LEAP: LATENT ENCODER ALIGNMENT FOR PHYSICS REWARDS IN VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Reward Feedback Learning is a powerful paradigm for generative modeling, yet its extension to Text-to-Video (T2V) generation is hindered by intractable computational overhead and the scarcity of differentiable reward signals. Although mapping rewards to the latent space reduces computational burdens, existing latent rewards rely heavily on human preference datasets, inherently lacking the rich physical and structural priors of foundational models. In this work, we introduce a novel framework that constructs a latent physics reward by extracting and leveraging priors from a latent world model. We overcome two fundamental hurdles to achieve this integration: 1) developing an alignment mechanism with a lightweight Aligner to connect the generative VAE latent space with the target prior embedding space, and 2) adapting the feedback learning algorithm to stably incorporate these latent priors during training. Our method is the first to demonstrate the viability of projecting T2V latent representations into the V-JEPA 2 vision embedding space for enforcing physical plausibility in generated videos. Empirical results on VideoPhy and VideoPhy2 show that our latent physics reward substantially improves the physical realism and temporal consistency of generated videos, while training within 12.9 GB, where backpropagating the same reward through pixel space would require hundreds of GB.

1 INTRODUCTION

Text-to-video (T2V) generation has achieved remarkable visual fidelity, yet even state-of-the-art models frequently violate basic physical laws: objects accelerate without cause, liquids flow upward, and collisions produce impossible trajectories. Reward Feedback Learning (ReFL) (Xu et al., 2023; Clark et al., 2024; Prabhudesai et al., 2023), which directly backpropagates reward gradients into the generative model, has proven highly effective for aligning image generators with human intent. Extending ReFL to video (Prabhudesai et al., 2024) is prohibitively expensive, since the gradient propagates through a temporally unrolled denoiser, a VAE decoder, and a pixel-space reward network, causing memory and compute costs to grow rapidly.

Recent works attempt to circumvent the cost by evaluating the reward in the latent space of the generator, replacing the pixel-space reward with a *latent reward model* (LRM) that scores generator latents directly (Mi et al., 2025; Ding et al., 2025). While the substitution eliminates the decoder pass, existing LRMs are typically trained via a Bradley–Terry objective (Mi et al., 2025; Ding et al., 2025) on paired human preference data, a supervisory signal that is subjective, label-sparse, and shaped more by aesthetics than by physical correctness.

As an alternative to human preference data, self-supervised video foundation models have proven an effective source of physics supervision, since their representations encode rich physical dynamics learned from large-scale video without annotation. Existing methods exploit such models in two distinct roles. As an alignment target, VideoREPA (Zhang et al., 2025b) distils physics knowledge from a frozen self-supervised video encoder into a T2V model through a token-level relational matching loss, improving physics commonsense in the generated videos , but its alignment loss is defined on training videos rather than on the samples the model generates. As a world-rule expert, WMReward (Yuan et al., 2026) instead scores generation with V-JEPA 2 (Assran et al., 2025), a latent world model whose predictor anticipates masked future embeddings from context, treating its prediction error on generated frames as a measure of physics adherence. The resulting signal is

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.

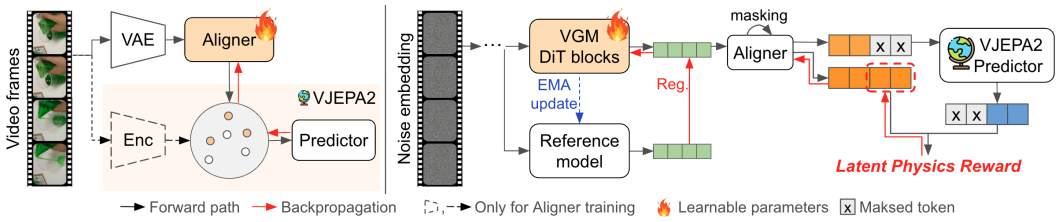


Figure 1: Pipeline of Latent Encoder Alignment for Physics reward (LEAP). (a) Aligner maps the Video Generative Model (VGM) latent representations to the pre-trained V-JEPA 2 feature space. (b) During feedback learning, active VGM features are aligned, masked, and predicted by a V-JEPA 2 Predictor. The prediction consistency serves as a physics-based reward to update the DiT blocks, while an EMA-updated Reference model provides regularization to velocity space.

more directly tied to physics than relational alignment, but computing it requires decoding generator latents into pixel-space video at every reward evaluation, which confines it to inference time.

Our key insight is that the world-rule-expert role can be lifted into the generator's latent space, turning V-JEPA 2's predictor surprise into a differentiable training-time reward without pixel decoding or human preference labels. Realizing this insight poses two technical challenges. First, the VAE latent space of a T2V model and V-JEPA 2's embedding space are structurally incompatible, differing in resolution, channel dimension, and semantic granularity. We bridge them with a lightweight *Aligner*. Its two learnable stages are a Q-Former (Li et al., 2023) cross-attention front-end whose queries absorb that mismatch and a stack of RoPE-based (Su et al., 2024) self-attention blocks that mirror V-JEPA 2's encoder, so the output is directly compatible with the frozen predictor; a parameter-free Scaler then restores V-JEPA 2's native embedding scale. The Aligner draws supervision from two sources: direct alignment with the encoder output, and reward-level alignment propagated through the frozen predictor.

