

Following Sec. 3.2.2, PAN produces a signed latent response $\hat{\mathbf{a}}^k \in \mathbb{R}^{B\times C_a\times T_a\times H_a\times W_a}$ from each clean rollout. We convert it into a nonnegative attribution map by taking the channelwise L_2 norm:

$$\hat{\mathbf{a}}^k = \mathrm{PAN}\left(\Delta \mathbf{z}_0^k, \mathrm{sg}(\mathbf{z}_\varnothing)\right), \qquad A^k = \left(\sum_{c=1}^{C_a} \left|\hat{a}^k\right|^2\right)^{\frac{1}{2}}, \tag{11}$$

Thus, $\mathbf{A}^k \geq 0$ measures the response magnitude at each latent location. We resize and temporally pool the PAN attribution map $\mathbf{A}^k$ to match the spatial patches and temporal tubelets of the V-JEPA features. The resulting nonnegative weight map $\mathbf{W}^k$ assigns an attribution weight to each predicted future token. Thus, each V-JEPA token is assigned a nonnegative attribution weight.

The aligned map is broadcast over the feature dimension and applied to the V-JEPA prediction error:

$$\mathcal{L}_{\mathrm{attr}}^k = \frac{\left\|\mathbf{W}^k \odot \left(\hat{\mathbf{H}}_\ell^{k,\mathrm{fut}} - \mathbf{H}_\ell^{k,\mathrm{fut}}\right)\right\|_1}{D\left\|\mathbf{W}^k\right\|_1 + \epsilon}, \qquad R_{\mathrm{phy}}^k = -\mathcal{L}_{\mathrm{attr}}^k, \tag{12}$$

This formulation evaluates local predictability at the V-JEPA token level, assigning higher rewards to predictable state changes in PAN-identified regions.

Dynamic Gate. The physics reward may assign high scores to degenerate rollouts, including static videos and motion caused by visual artifacts. We therefore introduce a reward-independent Dynamic Gate implemented by a VBench Judger. For each rollout $\mathbf{X}^k$, it evaluates the dynamic ratio d^k, temporal stability q_{temp}^k, and visual quality q_{vis}^k:

$$m^k = \mathbb{I}\left[\tau_d^- \leq d^k \leq \tau_d^+ \wedge q_{\mathrm{temp}}^k \geq \tau_{\mathrm{temp}} \wedge q_{\mathrm{vis}}^k \geq \tau_{\mathrm{vis}}\right], \qquad \mathcal{N} = \left\{k \mid m^k = 1\right\}. \tag{13}$$

The thresholds are calibrated on held-out SFT rollouts and remain fixed during reinforcement learning. The additional tolerance ϵ_A excludes empty or near-zero attribution maps. Only rollouts in $\mathcal{N}$ participate in reward normalization and optimization. The gate therefore prevents static collapse, excessive motion, visual artifacts, and invalid attribution maps from exploiting the physics reward.

Forward-process reinforcement learning. Following DiffusionNFT Zheng et al. (2025b), we convert the physics rewards within the eligible set into relative optimality probabilities:

$$\bar{R}_c = \frac{1}{|\mathcal{N}|}\sum_{j\in\mathcal{N}} R_{\mathrm{phy}}^j, \qquad r(\mathbf{X}_0^k, c) = \frac{1}{2} + \frac{R_{\mathrm{phy}}^k - \bar{R}_c}{2\max_{j\in\mathcal{N}}\left|R_{\mathrm{phy}}^j - \bar{R}_c\right| + \epsilon}, \quad k \in \mathcal{N}. \tag{14}$$

Here, rewards above and below the group mean are mapped above and below $1/2$, respectively. Conditions with fewer than two eligible rollouts are skipped, and rejected rollouts are excluded from both normalization and optimization. Noise is added to each eligible latent $\mathbf{z}_0^k$ to obtain $\mathbf{z}_t^k$, which is optimized using Eq. 1 and Eq. 2. The data-collection predictor v_{old} is updated from v_θ through exponential moving average.

4 Experiment

4.1 Experimental Settings

Implementation Details. PhysForce builds on the DiT-based Wan2.2-Fun-5B-Control model Peebles & Xie (2023); Wan et al. (2025) and is trained on eight NVIDIA RTX PRO 6000 GPUs. Stage 1 jointly optimizes LFIE, PAN, and a rank-64 DiT LoRA for 4,600 steps at up to 832×480 resolution. Stage 2 applies DiffusionNFT Zheng et al. (2025b) for 55 updates using a rank-32 policy LoRA initialized from the frozen Stage 1 model. Each update uses two conditions and six 30-step rollouts per condition. Further details are provided in Appendix D.

Dataset. **Training data.** Stage 1 uses approximately 12K point-force videos from GoalForce Gillman et al. (2026b), 15,359 wind-force videos from ForcePrompting Gillman et al. (2026a), and 3K

Table 1: Quantitative comparison and ablation study on the Force Prompting out-of-distribution benchmark. Best and second-best results are highlighted independently.

Evaluation	Method	Force Alignment					Video Quality				
		DCS ↑	DA@30° ↑	DA@45° ↑	MRR ↑	RMM ↑	CLIP-T ↑	PickScore ↑	Motion Smoothness ↑	Aesthetic Quality ↑	Imaging Quality ↑
Qualitative Evaluation	Wan-Move	0.333	0.392	0.499	0.210	0.206	0.295	19.942	**0.997**	47.413	62.237
	Force Prompting	0.575	0.714	0.748	0.810	1.453	0.306	20.797	0.996	47.568	62.238
	GoalForce	0.392	0.490	0.560	0.389	0.469	0.301	20.046	0.996	47.498	62.059
	PhysForce (Ours)	**0.782**	**0.835**	**0.863**	**0.952**	**2.488**	**0.315**	**21.065**	0.996	**49.048**	**64.352**
Ablation Study Stage 1	S1: Wan2.2-5B Base	0.010	0.190	0.252	0.608	1.154	**0.316**	**21.144**	**0.997**	**48.803**	**62.833**
	S2: S1 w/ LFIE	0.614	0.672	0.768	**0.908**	1.987	0.308	20.830	0.995	47.320	62.420
	S3: S2 w/ Direct Δz Loss	0.606	0.697	0.768	0.880	2.032	0.310	20.849	0.995	47.432	62.440
	S4: S1 w/ PAN	0.017	0.204	0.286	0.653	1.437	0.314	20.970	0.996	47.601	62.715
	S5: S1 w/ LFIE+PAN (Ours)	**0.625**	**0.717**	**0.787**	**0.908**	**2.063**	0.309	20.857	0.995	47.503	62.406
Ablation Study Stage 2	R1: R4 w/o. OOD Data	0.622	0.686	0.773	0.751	0.892	0.310	20.945	**0.997**	48.203	61.045
	R2: R4 w/o. Attribution Align	0.592	0.695	0.754	0.832	1.540	0.310	20.937	0.996	48.246	63.248
	R3: R4 w/o. Dynamic Gate	0.621	0.717	0.779	0.933	2.205	0.308	20.804	0.995	47.514	63.265
	R4: PARL (Ours)	**0.782**	**0.835**	**0.863**	**0.952**	**2.488**	**0.315**	**21.065**	0.996	**49.048**	**64.352**

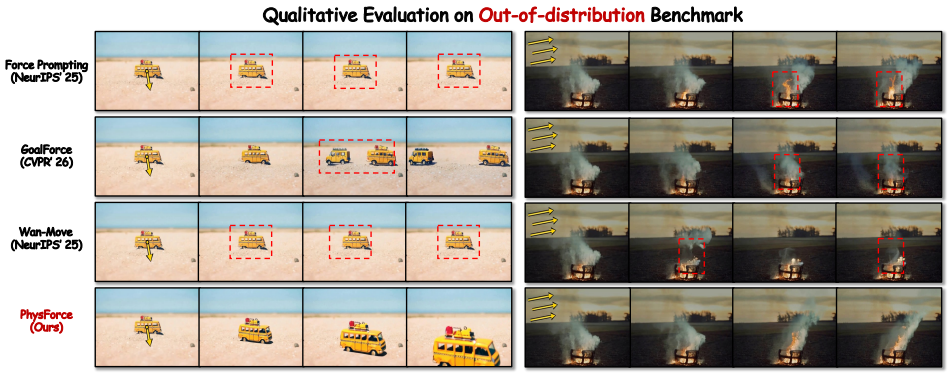

Figure 4: **Qualitative comparison of OOD Benchmark.**

force-free videos from WISA-80K Wang et al. (2026c). Stage 2 uses 100 static-camera WISA-80K videos without force annotations. **Evaluation benchmarks.** For OOD evaluation, we use the official ForcePrompting test set Gillman et al. (2026a), containing 305 point-force and 52 wind-force cases disjoint from training. For ID evaluation, GT80 contains 20 held-out cases each for balls, dominoes, plants, and wind. ForcePrompting evaluates force alignment and video quality, while GT80 additionally evaluates ground-truth physics alignment. Details are provided in Appendix F.