Second, we need a training signal that is physically grounded and stable to optimize. We use the predictor's surprise on the Aligner's output as the latent physics objective, minimizing it so that the generated dynamics become predictable under the world-model prior. Concretely, this surprise is the mismatch between the predictor's output and the actual future embeddings at the tokens we hold out. Optimizing this objective without constraint triggers the reward-hacking collapse familiar from ReFL (Xu et al., 2023; Prabhudesai et al., 2024). We therefore refine ReFL into *Latent Physics Feedback Learning* (LPFL), which couples the objective with an EMA velocity regularizer that anchors the generator to its pre-trained behavior. Together, the Aligner and LPFL form our framework, **LEAP** (Latent Encoder Alignment for Physics Rewards). The signal itself is unchanged, since V-JEPA 2's surprise is always computed in its embedding space; what LEAP changes is the route to that space, which the Aligner provides without VAE decoding and re-encoding. Moving rewards into latent space is an established remedy for the memory cost of backpropagating through the decoder (Prabhudesai et al., 2023; Clark et al., 2024; Prabhudesai et al., 2024; Ding et al., 2025), and with our 1.3B backbone the pixel route exhausts device memory before a single training step completes (Sec. 4.3).

We summarize our contributions as follows:

- A framework, LEAP, for extracting latent physics rewards from a frozen self-supervised latent world model (V-JEPA 2 (Assran et al., 2025)), eliminating both pixel-space decoding and human preference data from the reward feedback loop.
- An Aligner module that bridges the T2V VAE latent space and the V-JEPA 2 embedding space via a Q-Former (Li et al., 2023) cross-attention front-end, a RoPE (Su et al., 2024) self-attention back-end, and a parameter-free Scaler, trained with combined embedding-alignment, surprise-score, and predictor-consistency objectives while keeping the T2V backbone and V-JEPA 2 frozen.
- LPFL, a refined ReFL algorithm that operates entirely in latent space and combines an aligner-mediated physics reward with an EMA velocity regularizer and alternating supervised flow-matching updates to stably fine-tune the T2V generator.
- Consistent improvements over the baseline on VideoPhy (Bansal et al., 2024) and VideoPhy2 (Bansal et al., 2025), with the best Joint scores on both benchmarks and the best physical-

commonsense score on the harder VideoPhy2, demonstrating that world model priors are a principled and efficient alternative to human-preference rewards for physical video alignment.

2 Related Work

Reward Feedback Learning. Reward Feedback Learning (ReFL) aligns diffusion models with downstream objectives by backpropagating reward gradients from a differentiable reward model into the denoiser. ImageReward (Xu et al., 2023), AlignProp (Prabhudesai et al., 2023), DRaFT (Clark et al., 2024), and DRTune (Wu et al., 2024) establish the recipe on images, and follow-ups extend it to video (Prabhudesai et al., 2024), where the pixel-space backward pass becomes the main bottleneck, with memory and compute costs growing rapidly in frame count and resolution. We retain the differentiable-reward formulation but execute the entire feedback loop in latent space.

Latent-space alignment and rewards. A complementary line bypasses pixel-space rewards by scoring directly on diffusion latents. On images, this includes preference optimization over paired latent samples (Zhang et al., 2025a). On video, similar ideas appear as latent critics trained on human preference data (Ding et al., 2025; Mi et al., 2025) that score generated latents directly inside the fine-tuning loop. Such critics reflect perceptual preference and fail to govern the correlation with physical correctness. We replace the preference-trained latent critic with a frozen self-supervised world-model prior, exposing physical-plausibility gradients without any preference annotation.

Physics priors and video generation. Several works inject physics priors into video generation. One stream injects physics knowledge through vision-language models: ProPhy (Wang et al., 2025) transfers VLM physical reasoning into the generator via a Mixture-of-Physics-Experts, and PhysVid (Arani et al., 2026) fuses VLM-generated chunk-level physics annotations into the backbone. Self-supervised video models internalize physical dynamics without pixel reconstruction, including JEPA-style encoders (Bardes et al., 2024; Assran et al., 2025) and masked-modeling encoders (Tong et al., 2022; Wang et al., 2022). VideoREPA (Zhang et al., 2025b) treats the world model as a feature extractor, distilling physics knowledge into a T2V backbone via a token-level relational alignment loss against a frozen self-supervised video encoder. As this loss is computed on real videos, the encoder acts as a fixed alignment target rather than a critic of generated content, making it complementary to, and composable with, a reward on the model's own rollouts. WMReward (Yuan et al., 2026) instead uses the encoder–predictor surprise to steer denoising at inference time, which requires decoding frames to pixels. We instead propose a lightweight aligner that turns V-JEPA 2's predictive consistency into a tractable physics reward computed on the generator's own latents.

3 Method

3.1 Preliminary

3.1.1 V-JEPA 2

V-JEPA 2 is a self-supervised video world model trained via a joint-embedding predictive architecture (JEPA). It consists of an encoder E_θ and a predictor P_ϕ. The encoder maps an input video x to a sequence of spatiotemporal token embeddings $e = E_\theta(x) \in \mathbb{R}^{N\times d}$, where N is the number of spatiotemporal tokens and d is the embedding dimension. During training, a subset of input patches is designated as the context and the remaining patches are masked. The predictor P_ϕ then operates entirely in embedding space and takes the embedding of context e_{ctx} together with positional mask tokens at the target locations and predicts the future embeddings $\hat{e}_{\text{fut}} = P_\phi(e_{\text{ctx}})$.

By minimizing the prediction error between predicted future embeddings and target masked embeddings, V-JEPA 2 learns rich representations of physical dynamics, making the prediction error a natural measure of physical implausibility, which is formalized as a *surprise score* using cosine distance (Yuan et al., 2026).

3.1.2 ReFL & Latent Reward Model

Reward Feedback Learning. The ReFL method introduced by ImageReward (Xu et al., 2023), finetunes a pre-trained diffusion generator g_θ by backpropagating gradients of a differentiable reward

model R, a scorer trained on human preference comparisons, directly into the generator parameters θ. Given a prompt c, ReFL first runs the reverse diffusion process *without* gradients from z_T down to a randomly sampled late timestep $t \in [T_1, T_2]$, then performs a single gradient-tracked denoising step to get z_{t-1}. With noise scheduler, it obtains clean estimate $\hat{z}_0$ from z_{t-1} and decode it to pixel-space sequence $\hat{x}_0$. The generator is updated to maximize the expected reward,

$$\mathcal{L}_{\text{ReFL}}(\theta) = -\mathbb{E}_{c,\, t\in[T_1,T_2]}[R(\hat{x}_0, c)]\,, \tag{1}$$

restricting gradient flow to the final denoising step so that memory remains tractable while still pushing θ toward higher-reward samples. In practice, $\mathcal{L}_{\text{ReFL}}$ is jointly minimized with the original diffusion pre-training loss on real data to prevent reward over-optimization and preserve generation fidelity. While effective for image generation, this formulation becomes prohibitively expensive in the video setting even with one-step truncation

Latent Reward Model. Recent work (Mi et al., 2025; Ding et al., 2025; Zhang et al., 2025a) replaces the pixel-space reward $R(\hat{x}, c)$ with a *latent reward model* (LRM) R_η that operates directly on the generator's latent $\hat{z}$, eliminating the decoder pass entirely:

$$\mathcal{L}_{\text{LRM}}(\theta) = -\mathbb{E}_c[R_\eta(\hat{z}, c)]\,. \tag{2}$$

The latent reward R_η is typically trained on paired human preference data $\mathcal{D} = \{(\hat{z}^+, \hat{z}^-, c)\}$ via a Bradley–Terry objective (Liu et al., 2025), yielding a scalar critic that scores latents according to annotator preference. We instead replace this preference-trained critic with V-JEPA 2 as a latent physics reward, realized by an Aligner that projects T2V VAE latents into its embedding space (Sec. 3.2) and a refined feedback-learning algorithm (Sec. 3.3).

3.2 Latent Space Alignment

The VAE latent space of a T2V model and the embedding space of V-JEPA 2 are structurally incompatible: they differ in resolution, channel dimension, and semantic granularity. We introduce a lightweight **Aligner** module $\mathcal{A}_\psi$ that projects VAE latents into the V-JEPA 2 embedding space, enabling the frozen world model to provide physics reward signal without any pixel-space decoding.

3.2.1 Architecture

Let $z \in \mathbb{R}^{K\times h\times w\times c}$ denote the VAE latent of a generated video, where K is its temporal length. The Aligner has two stages. A **Q-Former (Li et al., 2023) cross-attention front-end** first maps the VAE latent into the V-JEPA 2 token grid: N learnable queries $Q \in \mathbb{R}^{N\times d}$, one per V-JEPA 2 spatiotemporal position, attend to the spatiotemporally flattened VAE latent tokens, producing N intermediate tokens. This stage absorbs the resolution and channel-dimension mismatch between the two latent spaces. A stack of L **self-attention blocks with rotary position embeddings (RoPE) (Su et al., 2024)** then refines these tokens to produce the aligned embeddings $\hat{e} = \mathcal{A}_\psi(z, m) \in \mathbb{R}^{N'\times d}$, where m specifies the subset of $N' \le N$ visible tokens to be processed. V-JEPA 2's spatiotemporal masking schedule is applied between the two stages: *mask tokens are dropped before the RoPE stack*, so the Aligner is trained under the same context/target split as V-JEPA 2 itself. The RoPE back-end mirrors V-JEPA 2's encoder design so that $\hat{e}$ is drop-in compatible with the frozen predictor P_ϕ, which was trained on RoPE-encoded inputs and uses this encoding to attend across masked target positions during prediction.

3.2.2 Feature Rescaling

Since the VAE encoder and V-JEPA 2 are trained with different normalization schemes (the VAE operates in a normalized embedding space while V-JEPA 2 does not), the two representations differ substantially in scale. This mismatch both destabilizes alignment training and harms the frozen predictor P_ϕ, which expects inputs at the encoder's native scale. To address both issues, we introduce a **Scaler** that maps the Aligner output to V-JEPA 2's native embedding scale via a per-channel affine transform parameterized by the empirical mean $\mu \in \mathbb{R}^d$ and standard deviation $\sigma \in \mathbb{R}^d$ of V-JEPA 2 encoder embeddings, precomputed over a representative set of real videos:

$$\tilde{e} = \hat{e} \cdot \sqrt{\sigma^2 + \epsilon} + \mu \in \mathbb{R}^{N\times d}, \tag{3}$$

where the multiplication and addition are broadcast across the N token dimension and ϵ is a small constant for numerical stability. The scaled embedding $\tilde{e} \in \mathbb{R}^{N\times d}$ is used in place of $\hat{e}$ for all

Algorithm 1 Latent Physics Feedback Learning (LPFL)

Require: Pre-trained generator g_θ, frozen Aligner $\mathcal{A}_\psi$, frozen V-JEPA 2 predictor P_ϕ, real-video set $\mathcal{D}_{\text{real}}$, EMA decay τ, weight β.

Initialization: EMA initializes $\theta_{\text{EMA}} \leftarrow \theta$; The number of noise scheduler time steps T and the time step range for finetuing $[T_1, T_2]$.
for $(c^i, x^i) \in \mathcal{D}_{\text{real}}$ **do**
  // SFT step
  Sample $z_0^i \sim \mathcal{N}(0, I)$, $t^i \in [0, T]$.
  $\mathcal{L}_{\text{SFT}} \leftarrow g_\theta(z_t^i, t^i, c^i)$ // Update g_θ using SFT loss
  // LPFL step
  $t_{mid} \leftarrow rand(K, T')$ // Pick a random time step t_{mid}.
  for $j = T, T-1, ..., t_{mid}+1$ **do**
    no grad: $z_{j-1}^i \leftarrow z_j^i - \Delta t \cdot g_\theta(z_j^i, j, c^i)$
  end for
  with grad: $z_{t_{mid}}^i \leftarrow z_{t_{mid}+1}^i - \Delta t \cdot g_\theta(z_{t_{mid}+1}^i, t_{mid}+1, c^i)$
  Compute $\mathcal{R}_{\text{phys}}$ via 8 and $\mathcal{L}_{\text{EMA}}$ via 9.
  Update θ on $\mathcal{L}_{\text{LPFL}}$ from Eq. (10).
  Update EMA: $\theta_{\text{EMA}} \leftarrow \tau\theta_{\text{EMA}} + (1-\tau)\theta$.
end for

subsequent loss computations. The Scaler introduces no learnable parameters and requires no gradient flow, serving purely as a stable alignment anchor between the two representation spaces.