Metrics. Following prior work Wang et al. (2026b); Gillman et al. (2026a;b), we evaluate ID and OOD performance using 14 metrics across three dimensions. **(1) Force Alignment.** *Directional Cosine Similarity (DCS)* measures the directional agreement between the applied force and generated motion. *DA@30°* and *DA@45°* report the percentages of samples with directional errors below 30° and 45°, respectively. *Motion Response Rate (MRR)* measures the frequency of valid motion responses, while *ROI Motion Magnitude (RMM)* quantifies motion strength within the affected region. **(2) Video Quality.** *CLIP-T* and *PickScore* Kirstain et al. (2023) evaluate text alignment and learned human preference, respectively. We further adopt *Motion Smoothness*, *Aesthetic Quality*, and *Imaging Quality* from VBench Huang et al. (2024); Zheng et al. (2025a) to assess temporal consistency, visual appeal, and image fidelity. **(3) Physics Alignment.** For cases with paired ground-truth videos, we follow the Physics-IQ evaluation protocol Motamed et al. (2026). *Spatial IoU*, *Spatiotemporal IoU (ST-IoU)*, and *Weighted IoU* measure agreement between the generated and ground-truth motion regions, while *MSE* measures their pixel-level discrepancy.

Compared Methods. To evaluate the effectiveness of PhysForce, we compare it with State-of-the-Art (SOTA) approaches. We adapt **Wan-Move** Chu et al. (2026), **ForcePrompting** Gillman et al. (2026a) and GoalForce Gillman et al. (2026b) to the Wan2.2-5B base model for fair comparison.

4.2 Qualitative Evaluation

We compare PhysForce with Wan-Move Chu et al. (2026), ForcePrompting Gillman et al. (2026a), and GoalForce Gillman et al. (2026b) in force response, video quality, and physical consistency. **(1) OOD Comparison.** As shown in Fig. 4, the baselines struggle with challenging force conditions. For point force, ForcePrompting and Wan-Move produce nearly static videos, while GoalForce duplicates the vehicle. PhysForce moves the vehicle toward the foreground with coherent perspective

Table 2: Quantitative comparison and ablation study on the GT80 in-distribution benchmark. Best and second-best results are highlighted independently.

Evaluation	Method	Force Alignment					GT-based Physics Alignment			
		DCS ↑	DA@30° ↑	DA@45° ↑	MRR ↑	RMM ↑	Spatial IoU ↑	ST-IoU ↑	Weighted IoU ↑	MSE ↓
Qualitative Evaluation	Wan-Move	0.464	0.488	0.550	0.450	0.392	0.373	0.120	0.252	0.005
	Force Prompting	0.433	0.563	0.600	0.663	0.856	0.657	0.255	0.522	0.004
	GoalForce	0.433	0.550	0.575	0.725	0.873	0.525	0.186	0.404	0.006
	PhysForce (Ours)	**0.697**	**0.700**	**0.775**	**0.763**	**1.447**	**0.727**	**0.332**	**0.614**	**0.003**
Ablation Study Stage 1	S1: Wan2.2-5B Base	0.112	0.138	0.238	0.613	1.148	0.367	0.197	0.293	0.007
	S2: S1 w/ LFIE	0.509	0.575	0.663	0.755	1.206	0.610	0.250	0.495	0.005
	S3: S2 w/ Direct Δz Loss	0.509	0.550	0.663	0.743	1.144	0.616	0.250	0.499	0.005
	S4: S1 w/ PAN	0.118	0.263	0.325	0.663	0.712	0.367	0.125	0.260	0.007
	S5: S1 w/ LFIE+PAN (Ours)	**0.652**	**0.638**	**0.738**	**0.758**	**1.308**	**0.654**	**0.275**	**0.538**	**0.004**
Ablation Study Stage 2	R1: R4 w/o. OOD Data	0.409	0.550	0.638	0.438	0.467	0.488	0.161	0.365	0.004
	R2: R4 w/o. Attribution Align	0.417	0.575	0.613	0.688	1.245	0.650	0.283	0.545	0.004
	R3: R4 w/o. Dynamic Gate	0.536	0.600	0.650	0.750	1.360	0.634	0.273	0.519	0.005
	R4: PARL (Ours)	**0.697**	**0.700**	**0.775**	**0.763**	**1.447**	**0.727**	**0.332**	**0.614**	**0.003**

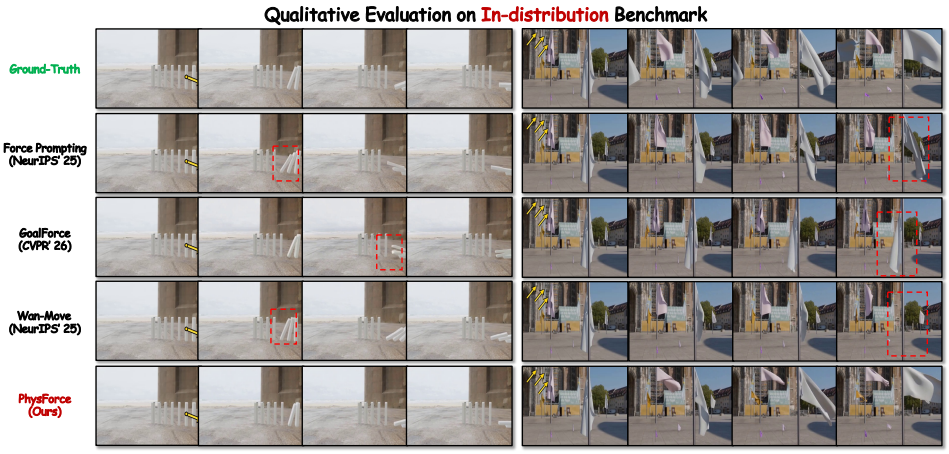

Figure 5: **Qualitative comparison of ID Benchmark.**

and scale changes, capturing the implicit 3D motion. For wind force, the baselines either separate the evolution of smoke and flame or fail to model the wind response. PhysForce preserves their interaction and generates motion aligned with the wind. **(2) ID Comparison.** As shown in Fig. 5, the baselines also struggle with force location and magnitude. In the domino example, ForcePrompting and Wan-Move respond at incorrect locations, while GoalForce generates an inconsistent fallen domino. PhysForce reproduces the ground-truth transition. In the wind example, ForcePrompting underestimates the response, GoalForce remains nearly static, and Wan-Move produces artifacts. PhysForce captures both the waving direction and the degree of flag extension.

4.3 Quantitative Evaluation

We compare PhysForce with representative force-controlled video generation methods, including Wan-Move Chu et al. (2026), Force Prompting Gillman et al. (2026a), and GoalForce Gillman et al. (2026b). The comparison covers force alignment, video quality, and ground-truth physics alignment on both OOD and ID benchmarks.

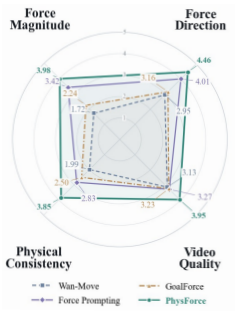

Figure 6: Quantitative User Study.