3.2.3 Training Objectives

Embedding alignment loss. We minimize the MSE between the scaled Aligner output and the V-JEPA 2 encoder output on the same frames:

$$\mathcal{L}_{\text{MSE}} = \|\tilde{e} - e\|_2^2 \tag{4}$$

This grounds the Aligner in the geometry of the V-JEPA 2 embedding space.

Latent surprise score loss. To imbue the Aligner with sensitivity to physical plausibility, we treat the V-JEPA 2 surprise score (Yuan et al., 2026), $s(a, b) := 1 - \cos(P_\phi(a),\ b)$, as a supervision target. We compute this score on the same video using V-JEPA 2's own encoder outputs ($e_{\text{ctx}}, e_{\text{fut}}$) and using the Aligner's outputs ($\tilde{e}_{\text{ctx}}, \tilde{e}_{\text{fut}}$), and minimize the L1 distance between the two:

$$\mathcal{L}_{\text{SS}} = \left\| s(\tilde{e}_{\text{ctx}}, \tilde{e}_{\text{fut}}) - s(e_{\text{ctx}}, e_{\text{fut}}) \right\|_1. \tag{5}$$

Minimizing $\mathcal{L}_{\text{SS}}$ makes the Aligner-induced surprise a faithful surrogate for the V-JEPA 2 prior.

Predictor consistency loss. We further enforce that the predictor's output is consistent regardless of whether its input came from the Aligner or the true V-JEPA 2 encoder:

$$\mathcal{L}_{\text{pred}} = \|P_\phi(\tilde{e}) - P_\phi(e)\|_1 \tag{6}$$

This L1 loss closes the gap between the two input pathways at the level of predictor outputs, providing a stronger functional alignment signal than embedding-space MSE alone.

The total training objective is:

$$\mathcal{L} = \mathcal{L}_{\text{MSE}} + \mathcal{L}_{\text{SS}} + \mathcal{L}_{\text{pred}} \tag{7}$$

3.3 LPFL: Latent Physics Feedback Learning

With the Aligner $\mathcal{A}_\psi$ in place, the V-JEPA 2 prior is now exposed as a differentiable function on the T2V VAE latent space. We then propose a procedure we call **Latent Physics Feedback Learning (LPFL)**. LPFL consists of two complementary terms: (i) a *latent physics reward* that supplies physical-plausibility gradients through the frozen Aligner, and (ii) an *EMA velocity regularizer* that anchors the generator to its pre-trained behavior, preventing the visual quality collapse that is endemic to reward-gradient fine-tuning.

Let $g_\theta(z_t, t, c)$ denote the flow-matching velocity field of the T2V backbone, parameterized by θ and conditioned on prompt c and noise level t. Following the truncated-backpropagation strategy of ReFL, we run $t_{mid} \in [T_1, T_2]$ sampling steps without gradients and only one gradient-tracked step; after that we recover a one-step clean latent estimate $\hat{z}_0$. Critically, $\hat{z}_0$ is never decoded to pixels; the entire reward computation is performed in latent space.

Latent physics reward. The surprise score (Yuan et al., 2026) originally requires pixel-space inputs to use V-JEPA 2 encoder. With Aligner A_ψ, we can build latent surprise score as in eq. (5) and apply it as latent physics reward. We construct mask $\{m_{ctx}, m_{pred}\}$, referring to positions of context and future tokens. A_ψ is applied to $\hat{z}_0$ with m_{ctx} and the full mask to obtain the context and full embeddings in the aligned space, respectively.

The reward is calculated as follows:

$$\begin{aligned} &\mathcal{R}_{\text{phys}}(\hat{z}_0) = 1 - \cos(sg[P_\phi(\tilde{e}_{\text{ctx}})],\ \tilde{e}_{\text{fut}})\,,\ \ where \\ &\tilde{e}_{\text{ctx}} = A_\psi(\hat{z}_0, m_{ctx}),\ \ \tilde{e}_{\text{fut}} = Mask(A_\psi(\hat{z}_0, m_{full}), m_{pred}), \end{aligned} \tag{8}$$

where m_{full} refers to a whole mask and function $Mask(\cdot)$ keeps the reference future tokens. We stop gradients of $P_\phi(\tilde{e}_{\text{ctx}})$ by $sg[\cdot]$ and the gradients of $\mathcal{R}_{\text{phys}}$ propagate through Aligner into $\hat{z}_0$ and ultimately into the generator parameters θ. Minimizing $\mathcal{R}_{\text{phys}}$ thus drives the T2V backbone to produce latents whose induced dynamics are predictable under the V-JEPA 2 prior, an objective, physics-grounded alignment signal that requires no human preference data and no pixel-space decoding.

EMA velocity regularization. Optimizing $\mathcal{R}_{\text{phys}}$ in isolation triggers the well-known reward-hacking failure mode (Amodei et al., 2016; Xu et al., 2023) of ReFL: the generator drifts off the natural-video manifold and produces degenerate samples that nevertheless score highly on the designed reward. To suppress this collapse, we maintain an exponential moving average θ_{EMA} of the generator weights and penalize the mismatch between the current and EMA velocity predictions on the same noisy latent:

$$\mathcal{L}_{\text{EMA}} = \left\| g_\theta(z_t, t, c) - g_{\theta_{\text{EMA}}}(z_t, t, c) \right\|_2^2. \tag{9}$$

This term acts as a soft anchor to the pre-trained generator, playing a role analogous to the KL penalty in RLHF but applied directly in velocity space. It preserves the visual fidelity and motion characteristics of the base model while still permitting the physics reward to reshape dynamics in physically meaningful directions.

Combined reward objective. We combine the two terms with a single trade-off coefficient β to form the LPFL reward objective:

$$\mathcal{L}_{\text{LPFL}} = \mathcal{R}_{\text{phys}}(\hat{z}_0) + \beta\, \mathcal{L}_{\text{EMA}}. \tag{10}$$

Alternating SFT regularization. To prevent overoptimization to the reward, we follow (Mi et al., 2025; Xu et al., 2023) and further regularize the fine-tuning loop with the original flow-matching supervised objective:

$$\mathcal{L}_{\text{SFT}} = \mathbb{E}_{t,(z,c)} \left[\left\| g_\theta(z_t, t, c) - (z_1 - z_0) \right\|_2^2 \right], \tag{11}$$

computed on real-video latents z_1 and noise z_0 along the standard linear interpolation $z_t = (1 - t)z_0 + tz_1$. $\mathcal{L}_{\text{SFT}}$ and $\mathcal{L}_{\text{LPFL}}$ work alternatively, as shown in Algorithm 1.

Only the generator parameters θ are updated during LPFL; the Aligner $\mathcal{A}_\psi$, the V-JEPA 2 predictor P_ϕ, and the VAE all remain frozen.

4 Experiments

4.1 Implementation Details

The Aligner $\mathcal{A}_\psi$ maps VAE latents of shape $K_v \times h \times w \times c = 21 \times 60 \times 104 \times 16$ to V-JEPA 2's $N \times d = 8192 \times 1408$ token grid via a Q-Former cross-attention front-end with N learnable queries of dimension d, followed by four RoPE self-attention blocks with $H = 22$ attention heads. We employ a joint masking strategy consisting of three complementary configurations: short-range tube masking ($\mathcal{M}_1$), long-range future blocks ($\mathcal{M}_2$), and future single-frame blocks ($\mathcal{M}_3$). The details are

Table 1: Comparison on VideoPhy and VideoPhy2 (%). **Bold**: best; underline: second best; "-": not reported in the original paper and not reproducible (no public code). WMReward is run on our baseline with best-of-4 selection.

Model	Size (B)	VideoPhy (%)			VideoPhy2 (%)		
		PC	SA	Joint	PC	SA	Joint
Baseline	1.3	23.55	48.26	12.21	62.44	29.55	25.38
SFT	1.3	26.16	48.84	12.50	63.61	31.39	27.21
PhysVid (Arani et al., 2026)	1.7	**32.00**	43.00	-	64.00	28.00	-
ProPhy (Wang et al., 2025)	1.3	-	-	-	65.00	**32.00**	26.50
WMReward (Yuan et al., 2026)	1.3	29.07	40.41	8.14	65.28	28.88	26.38
LEAP	1.3	29.36	**49.13**	**15.12**	**68.45**	31.55	**27.88**

Table 2: Pairs at each human-preference level for LEAP over Baseline.

Raters for LEAP	# pairs
100%	9
92%	1
80%	7
75%	3
67%	2
60%	6
20%	1
0%	1



Figure 2: Qualitative comparison across Baseline (Wan 2.1 1.3B), SFT-only model and LEAP.

Appendix A.1. Aligner training is done on real videos from UltraVideo (Xue et al., 2025); it runs for 5 epochs with batch size 16 and the Scaler statistics (μ, σ) are precomputed once on the same corpus.

We use *Wan2.1-1.3B* as the base T2V generator (480P, 81 frames), with V-JEPA 2 ViT-G/16 (Assran et al., 2025) (patch size 16, tubelet length 2) as the frozen physics prior; the V-JEPA 2 encoder E_θ and predictor P_ϕ remain frozen throughout, and we use the VAE from the base T2V model. LPFL fine-tunes all parameters of the base T2V generator , using noise scheduler timestep $T = 40$, finetuning range $[T_1 = 2, T_2 = 20]$. To calculate latent physics reward, we use sliding window masking strategy following (Yuan et al., 2026), that is, within a window of size W, latents of the first $W/2$ frames are flagged as context and the successive ones are masked for future prediction. We set $W = 16$ with stride of 8 throughout the experiment section. The other training details are stated in Appendix A.2.

Table 3: Cost of computing the V-JEPA 2 surprise through different routes (Wan2.1-1.3B, 480P, 81 frames, 64 GB accelerators). Memory and s/step are training-time costs, except for WMReward, whose time is per video (best-of-4); inference cost is relative to the base generator. ReFL memory is extrapolated and s/step a lower bound.

	WMReward	ReFL	**LEAP**
Reward space	pixel	pixel	latent
Peak mem (GB)	41.6	≈740 (OOM)	**12.9**
s/step	1290.8	≥357	**245.8**
Rel. infer. cost	4.11×	1.00×	**1.00**×

Table 4: Aligner ablation on VideoBench: Scaler (a), architecture (b–c), loss (d), and masking (e).

	VideoBench-test			
	ρ	τ	r	Acc
V-JEPA 2	-	-	-	<u>57.7</u>
(a) w/o Rescaling	-0.5916	-0.4070	0.5971	**58.8**
(b) w/o Learnable Query	<u>0.7339</u>	<u>0.5215</u>	<u>0.6990</u>	<u>57.7</u>
(c) w/o RoPE	0.4422	0.2968	0.4623	40.0
(d) MSE only	-0.6247	-0.4348	0.6345	54.1
(e) w/o $\mathcal{M}_3$	0.6293	0.4448	0.6336	47.1
LEAP	**0.7370**	**0.5302**	**0.7185**	54.1

4.2 Main Results: Physical Alignment of T2V Models

Metrics. We evaluate on two physical-plausibility benchmarks: VideoPhy (Bansal et al., 2024) and VideoPhy2 (Bansal et al., 2025), which includes 344 simple prompts and 600 harder prompts respectively. We use their auto-evaluation model, VideoCon-Physics from VideoPhy and VideoPhy-2-Autoeval from VideoPhy2 to measure both physical commonsense (PC) and semantic adherence (SA), and the joint score.