Objective Evaluation. As shown in Tab. 1 and 2, PhysForce outperforms existing methods across the evaluated dimensions. **(1) OOD Comparison.** On the OOD benchmark, PhysForce achieves the best results across all force alignment metrics. Compared with the best baseline, DCS improves from 0.575 to 0.782, MRR from 0.810 to 0.952, and RMM from 1.453 to 2.488. It also achieves the highest CLIP-T, PickScore, aesthetic quality, and imaging quality while maintaining comparable motion smoothness, demonstrating robust force response and video quality in unseen scenarios. **(2) ID Comparison.** On the GT80 benchmark, PhysForce achieves the best results across all force and physics alignment metrics. Compared with the strongest baselines, DCS improves from 0.464 to 0.697, RMM from 0.873 to 1.447, Spatial IoU from 0.657 to 0.727, and ST-IoU from 0.255 to 0.332,

while MSE decreases from 0.004 to 0.003. These results demonstrate state changes consistent with both the applied force and ground-truth physical evolution.

User Study. We recruited 150 participants to evaluate four methods on 40 scenes, including 20 point-force and 20 wind-force cases. Participants rated force magnitude response, direction adherence, video quality, and physical consistency on a five-point scale. Method names were hidden and presentation orders randomized. As shown in Fig. 6, PhysForce ranks first in all dimensions, exceeding the strongest baseline by 0.56, 0.45, 0.68, and 1.02 points, with the largest gain in physical consistency. More details in Appendix C.

4.4 Ablation Study

We evaluate PhysForce components on the ID and OOD benchmarks in Tables 2 and 1. Additional ablations on force magnitude and direction are provided in Appendix B.

Effect of LFIE in Stage 1. [S2] The base model frequently produces motion (MRR 0.608) but shows little force alignment (DCS 0.010). LFIE raises DCS to 0.614, DA45 to 0.768, and MRR to 0.908, while improving ID Spatial IoU from 0.367 to 0.610 and Weighted IoU from 0.293 to 0.495. These gains show that LFIE transforms generic motion into force-aligned responses without prescribing future displacement. The base model retains better OOD video quality because simulation-based fine-tuning introduces a domain shift, motivating real-video refinement in Stage 2.

Effect of PAN in Stage 1. [S3, S4, S5] (1) **S3** tests direct temporal-difference supervision without PAN. Despite increasing RMM, it yields marginal ID IoU gains and reduces OOD DCS from 0.614 to 0.606 and MRR from 0.908 to 0.880. Thus, direct difference matching may amplify motion without isolating force-consistent responses. (2) **S4** evaluates PAN without LFIE. Although MRR and RMM improve, OOD DCS remains near zero. PAN models the resulting state changes but requires LFIE to provide an effective intervention signal, confirming their complementary roles. (3) **S5** combines LFIE and PAN. With nearly unchanged MRR, ID DCS increases from 0.509 to 0.652 and Weighted IoU from 0.495 to 0.538. The gains reflect better aligned and propagated responses rather than more motion. Unlike direct difference matching, PAN learns a dedicated force-conditioned response representation that localizes how the intervention affects the scene, providing localized supervision in Stage 1 and reusable regional guidance for Stage 2 rewards.

Effect of OOD Data in Stage 2. [R1] Removing force-free OOD videos reduces OOD DCS from 0.782 to 0.622 and RMM from 2.488 to 0.892, while decreasing ID ST-IoU from 0.332 to 0.161 and Weighted IoU from 0.614 to 0.365. The degradation across both benchmarks shows that PARL benefits from exposing the shared generator to diverse real-world dynamics. The gains reflect the transfer of a refined dynamics prior rather than additional force supervision.

Effect of Attribution Alignment in Stage 2. [R2] R2 tests whether global V-JEPA prediction error alone is sufficient. Without attribution alignment, static backgrounds and unrelated regions can dominate the reward. PAN-guided alignment focuses the reward on relevant state changes, improving OOD DCS from 0.592 to 0.782 and MRR from 0.832 to 0.952. It also increases ID ST-IoU from 0.283 to 0.332 and Weighted IoU from 0.545 to 0.614. PAN therefore provides the regional bridge that converts the global V-JEPA signal into a response-aware physics reward.

Effect of the Dynamic Gate in Stage 2. [R3] Without the Dynamic Gate, static or degraded rollouts may exploit the predictability reward. The gate improves OOD DCS from 0.621 to 0.782, RMM from 2.205 to 2.488, and imaging quality from 63.265 to 64.352. It also raises ID Weighted IoU from 0.519 to 0.614 and reduces MSE from 0.005 to 0.003, demonstrating that it mitigates reward hacking while improving physical and visual consistency.

5 Conclusion

We present PhysForce, a two-stage framework for physics-aware force control in interactive video world models. LFIE encodes forces without exposing future motion, while PAN learns localized physical responses from temporal changes. PARL further refines the dynamics prior on force-free OOD videos using PAN-guided rewards and a Dynamic Gate. Experiments show improved force alignment, physical consistency, OOD generalization, and video quality. Future work will explore explicit 3D modeling and causal streaming generation for real-time interaction.

AI USE STATEMENT

We used generative AI tools, including OpenAI GPT-5.6-Sol and GPT-6 Astra, to assist with methodological and experimental refinement, mathematical presentation, code development and debugging, dataset processing, translation, manuscript editing, and scientific figure preparation.

All AI-assisted outputs were reviewed and revised by the authors. Generated data and evaluation conditions were manually inspected, AI-assisted code was tested for correctness, and all references, mathematical formulations, experimental results, and conclusions were independently verified by the authors. Generative AI was not treated as a source of empirical evidence. We take full responsibility for the final content of this work, including all text, claims, code, data, and artifacts produced with the assistance of generative AI.

ETHICS STATEMENT

This work studies force-controlled video generation and physical reasoning for research purposes. While it may benefit interactive world models and robotic data generation, realistic synthetic videos could be misused to fabricate physical events or misrepresent model capabilities. We therefore recommend clearly identifying generated content and avoiding deployment in safety-critical settings without human oversight and real-world validation. Our user study involved voluntary perceptual evaluation without collecting sensitive personal information. We will follow the licenses of all source datasets and document the intended use, limitations, and potential misuse risks in any public release.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of LFIE, PAN, PARL, rollout construction, training, inference, and evaluation in the main paper and appendix. We also report dataset splits, baseline adaptations, Dynamic Gate thresholds, checkpoint selection, and metric definitions. We will release the code, data-processing scripts, benchmark annotations, model configurations, and evaluation tools to support reproduction of our results.

REFERENCES

Mahmoud Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

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.

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 *Forty-first International Conference on Machine Learning*, 2024.

Taiye Chen, Xun Hu, Zihan Ding, and Chi Jin. Learning world models for interactive video generation. *Advances in Neural Information Processing Systems*, 38:154456–154483, 2026.

Ruihang Chu, Yefei He, Zhekai Chen, Shiwei Zhang, Xiaogang Xu, Dingdong WANG, Hongwei Yi, Xihui Liu, Hengshuang Zhao, Yu Liu, et al. Wan-move: Motion-controllable video generation via latent trajectory guidance. *Advances in Neural Information Processing Systems*, 38:404–432, 2026.

Lin Geng Foo, Mark He Huang, Alexandros Lattas, Stylianos Moschoglou, Thabo Beeler, and Christian Theobalt. Physical simulator in-the-loop video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4301–4311, 2026.

Nate Gillman, Charles Herrmann, Michael Freeman, Daksh Aggarwal, Evan Luo, Deqing Sun, and Chen Sun. Force prompting: Video generation models can learn and generalize physics-based control signals. *Advances in Neural Information Processing Systems*, 38:103174–103201, 2026a.

Nate Gillman, Yinghua Zhou, Zitian Tang, Evan Luo, Arjan Chakravarthy, Daksh Aggarwal, Michael Freeman, and Chen Sun. Goal force: Teaching video models to accomplish physics-conditioned goals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20077–20087, 2026b.

Alison Gopnik, Clark Glymour, David M Sobel, Laura E Schulz, Tamar Kushnir, and David Danks. A theory of causal learning in children: causal maps and bayes nets. *Psychological review*, 111 (1):3, 2004.

Bohai Gu, Taiyi Wu, Yueyang Yuan, Jian Liu, Xiaocheng Lu, Dazhao Du, Jie Zhang, Jinxiang Lai, Shuai Yang, Xiaotong Zhao, et al. Worldcraft: From camera navigation to object manipulation in interactive video world models. *arXiv preprint arXiv:2605.25077*, 2026.

Xianglong He, Chunli Peng, Zexiang Liu, Boyang Wang, Yifan Zhang, Qi Cui, Fei Kang, Biao Jiang, Mengyin An, Yangyang Ren, et al. Matrix-game 2.0: An open-source real-time and streaming interactive world model. *arXiv preprint arXiv:2508.13009*, 2025.

Siqiao Huang, Jialong Wu, Qixing Zhou, Shangchen Miao, and Mingsheng Long. Vid2world: Crafting video diffusion models to interactive world models. *arXiv preprint arXiv:2505.14357*, 2025.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818. IEEE, 2024.

Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15954–15964, 2023.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663, 2023.

Guangyuan Li, Bo Li, Jinwei Chen, Xiaobin Hu, Lei Zhao, and Peng-Tao Jiang. Magicworld: Towards long-horizon stability for interactive video world exploration. *arXiv preprint arXiv:2511.18886*, 2025.

Yanxun Li, Hao Wen, Bingze Song, Jiashu Zhu, Aiming Hao, Chubin Chen, Jintao Chen, Jiahong Wu, Xiangxiang Chu, and Miao Wang. Learning explicit physical parameter control and benchmarking for video generation. *arXiv preprint arXiv:2607.18924*, 2026.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *Advances in neural information processing systems*, 38:40783–40818, 2026a.

Shaowei Liu, Zhongzheng Ren, Saurabh Gupta, and Shenlong Wang. Physgen: Rigid-body physics-grounded image-to-video generation. In *European Conference on Computer Vision*, pp. 360–378. Springer, 2024.

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*, 2026b.

Xin Ma, Yaohui Wang, Xinyuan Chen, Gengyun Jia, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *arXiv preprint arXiv:2401.03048*, 2024.

Yue Ma, Kunyu Feng, Zhongyuan Hu, Xinyu Wang, Yucheng Wang, Mingzhe Zheng, Xuanhua He, Chenyang Zhu, Hongyu Liu, Yingqing He, et al. Controllable video generation: A survey. *arXiv preprint arXiv:2507.16869*, 2025.

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 (WACV)*, 2026.

Lorenzo Mur-Labadia, Matthew Muckley, Amir Bar, Mahmoud Assran, Koustuv Sinha, Michael Rabbat, Yann LeCun, Nicolas Ballas, and Adrien Bardes. V-jepa 2.1: Unlocking dense features in video self-supervised learning. *arXiv preprint arXiv:2603.14482*, 2026.

Sriram Narayanan, Ziyu Jiang, Srinivasa Narasimhan, and Manmohan Chandraker. Phyco: Learning controllable physical priors for generative motion. *arXiv preprint arXiv:2604.28169*, 2026.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Jean Piaget and Margaret Trans Cook. The origins of intelligence in children. 1952.

Joonghyuk Shin, Zhengqi Li, Richard Zhang, Jun-Yan Zhu, Jaesik Park, Eli Shechtman, and Xun Huang. Motionstream: Real-time video generation with interactive motion controls. In *International Conference on Learning Representations*, volume 2026, pp. 7381–7406, 2026.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. In *The Eleventh International Conference on Learning Representations*, 2023.

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

Wenqiang Sun, Haiyu Zhang, Haoyuan Wang, Junta Wu, Zehan Wang, Zhenwei Wang, Yunhong Wang, Jun Zhang, Tengfei Wang, and Chunchao Guo. Worldplay: Towards long-term geometric consistency for real-time interactive world modeling. *arXiv preprint arXiv:2512.14614*, 2025.

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. *Advances in Neural Information Processing Systems*, 38:167907–167932, 2026a.

Hanhui Wang, Yiming Xie, Haiwen Feng, Zhaoyang Lv, Shenlong Wang, and Huaizu Jiang. Streaming video generation with streaming force control. *arXiv preprint arXiv:2606.07508*, 2026b.

Jing Wang, Ao Ma, Ke Cao, Jun Zheng, Jiasong Feng, Zhanjie Zhang, Wanyuan Pang, and Xiaodan Liang. Wisa: World simulator assistant for physics-aware text-to-video generation. *Advances in Neural Information Processing Systems*, 38:5388–5416, 2026c.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14492–14501, 2026d.

Jiannan Xiang, Yi Gu, Zihan Liu, Zeyu Feng, Qiyue Gao, Yiyan Hu, Benhao Huang, Guangyi Liu, Yichi Yang, Kun Zhou, et al. Pan: A world model for general, interactable, and long-horizon world simulation. *arXiv preprint arXiv:2511.09057*, 2025.

Xuancheng Xu, Yaning Li, Sisi You, and Bing-Kun Bao. Smrabooth: Subject and motion representation alignment for customized video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16130–16141, 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.

Deheng Ye, Fangyun Zhou, Jiacheng Lv, Jianqi Ma, Jun Zhang, Junyan Lv, Junyou Li, Minwen Deng, Mingyu Yang, Qiang Fu, et al. Yan: Foundational interactive video generation. *arXiv preprint arXiv:2508.08601*, 2025.

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

Shenghai Yuan, Yuanyang Yin, Zongjian Li, Xinwei Huang, Xiao Yang, and Li Yuan. Helios: Real real-time long video generation model. *arXiv preprint arXiv:2603.04379*, 2026b.

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. *arXiv preprint arXiv:2509.21309*, 2025.

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*, 2026a.

Shiyi Zhang, Junhao Zhuang, Zhaoyang Zhang, Ying Shan, and Yansong Tang. Flexiact: Towards flexible action control in heterogeneous scenarios. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–11, 2025.

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 *European Conference on Computer Vision*, pp. 388–406. Springer, 2024.

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. *Advances in Neural Information Processing Systems*, 38:122647–122676, 2026b.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, et al. Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025a.

Kaiwen Zheng, Huayu Chen, Haotian Ye, Haoxiang Wang, Qinsheng Zhang, Kai Jiang, Hang Su, Stefano Ermon, Jun Zhu, and Ming-Yu Liu. Diffusionnft: Online diffusion reinforcement with forward process. *arXiv preprint arXiv:2509.16117*, 2025b.

APPENDIX

A Discussion **S1**

A.1 Extension to Multi-Force Control S1

A.2 Why Temporal Differences Enhance Physics Attribution? S1

A.3 Why V-JEPA Improves Physics Consistency? S2

A.4 How Force-Free PARL Improves Force-Conditioned Generation? S3

B Detailed ablation illustration **S3**

B.1 Ablation Settings S3

B.2 More ablation studies S4

C Detailed User Study Instructions **S5**

D Additional Implementation Details **S6**

E Detailed Metric Definitions **S8**

F Detailed Dataset Description **S9**

G Limitations and Future Work **S10**

A DISCUSSION

A.1 EXTENSION TO MULTI-FORCE CONTROL

As illustrated in Fig. 1, PhysForce naturally extends to multiple simultaneous forces. In the first example, two opposing forces are applied to separate ornaments on a Christmas tree. Each ornament moves independently in the specified direction without interfering with the other. In the second example, two opposing forces of different magnitudes are applied to the ends of a domino chain. Both sides initially fall toward the center; after their effects meet, the stronger force determines the subsequent propagation direction and eventually overturns the remaining dominoes. These examples demonstrate that PhysForce can combine multiple force conditions and generate plausible interactions between their induced responses.

A.2 WHY TEMPORAL DIFFERENCES ENHANCE PHYSICS ATTRIBUTION?

PAN is primarily trained on controlled force-interaction videos. Because these videos use static cameras and the applied force is the dominant source of state change, temporal differences suppress static appearance and reveal where and how strongly the scene evolves. They provide dense supervision for the complete response, including initial motion, deformation, collisions, and subsequent motion propagation. Conditioned on $\mathbf{z}_f$, PAN transforms these differences into a localized response representation associated with the applied force.