Quantitative comparison. Table 1 compares LEAP with the baseline, its SFT variant, and three physics-oriented methods. LEAP improves every metric over the baseline and achieves the best Joint score on both benchmarks, as well as the best PC on the harder VideoPhy2. It trails only PhysVid on VideoPhy PC, a model about 30% larger, and ProPhy on VideoPhy2 SA by 0.45, consistent with our reward targeting physical dynamics rather than semantic adherence. WMReward, which uses the same V-JEPA 2 surprise for best-of-4 selection at inference time, improves physics but substantially reduces SA (e.g., $48.26 \rightarrow 40.41$ on VideoPhy), whereas LEAP outperforms it on all six metrics without inference overhead (Sec. 4.3).

Human evaluation. We run a two-alternative forced-choice study on 30 baseline–LEAP pairs with matched prompts, collecting 220 judgments from 22 participants (protocol in Appendix B.2). LEAP is preferred at a mean per-pair rate of $\mathbf{76.3}\%$ and wins the majority vote on $\mathbf{28}$ of 30 pairs ($p < 0.001$, binomial test); Table 2 reports the distribution. No human judgment enters the reward at any stage.

Qualitative comparison. Figure 2 presents side-by-side video strips comparing the baseline, the SFT-only model, and our LEAP model on prompts that exercise physics regimes. For the first case, ours is more semantically coherent and demonstrates vivid waves; for the second, our generation shows realistic walking patterns in muddy water. Videos for these comparisons, along with additional examples, are available in the supplementary material.

4.3 Training and Inference Efficiency

We measure the peak memory and wall-clock time needed to obtain the V-JEPA 2 surprise along three routes that compute the same signal: WMReward's inference-time pixel route, a training-time pixel route (ReFL), and LEAP's latent route. As Table 3 shows, the pixel route does not fit on a 64 GB device at our training resolution, whereas LEAP trains at 12.9 GB and in less time per step; WMReward avoids training but multiplies inference cost by $4.11\times$. Measurement details and the scope of this comparison are given in Appendix A.3.

4.4 Ablations and Analysis

4.4.1 Aligner Ablation Study

Dataset and metrics. We evaluate on VideoBench-Test (Han et al., 2025), an out-of-distribution human preference set, forming a composite realism rating from its action-consistency, temporal-consistency, and motion-effects labels. The Pearson (r), Spearman (ρ), and Kendall (τ) correlations between Aligner surprise and this rating are our primary metric, as they reflect fidelity to the V-JEPA 2

Table 5: Ablations on LPFL training. We ablate the EMA regularizor (Alt.1), number of T_2 for step sampling (Alt.2) and the criterion (Cri.) of selecting Aligner A_ψ (Alt.3).

Method	$\mathcal{L}_{\text{EMA}}$	T_2	Cri.	VideoPhy			VideoPhy2		
				PA	SA	Joint	PA	SA	Joint
Alt.1	w/o	20	co.	20.64	47.09	12.21	62.94	32.22	28.21
Alt.2	w/	30	co.	22.09	45.06	10.76	63.94	30.55	27.05
Alt.3	w/	20	acc.	22.08	46.43	11.36	66.28	29.38	26.04
LEAP	w/	20	co.	29.36	49.12	15.12	68.45	31.55	27.88

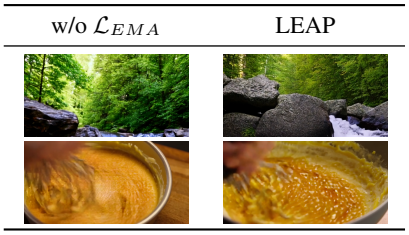


Figure 3: Degeneration without $\mathcal{L}_{EMA}$.

surprise geometry used to select the Aligner (Sec. 4.4.2); pairwise accuracy (Acc) is a secondary diagnostic.

Results. As Table 4 shows, the full Aligner achieves the highest correlations on all three measures. Removing feature rescaling (a) flips the rank correlations to negative, confirming that the distributional mismatch between the VAE and V-JEPA 2 spaces must be corrected explicitly. Replacing the learnable queries with non-parametric pooling (b) lowers every correlation despite a higher Acc, and removing RoPE (c) yields the lowest Acc of all variants, indicating that temporal positional structure is critical for dynamics-aware surprise. Training with $\mathcal{L}_{\text{MSE}}$ alone (d) inverts the rank correlations, highlighting the complementary roles of $\mathcal{L}_{\text{SS}}$ and $\mathcal{L}_{\text{pred}}$, while dropping the mask configuration $\mathcal{M}_3$ (e) lowers every correlation, showing that multiple masking regimes broaden the learned surprise signal. The full model does not lead on Acc, where raw V-JEPA 2 scores higher; since Acc rewards coarse pairwise ordering rather than fidelity to the surprise geometry, we treat it as secondary, consistent with Sec. 4.4.2, where selecting the Aligner by accuracy degrades downstream physical plausibility.

4.4.2 LPFL Ablation Study

EMA velocity regularizer. We compare LPFL with and without the EMA velocity regularizer $\mathcal{L}_{\text{EMA}}$ to isolate its contribution to preventing quality collapse. Table 5 (Alt.1) reports the performance on videophy&videophy2 and 3 show the visual results with and without $\mathcal{L}_{\text{EMA}}$, which introduce grid-like artifacts. Both quantitatively and qualitatively, $\mathcal{L}_{EMA}$ is important for anchoring the latent space and optimize the model to an physics-plausible direction.