The ablations in Tables 2 and 1 show that this benefit cannot be reproduced by temporal differences alone. Compared with LFIE (S2), directly applying a Δz loss (S3) leaves ID DCS unchanged at 0.509 and yields only marginal gains in Spatial IoU from 0.610 to 0.616 and Weighted IoU from 0.495 to 0.499. It also reduces OOD DCS from 0.614 to 0.606 and MRR from 0.908 to 0.880. In contrast, PAN with LFIE (S5) increases ID DCS to 0.652, Spatial IoU to 0.654, and Weighted IoU to 0.538, while achieving an OOD DCS of 0.625 and DA@30 of 0.717. Meanwhile, PAN without LFIE (S4) achieves only 0.118 ID DCS and 0.017 OOD DCS, confirming that PAN complements rather than replaces force conditioning.

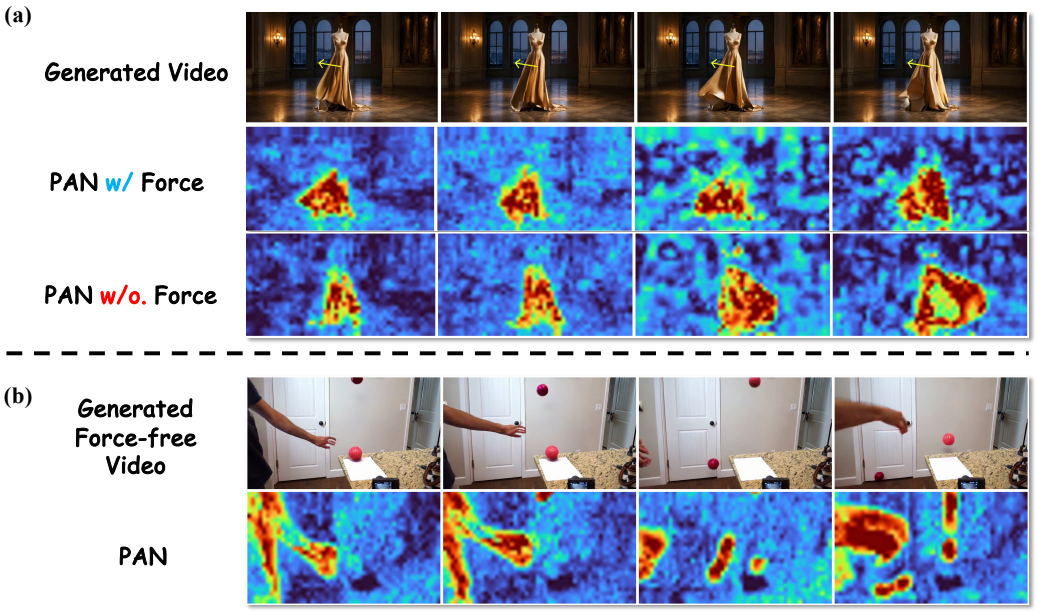

Figure 1: **Visualization of PAN** (a) Force-conditioned response localization. We keep the generated video and LFIE unchanged and remove only the force embedding provided to PAN. With the force condition, PAN focuses on state changes associated with the intervention; removing it substantially alters the response map, showing that PAN uses the force information for localization. (b) Force-free response localization. Given a completely force-free generated video, PAN identifies regions undergoing temporal state changes, supporting region-aware reward evaluation during PARL.

Figure 1(a) further isolates the role of force information within PAN. We keep the generated video and LFIE fixed and remove only the force embedding provided to PAN. The resulting change in the response map shows that PAN uses the intervention condition rather than acting as an unconditional temporal-difference filter.

We additionally include force-free OOD videos and drop the force condition for 10% of simulated samples during Stage 1. This extends the learned localization prior to scenes without explicit interventions. Conditioned on $\mathbf{z}_\varnothing$, PAN identifies physically relevant state changes in force-free videos, as shown in Fig. 1(b), and Stage 2 uses these regions to weight the V-JEPA prediction error. The Stage 2 ablations provide further evidence: compared with removing attribution alignment (R2), PAN-guided alignment (R4) improves OOD DCS from 0.592 to 0.782 and MRR from 0.832 to 0.952. On GT80, it increases ST-IoU from 0.283 to 0.332 and Weighted IoU from 0.545 to 0.614. These results support transferring the response-localization prior learned from supervised force interactions to force-free physical refinement.

A.3 Why V-JEPA Improves Physics Consistency?

V-JEPA prediction error should not be interpreted as a generic measure of temporal smoothness. As a latent world model, V-JEPA predicts masked future representations from preceding observations. Learning in representation space reduces sensitivity to appearance details and encourages its features to capture object dynamics, structural continuity, interactions, and state transitions. Consequently, even visually smooth motion can receive a low reward when its future representation deviates from the dynamics learned from real videos. Prior work further shows that V-JEPA-based rewards consistently improve PhysicsIQ scores and human-perceived physical plausibility across different video generators, outperforming pixel-reconstruction and VLM-based reward signals Assran et al. (2025); Yuan et al. (2026a).

PhysForce transfers this learned dynamics prior through a region-aware reward. PAN restricts the V-JEPA prediction error to regions undergoing relevant state changes, preventing static backgrounds

from dominating the reward, while the Dynamic Gate removes static, excessively dynamic, and visually degraded shortcuts. Our ablations further show that both attribution alignment and the Dynamic Gate improve ground-truth physics-alignment metrics. Therefore, PARL uses V-JEPA not merely to favor predictable videos, but to reinforce state evolution consistent with a learned physical dynamics prior. Although this reward is not an analytical physics solver, it provides an effective and empirically validated signal for improving physical consistency.

A.4 How Force-Free PARL Improves Force-Conditioned Generation?

PARL does not receive explicit force conditions or optimize a reward based on force-value regression. Our premise is that Stage 1 has already learned the basic relationship between force conditions and object responses. Rather than relearning this mapping, Stage 2 uses a physics-consistency reward on force-free videos to refine the generator's dynamics prior, temporal stability, and generalization to out-of-distribution scenes.

Although Stage 2 uses the force-free embedding, both stages share the same generator parameters. Updates from PARL can therefore transfer to force-conditioned generation. Specifically, PARL evaluates the spatiotemporal consistency of object motion through V-JEPA prediction errors within PAN-localized regions, discouraging discontinuous state transitions, structural degradation, and static collapse. When force conditions are reintroduced, the model retains the condition-response mapping learned in Stage 1 while benefiting from a more stable dynamics prior for propagating the induced response.

Our qualitative results reveal two main improvements. First, once motion is initiated, the Stage 2 model is less likely to exhibit abrupt directional changes, interrupted motion, or structural collapse. Second, in unseen scenes, it better preserves motion patterns consistent with the depicted objects and is less likely to generate static videos or ignore the initial dynamic state. These observations suggest that PARL primarily improves the stability of state propagation following a force response, rather than implicitly regressing an absent external force.

The quantitative results are consistent with this interpretation. From the complete Stage 1 model to Stage 2, OOD DCS increases from 0.625 to 0.782, while ID ST-IoU improves from 0.275 to 0.332, indicating better directional force alignment in unseen scenes and closer spatiotemporal agreement with ground-truth motion regions. Meanwhile, ID MRR increases slightly from 0.758 to 0.763, suggesting that the improvement in ID motion-region alignment is accompanied by only a small change in motion-response frequency. Taken together, these results support a transfer benefit of physics-consistency post-training to force-conditioned generation, reflected in improved spatiotemporal motion alignment and stronger OOD generalization, rather than providing direct evidence that PARL improves force-magnitude estimation.

B Detailed ablation illustration

B.1 Ablation Settings

We detail the configurations of the ablations reported in Tables 2 and 1. Unless otherwise specified, all variants use the same backbone, training data, and optimization settings.

Stage 1 ablations. S1 is the Wan2.2-5B base model without LFIE or PAN. **S2** adds LFIE to evaluate the effect of force conditioning without response attribution. **S3** removes PAN and directly applies an ℓ_1 loss between the predicted and ground-truth latent temporal differences, testing whether direct difference matching can replace the learned response representation. **S4** retains PAN but removes LFIE, examining response supervision without an effective force-conditioning interface. **S5** combines LFIE and PAN as the complete Stage 1 model.