Sampling step range. When random sample t_{mid} from $[T_1, T_2]$, T_2 controls the minimum number of sampling steps without gradient tracking and affect the quality of estimated clean latent $\hat{z}_0$. Table 5 (Alt.2) compares two T_2 values and shows more sampling steps results in better performance, where we can see a lower T_2 benefits the physics learning. Considering decreasing T_2 will slow the training, we did not use values lower than 20.

High-accuracy vs. high-correlation Aligner. The Aligner can be selected by either (i) its accuracy on the pairwise human-realism preference set (VideoBench) or (ii) its correlation with V-JEPA 2 surprise scores on VideoBench. We compare LPFL fine-tuning under both Aligner variants while keeping all other LPFL hyperparameters fixed. Table 5 (Alt.3) reports downstream VideoPhy / VideoPhy2 scores for each variant. The high-correlation Aligner yields better downstream physical plausibility, suggesting that fidelity to the V-JEPA 2 surprise geometry, not raw human-preference accuracy, is the more useful selection criterion for LPFL.

5 Conclusion

We presented **LEAP**, a framework that aligns text-to-video generators with the physical priors of a frozen self-supervised world model. It pairs a lightweight *Aligner* that projects T2V VAE latents into V-JEPA 2's embedding space with *Latent Physics Feedback Learning* (LPFL), a refined ReFL algorithm that adapts V-JEPA 2's surprise score into a differentiable latent reward regularized by an EMA velocity term. Together, they remove both the pixel-space decoder and the human-preference critic from the reward feedback loop, and we hope LEAP motivates a shift from human-preference-trained latent critics to self-supervised, domain-specific reward priors that can be stitched into reward feedback learning without ever leaving latent space.

AI USE STATEMENT

In this work, we used generative AI tools for editing only, grammar, spelling, and word choice. We did not use generative AI tools for research ideation, for deriving or verifying any of the methods, or for generating or analysing any of the experimental results reported here. All text has been reviewed by the authors, and we take responsibility for the final content of this work, including all claims and artifacts.

REFERENCES

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

Elahe Arani, Mykola Pechenizkiy, Bahram Zonooz, et al. Physvid: Physics aware local conditioning for generative video models. *arXiv preprint arXiv:2603.26285*, 2026.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, 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 Goldberg, Yizhou Sun, Kai-Wei Chang, Aditya Grover, and Chenfanfu Jiang. VideoPhy-2: A challenging action-centric physical commonsense evaluation in video generation. *arXiv preprint arXiv:2503.06800*, 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.

Kevin Clark, Paul Vicol, Kevin Swersky, and David J. Fleet. Directly fine-tuning diffusion models on differentiable rewards. In *International Conference on Learning Representations (ICLR)*, 2024.

Zihan Ding, Chi Jin, Difan Liu, Haitian Zheng, Krishna Kumar Singh, Qiang Zhang, Yan Kang, Zhe Lin, and Yuchen Liu. Dollar: Few-step video generation via distillation and latent reward optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17961–17971, 2025.

Hui Han, Siyuan Li, Jiaqi Chen, Yiwen Yuan, Yuling Wu, Yufan Deng, Chak Tou Leong, Hanwen Du, Junchen Fu, Youhua Li, et al. Video-bench: Human-aligned video generation benchmark. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18858–18868, 2025.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Menghan Xia, Xintao Wang, et al. Improving video generation with human feedback. *arXiv preprint arXiv:2501.13918*, 2025.

Xiaoyue Mi, Wenqing Yu, Jiesong Lian, Shibo Jie, Ruizhe Zhong, Zijun Liu, Guozhen Zhang, Zixiang Zhou, Zhiyong Xu, Yuan Zhou, Qinglin Lu, and Fan Tang. Video generation models are good latent reward models. *arXiv preprint arXiv:2511.21541*, 2025.

Mihir Prabhudesai, Anirudh Goyal, Deepak Pathak, and Katerina Fragkiadaki. Aligning text-to-image diffusion models with reward backpropagation. *arXiv preprint arXiv:2310.03739*, 2023.

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

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, Sen Xing, Guo Chen, Junting Pan, Jiashuo Yu, Yali Wang, Limin Wang, and Yu Qiao. InternVideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022.

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. *arXiv preprint arXiv:2512.05564*, 2025.

Xiaoshi Wu, Yiming Hao, Manyuan Zhang, Keqiang Sun, Zhaoyang Huang, Guanglu Song, Yu Liu, and Hongsheng Li. Deep reward supervisions for tuning text-to-image diffusion models. In *European Conference on Computer Vision (ECCV)*, 2024.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. ImageReward: Learning and evaluating human preferences for text-to-image generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Zhucun Xue, Jiangning Zhang, Teng Hu, Haoyang He, Yinan Chen, Yuxuan Cai, Yabiao Wang, Chengjie Wang, Yong Liu, Xiangtai Li, et al. Ultravideo: High-quality uhd video dataset with comprehensive captions. *arXiv preprint arXiv:2506.13691*, 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*, 2026.

Tao Zhang, Cheng Da, Kun Ding, Huan Yang, Kun Jin, Yan Li, Tingting Gao, Di Zhang, Shiming Xiang, and Chunhong Pan. Diffusion model as a noise-aware latent reward model for step-level preference optimization. *arXiv preprint arXiv:2502.01051*, 2025a.

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

A ADDITIONAL IMPLEMENTATION DETAILS

A.1 MASKING CONFIGURATION

We sample several mask configurations jointly for each training Aligner. We use three complementary configurations, summarized in Table 6. Each configuration specifies the number of blocks N_{blocks}, the (relative) spatial and temporal extents of each block, and the block aspect-ratio range. Spatial and temporal scales are expressed as fractions of the latent grid along each axis; e.g. a spatial scale of 0.15 corresponds to a block whose height and width are each 15% of the latent spatial extent. Configurations $\mathcal{M}_2$ and $\mathcal{M}_3$ further constrain the masked region to lie strictly in the future of the visible context, whereas $\mathcal{M}_1$ is temporally unconstrained. Throughout, we set `strict_past_ratio`$=0$ and `max_temporal_keep`$=1.0$ and impose no upper bound on the number of retained tokens (`max_keep`$=$`null`).