Stage 2 ablations. All Stage 2 variants are initialized from S5. **R1** removes the additional force-free OOD videos while retaining the PARL objective. **R2** removes PAN-guided attribution alignment and globally aggregates the V-JEPA prediction error. **R3** disables the Dynamic Gate, allowing all sampled rollouts to participate in reward normalization and optimization. **R4** is the complete PARL model, combining force-free OOD data, PAN-guided regional rewards, and the Dynamic Gate.

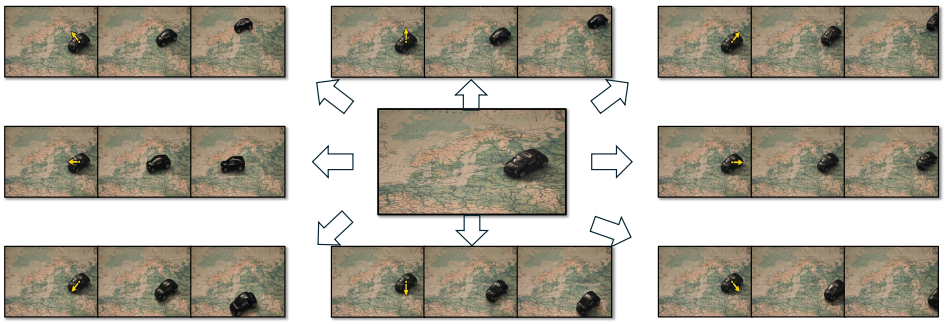

Figure 2: **Ablation across force directions.** We apply forces in eight directions to the same object and examine the resulting state changes. PhysForce consistently generates motion aligned with each applied direction, demonstrating accurate directional control.



Figure 3: **Ablation across force magnitudes.** We apply forces of varying magnitudes to the same object under identical conditions. Stronger forces produce progressively larger motion responses, showing that PhysForce captures the relationship between force magnitude and state change.

B.2 More ablation studies

We further examine how PhysForce responds to changes in force direction and magnitude. In each experiment, we vary only one force factor while keeping the initial scene, application point, and remaining conditions fixed.

Force Direction. As shown in Fig. 2, we apply forces in eight directions to the same object with a fixed magnitude. The generated motion consistently follows the specified direction while preserving the scene content. This demonstrates that PhysForce accurately distinguishes force directions and translates them into corresponding state changes.

Force Magnitude. As shown in Fig. 3, we fix the force direction and vary its magnitude. The resulting motion becomes progressively stronger as the applied force increases, consistent with expected physical behavior. This indicates that PhysForce models force magnitude as a continuous control signal rather than a binary motion trigger.

Together, these results show that PhysForce captures both the direction and magnitude of a force without relying on prescribed trajectories, enabling precise and physically consistent control.

Role of Force Conditioning in PAN. As shown in Fig. 1(a), we keep the generated video, latent temporal differences, generator, and LFIE unchanged, and remove only the force embedding provided to PAN. With the correct force condition, PAN concentrates on localized state changes associated with the applied intervention. Replacing it with the null embedding produces a substantially different response map that captures broader object motion. Because the visual input and temporal differences are identical in both settings, this difference can only arise from PAN's force-conditioning branch.

The experiment therefore shows that PAN uses the intervention information rather than functioning merely as an unconditional temporal-difference filter.

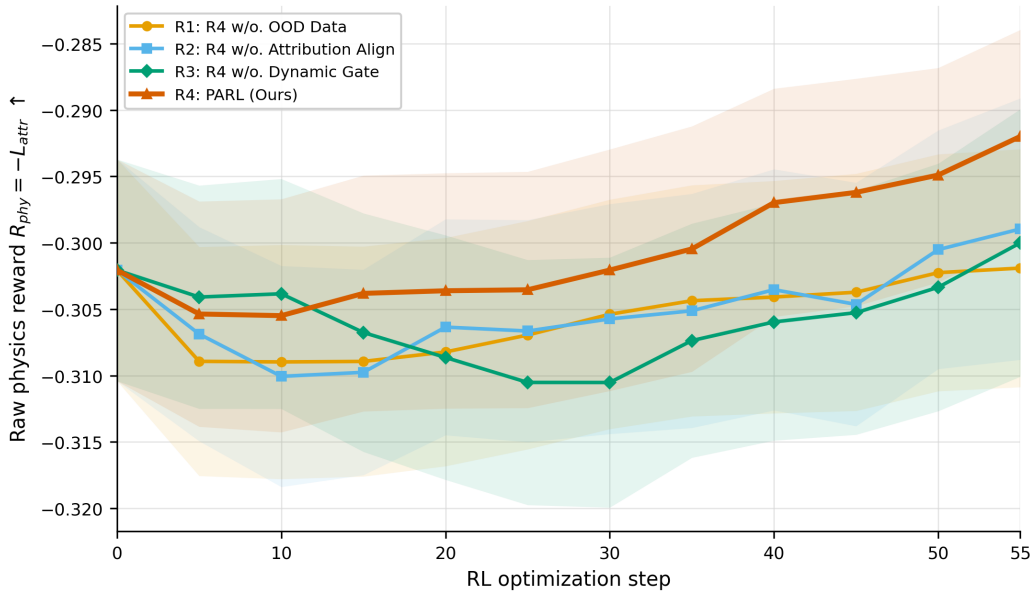

Figure 4: We visualize the reward trajectories of PARL and its ablated variants. The full model improves steadily and reaches its peak at step 55, consistently outperforming all ablations throughout optimization.

PAN is also trained on force-free videos and simulated samples with force dropout. Consequently, when no intervention is available, it operates in a complementary force-free mode. As shown in Fig. 1(b), PAN conditioned on $\mathbf{z}_\varnothing$ identifies naturally evolving regions in a completely force-free video, providing localized guidance for PARL. Thus, PAN represents force-associated responses when $\mathbf{z}_f$ is available and general physical state changes under the null condition. This evidence supports PAN as a force-conditioned response representation, although it should not be interpreted as formal counterfactual causal identification.

Ablation of PARL Components. As shown in Fig. 4, full PARL recovers from an initial decline and steadily improves its raw physics reward from -0.302 to -0.292, reaching the highest value at update 55. Removing OOD data, attribution alignment, or the Dynamic Gate consistently yields lower final rewards. In particular, removing the Dynamic Gate causes the largest intermediate degradation and a slower recovery. During training, the Dynamic Gate retains 65.3% of rollout groups on average. These results demonstrate that OOD data, attribution-guided alignment, and rollout filtering jointly improve the effectiveness and stability of PARL optimization.

C Detailed User Study Instructions

We conduct a user study to compare videos generated by Wan-Move, ForcePrompting, GoalForce, and PhysForce. Each participant evaluates 40 scenes, including 20 point force and 20 wind force scenarios. For each scene, the initial state, force condition, and four generated videos are presented together. A single arrow represents a point force, while three arrows represent wind; the arrow direction and length indicate the force direction and magnitude, respectively. Method identities are hidden and presentation orders are randomized to reduce bias. The questionnaire format is shown in Fig. 5. Participants rate each video from 1 to 5 according to four criteria: **(1) Force Magnitude Response.** Whether the intensity of the generated motion appropriately reflects the magnitude of the applied force. **(2) Force Direction Adherence.** Whether the generated motion follows the specified force direction. **(3) Video Quality.** Whether the video maintains visual fidelity, temporal

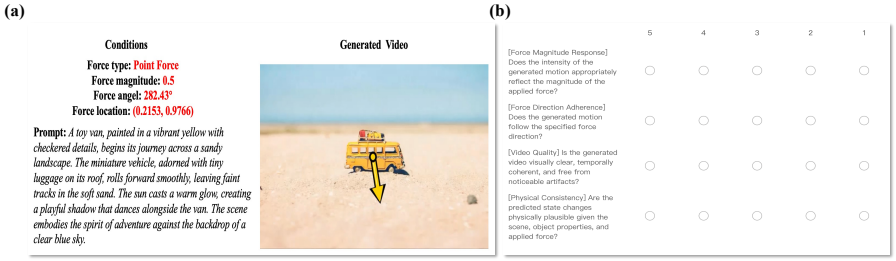

Figure 5: **Human evaluation questionnaire format.** (a) presents the scene prompt, force type, magnitude, direction, location, and generated video. (b) asks participants to rate the video from 1 to 5 in terms of force magnitude response, force direction adherence, video quality, and physical consistency.

coherence, and freedom from noticeable artifacts. **(4) Physical Consistency.** Whether the predicted state changes are physically plausible given the scene, object properties, and applied force.

D Additional Implementation Details

PARL Training and Checkpoint Selection. We select WISA-80K videos using predefined criteria that require a static camera, observable object motion, and no severe visual artifacts. The ID and OOD test benchmarks are not used for video selection, Dynamic Gate calibration, hyperparameter tuning, or checkpoint selection.

PARL is trained for 100 updates, with checkpoints evaluated every five updates on the fixed validation split. We select the checkpoint at update 55 because it achieves the highest mean validation physics-consistency reward, after which the reward plateaus without further improvement. This checkpoint is then evaluated once on the ID and OOD test benchmarks. Figure 4 reports the mean reward, with the shaded regions indicating standard deviation/95% confidence intervals. All variants use the same data split, initialization, rollout seeds, and checkpoint-selection protocol.

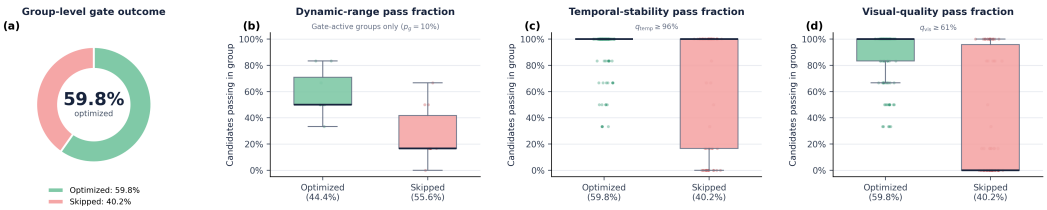

Figure 6: **Analysis of the Dynamic Gate.** Up to the selected step-55 checkpoint, the gate retains 59.8% of rollout groups for optimization and skips 40.2%. The remaining panels compare the dynamic-range pass fraction for gate-active groups and the temporal-stability and visual-quality pass fractions for all groups, showing that the gate effectively excludes degenerate rollouts.

Dynamic Gate. We analyze the Dynamic Gate in Fig. 6. For every rollout, the VBench Judger measures temporal stability and visual quality. It additionally computes a dynamic ratio by normalizing the rollout motion intensity against that of the frozen Stage-II policy under the same condition. A rollout is eligible if its temporal stability exceeds 96%, its visual quality exceeds 61%, and, when the dynamic criterion is active, its dynamic ratio lies within $[60\%, 150\%]$. A six-candidate group is optimized only if at least two rollouts are eligible; otherwise, it is skipped.

All thresholds are calibrated on held-out SFT-policy rollouts and remain fixed throughout PARL. Up to the selected step-55 checkpoint, 110 of 184 attempted groups are optimized, corresponding to a retention rate of 59.8%. As shown in Fig. 6, skipped groups exhibit substantially lower pass fractions under the active validity criteria. The gate neither rescales the physics reward nor operates during inference; it only prevents invalid rollout groups from contributing gradients, thereby mitigating reward hacking caused by static, excessively dynamic, flickering, or visually degraded samples.

Compared Methods. We compare PhysForce with Wan-Move Chu et al. (2026), ForcePrompting Gillman et al. (2026a), and GoalForce Gillman et al. (2026b). For a controlled comparison, all methods are implemented on the same Wan2.2-5B backbone and trained using identical data splits. Each method receives the same initial image, text prompt, and force annotation, including the application point, direction, and magnitude. The force annotation is converted into the conditioning representation required by each method following its original design.

We preserve the conditioning modules and objectives of each baseline while keeping the remaining settings identical. All methods are trained at 832×480 resolution for 4,600 steps using the same sampling strategy, a learning rate of 1×10^{-4}, and a force-condition dropout probability of 10%. LFIE, PAN, and PARL are used only in PhysForce.

During inference, all methods generate videos at 832×480 resolution using 50 denoising steps and a classifier-free guidance scale of 5.0.

Stage 1: Force-conditioned supervised fine-tuning. In Stage 1, we introduce the Latent Force Intervention Encoder (LFIE) and Physics Attribution Network (PAN). LFIE derives the intervention magnitude from an acceleration-related proxy. According to Newton's second law, $a = F/m$, and the initial velocity change over a fixed interval is proportional to acceleration. We therefore define the raw intervention strength as

$$\tilde{s} = \begin{cases} F/m, & \text{for balls with mass annotations,} \\ F, & \text{for dominoes and plants,} \\ F_{\text{wind}}, & \text{for wind.} \end{cases} \tag{15}$$

For balls, F/m directly represents the applied acceleration. For dominoes and plants, object mass is not exposed by the simulator, so we use its native force parameter, with the resulting response jointly determined by the simulated physical properties. Similarly, F_{wind} denotes the native wind-control strength.

Because these quantities have different physical meanings and units, we normalize them independently within each dataset subset $\mathcal{D}$:

$$s = \text{clip}\left(\frac{\tilde{s} - \tilde{s}^{\mathcal{D}}_{\min}}{\tilde{s}^{\mathcal{D}}_{\max} - \tilde{s}^{\mathcal{D}}_{\min} + \epsilon}, 0, 1 \right). \tag{16}$$

The normalization statistics are computed only from the corresponding training split. Therefore, s is a dimensionless intervention strength that preserves relative magnitudes within each physical domain, rather than a shared force unit across different categories.

After converting the source direction into image coordinates, we denote its unit vector as $\mathbf{d} = (d_x, d_y) = (\cos\theta, \sin\theta)$. Given an application point $\mathbf{p}$, LFIE constructs the factorized force field

$$\mathcal{F}(\mathbf{p}) = [M(\mathbf{p}), W(\mathbf{p})d_x, W(\mathbf{p})d_y, W(\mathbf{p})s], \tag{17}$$

where M is a truncated Gaussian centered at $\mathbf{p}$ and W is its binary support. Their spatial extents are fixed and independent of s. For global wind, we set $M = W = 1$.

The force field is broadcast across the temporal latent dimension without changing its spatial support. This broadcasting describes how the condition is provided to the generator rather than how long the physical force is applied. A localized point force represents an impulse at the beginning of the interaction, whereas global wind remains active throughout the video. The current representation does not explicitly control arbitrary onset times or force durations. Since its center remains fixed, the condition contains no future displacement or object trajectory.

LFIE projects $\mathcal{F}$ into the video latent space through two complementary branches. The frozen branch constructs the magnitude-aware Cartesian components

$$F_x = Msd_x, \qquad F_y = Msd_y, \tag{18}$$

which are converted into the native control format and encoded by the frozen video VAE to obtain the Directional Force Embedding $\mathbf{z}_{\text{DFE}}$. In parallel, a trainable adapter $\mathcal{A}_\psi$ with a hidden dimension of 32 projects the complete four-channel field onto the 48-channel Wan2.2 latent grid. The final force embedding is

$$\mathbf{z}_f = \mathbf{z}_{\text{DFE}} + \mathcal{A}_\psi(\mathcal{F}), \tag{19}$$

where the output projection of $\mathcal{A}_\psi$ is initialized to zero.

PAN has a hidden dimension of 64 and aligns the predicted inter-frame latent response with the corresponding ground-truth latent difference using an ℓ_1 objective. We use a temporal interval of one and set the attribution-loss weight to $\lambda_{\text{attr}} = 0.02$. The LFIE adapter, PAN, and rank-64 DiT LoRA modules are jointly optimized, while the remaining DiT parameters are frozen. LoRA modules are inserted into the query, key, value, and output projections of every self-attention layer and the two feed-forward projections.

This phase is trained for 4,600 optimization steps using AdamW with a learning rate of 1×10^{-4}, a weight decay of 0.01, a cosine schedule, 100 warm-up steps, and four-step gradient accumulation. We sample 49-frame and 81-frame clips with probabilities of 0.7 and 0.3, respectively, at a fixed resolution of 832×480. Clips always start from the first frame, and the force condition are independently dropped with probability 10%. Force-free WISA-80K samples use an all-zero force field.

Stage 2: Physics-Aware Reinforcement Learning. PARL is initialized from the 4,600-step Stage-1 checkpoint and optimized with DiffusionNFT. We use 100 camera-stationary, force-free videos selected from WISA-80K. Each rollout contains 49 frames at 832×480 resolution and is generated using 30 denoising steps, a classifier-free guidance scale of 5.0, and a flow-shift value of 5.0. Each condition produces six candidate rollouts, and two condition groups are processed per update, yielding 12 rollouts per optimization step.

We train an additional rank-32 policy LoRA with an alpha of 64 and zero dropout. AdamW is used with a constant learning rate of 5×10^{-5}, a weight decay of 1×10^{-4}, $(\beta_1, \beta_2) = (0.9, 0.999)$, and a maximum gradient norm of 1.0. The DiffusionNFT interpolation coefficient is $\beta = 0.1$, and the reference regularization coefficient is 1×10^{-4}. Rewards are normalized within each condition group.

The physics reward is computed using a frozen V-JEPA 2.1 model. Each temporal window contains 16 frames, comprising eight context frames and eight prediction frames. A stride of eight produces five windows for each 49-frame rollout. Before encoding, the longest spatial side is resized to 384 pixels, followed by the native V-JEPA 2.1 preprocessing. The reward is defined as the negative ℓ_1 prediction error between the predictor and target-encoder features.

A VBench-based Dynamic Gate prevents degenerate rollouts from exploiting this predictability reward. Temporal stability and imaging quality are evaluated for every rollout, with thresholds of 0.96 and 0.61, respectively. When activated, an eligible rollout must have a dynamic ratio between 0.6 and 1.5 relative to the frozen Stage-1 policy under the same condition. Groups with fewer than two eligible rollouts are skipped.

Within each valid group, physics rewards are centered over the eligible rollouts and continuously mapped to optimality probabilities following Eq. 14. DiffusionNFT optimization is performed only over these eligible rollouts, while rejected rollouts contribute neither to the group statistics nor to the training objective.

Separately, we monitor drift from the DiffusionNFT regularization reference and refresh it when the drift exceeds 0.03, subject to a minimum interval of three updates and a maximum interval of ten updates. PARL is trained for 100 updates, with checkpoints evaluated every five updates on a fixed validation split. The update-55 checkpoint achieves the highest validation physics reward and is used once for the reported ID and OOD evaluations.

E Detailed Metric Definitions

We evaluate the generated videos using 14 metrics grouped into force alignment, video quality, and physics alignment.

(1) Force Alignment. These metrics evaluate whether the generated motion exhibits a valid response to the input force and whether its initial direction agrees with the specified force direction. For a local point force, motion is estimated within the force-affected region using dense optical flow, after subtracting the median background flow to compensate for global camera translation. For a global force, motion is evaluated over the entire frame. Directional metrics are computed over the initial

temporal window to avoid penalizing physically plausible direction changes caused by subsequent collisions or interactions.

Directional Cosine Similarity (DCS) is the cosine similarity between the dominant initial motion vector and the target force vector. A larger value indicates stronger directional agreement. **Direction Accuracy at** 30° **(DA@30)** and 45° **(DA@45)** report the proportions of videos whose angular errors are no greater than 30° and 45°, respectively. **Motion Response Rate (MRR)** measures the proportion of videos whose average motion magnitude within the evaluated region exceeds a predefined response threshold. **ROI Motion Magnitude (RMM)** reports the average optical-flow displacement within the force-affected region in pixels per frame, reflecting the strength of the generated motion response.

(2) Video Quality. We evaluate both semantic consistency and perceptual video quality. **CLIP-Text Score (CLIP-T)** is the mean cosine similarity between the text prompt and the individual generated frames. **PickScore** measures text-conditioned human-preference alignment by averaging the frame-level preference scores produced by the PickScore model. **Motion Smoothness** follows the AMT-based VBench implementation, which predicts intermediate frames between neighboring frames and compares them with the observed frames to evaluate temporal interpolation consistency. **Aesthetic Quality** measures the visual appeal of the generated content using the VBench aesthetic predictor. **Imaging Quality** evaluates low-level visual fidelity—including sharpness, distortion, and image degradation—using the VBench implementation.

(3) Physics Alignment. For GT80 cases with paired ground-truth videos, we follow the metric definitions and official implementation of Physics-IQ Motamed et al. (2026). Generated and reference videos are temporally aligned and resized to the same spatial resolution before evaluation. Motion masks are extracted using the official background-subtraction pipeline without optical flow or additional spatial registration.

Spatial IoU measures the overlap between temporally aggregated motion regions. *Spatiotemporal IoU (ST-IoU)* averages the frame-wise overlap between motion masks. *Weighted IoU* compares motion-frequency maps, assigning greater importance to persistently active regions. *Mean Squared Error (MSE)* measures the RGB difference between aligned generated and reference frames. We report these metrics separately because GT80 provides only one reference outcome per case and therefore does not support the physical-variance normalization required by the official Physics-IQ Score.

F Detailed Dataset Description

Stage 1 Training Data. The force-conditioned training set contains four datasets. GoalForce-Balls Gillman et al. (2026b) contains approximately 6K videos of ball interactions, while GoalForce-Dominos and GoalForce-Plants contain approximately 3K videos each. These datasets provide localized point-force conditions. We additionally use 15,359 videos from the ForcePrompting training data Gillman et al. (2026a), which provide global wind-force conditions. Stage 1 therefore contains approximately 27K force-conditioned videos covering both localized and global forces. Each sample provides a source video, text description, reference image, and temporally aligned force condition.

We further include 3K force-free videos from WISA-80K Wang et al. (2026c) to improve robustness to natural scenes and support the subsequent reward-based training. These videos do not provide force annotations or paired physical outcomes.

Stage 2 Training Data. For reward-based post-training, we select static-camera videos from WISA-80K Wang et al. (2026c). Videos with noticeable camera motion are excluded because camera displacement may be incorrectly interpreted as object dynamics by the reward model. Stage 2 uses neither force annotations nor paired target videos. No sample from either evaluation benchmark is used in Stage 1 or Stage 2.

OOD Evaluation. Following the official ForcePrompting protocol Gillman et al. (2026a), we use its complete OOD test set, which contains 357 cases comprising 305 point-force cases and 52 wind-force cases. This test set is disjoint from the force-conditioned training data and contains unseen

scenes and objects. Each case specifies the force location, direction, and magnitude. We use this benchmark to evaluate force alignment and video quality.

ID Evaluation. We construct GT80 from 80 videos held out from the four force-conditioned training domains before training. It contains 20 ball, 20 domino, 20 plant, and 20 wind cases, none of which are included in the training set. Because these cases follow the same scene and interaction distributions as the training data, GT80 serves as the in-distribution benchmark. Its paired ground-truth videos enable the evaluation of force alignment, video quality, and GT-based physics alignment.

Demo Dataset. To examine generalization under more challenging conditions, we use GPT-6 Astra Singh et al. (2025) to construct a demo set following the ForcePrompting test format. This set is used exclusively for qualitative demonstration and is excluded from all quantitative evaluations. Examples on the project page show that PhysForce generalizes across diverse scenes and predicts plausible state changes according to object properties and applied forces.

G Limitations and Future Work

Limitations. Although PhysForce produces perspective-consistent and scale-consistent motion that suggests implicit 3D reasoning, it operates entirely in the 2D video latent space. The force location and direction are defined in image coordinates, without explicitly modeling scene depth, 3D geometry, camera pose, or out-of-plane forces. Consequently, the model may struggle with interactions involving depth-dependent contact, occlusion, or complex 3D trajectories. Moreover, PhysForce generates each video through offline bidirectional denoising, requiring the force condition to be specified before generation and preventing users from adjusting it in response to the evolving scene.

Future Work. We plan to extend LFIE with 3D-aware scene and force representations, including depth, camera geometry, and force vectors defined in 3D coordinates. Incorporating object geometry, material properties, and contact constraints may further improve reasoning about out-of-plane and multi-object interactions. We also aim to develop a causal streaming version of PhysForce that generates frames incrementally and accepts updated force locations, directions, and magnitudes during generation. Combined with diffusion distillation and temporal-state caching, this could enable low-latency, real-time interaction with generated physical worlds.