Table 6: Mask configurations used during training. Spatial and temporal scales are fractions of the latent grid along each axis. $\mathcal{M}_2$ and $\mathcal{M}_3$ are restricted to the future of the visible context; $\mathcal{M}_1$ is unconstrained.

Config	N_{blocks}	Spatial scale	Temporal scale	Aspect ratio
$\mathcal{M}_1$ (short-range)	8	0.15	1.00	$[0.75, 1.5]$
$\mathcal{M}_2$ (long-range future)	2	0.70	0.25	$[0.75, 1.5]$
$\mathcal{M}_3$ (future-frame)	2	0.70	0.00	$[0.75, 1.5]$

$\mathcal{M}_1$: short-range tube masking. Eight thin spatiotemporal tubes, each covering roughly 2% of the spatial area but spanning the full temporal extent of the **token sequence**, are masked. Because each tube spans the entire temporal axis, the predictor cannot rely on neighboring frames at the same spatial location and must instead exploit the surrounding spatial context. This configuration encourages appearance- and texture-level representations and serves as a regularizer against the trivial temporal-interpolation shortcut.

$\mathcal{M}_2$: long-range future block. Two large blocks, each covering roughly 49% of the spatial area and 25% of the temporal extent, are masked and constrained to lie strictly in the future of the visible context. This setting most directly exercises the world-modelling objective: given a partial past observation, the predictor must produce a substantial spatiotemporal chunk of the future, encouraging it to capture longer-horizon dynamics.

$\mathcal{M}_3$: future single-frame block. With `temporal_scale`$=0$ each block collapses to a single latent time step,[1] yielding two large spatial blocks within a single future frame. This produces a short-horizon, frame-level causal prediction task that complements $\mathcal{M}_2$ by emphasizing fine-grained next-frame dynamics rather than longer-horizon evolution.

Combined objective. At training time, masks from $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$ are sampled independently for each **input sequence** and applied jointly: every training **sample** is supervised on the union of all three masks.

A.2 OTHER TRAINING DETAILS

In LPFL training stage, we set EMA decay $\tau = 0.999$, and EMA-regularization weight $\beta = 0.001$; SFT alternation videos and LPFL rollouts also draw from UltraVideo. We chose the best checkpoint for each benchmark.
Both Aligner training and LPFL stages are optimized with AdamW (learning rate 5e-6, weight decay 0.01), with the Aligner using 8 Ascend NPUs ($\sim$8 hours) and LPFL using 32 Ascend NPUs ($\sim$24 hours).

[1] We clamp the temporal extent to a minimum of one latent frame.

A.3 Efficiency measurement methodology

This section specifies how each entry of Table 3 was obtained, and which quantities are measured rather than estimated.

Common setup. Except for the ReFL row, all numbers are measured on the same hardware under the paper's training configuration: 81 frames at 832×480, i.e. 21 VAE latent frames, with batch size 2. The two pipelines differ in exactly one stage. The ReFL route is VAE latent $\rightarrow$ VAE decoder $\rightarrow$ V-JEPA 2 encoder $\rightarrow$ predictor; the LEAP route is VAE latent $\rightarrow$ Aligner $\rightarrow$ predictor.

ReFL peak memory (extrapolated). The full configuration exhausts memory before a single step completes, so we measured it at reduced temporal length instead: 13.2 GB at one latent frame and 31.0 GB at two latent frames (batch size 1), a slope of $\approx$17.9 GB per additional latent frame. Linear extrapolation to 21 latent frames gives $\gtrsim 370$ GB at batch size 1 and $\approx$740 GB at batch size 2, assuming activations scale linearly. This is the only extrapolated quantity in the table.

ReFL seconds per step (estimated lower bound). Since the full ReFL path cannot run, we estimate it by substitution. The two routes are identical except that LEAP's Aligner pass is replaced by a VAE decode followed by a V-JEPA 2 encode; we measured that replacement at the same latent shape at 111.2 s (forward, plus backward with the VAE output detached so that the measurement itself does not exhaust memory). The reported $245.8 + 111.2 = 357.0$ s is therefore a *conservative lower bound*: it excludes the backward pass through the VAE decoder, which a real ReFL step cannot skip and which would add substantially to the total. The Aligner pass it replaces is negligible by comparison.

WMReward accounting. WMReward is applied as best-of-N selection: N candidates are generated per prompt and the one with the lowest V-JEPA 2 surprise is kept (N=4 here). Its cost per generated video is $(314 \text{ s generation} + 8.7 \text{ s scoring}) \times N$. Note that the s/step column therefore has different units across rows: it is per *training* step for ReFL and LEAP, but per *generated video* for WMReward, which requires no training at all.

Scope of the comparison. The claim is scoped to reward feedback learning, i.e. methods that evaluate a reward on the model's own generations and backpropagate through it; what LEAP removes is the VAE decoder and the pixel-space reward network from the backward graph. VideoREPA is not in this class and we make no efficiency claim relative to it: its Token Relation Distillation loss aligns intermediate generator features against real videos and requires no sampling rollout, so its per-step cost is lower than that of any reward-based method. That is a property of the objective class, not an efficiency result in either direction.

B Additional experiments

Table 7: Effect of the number of query blocks in the Aligner, evaluated on VideoBench-test.

	VideoBench-test			
# Blocks	ρ	τ	r	Acc
1	0.4896	0.3427	0.5008	**57.7**
4	**0.7370**	**0.5302**	**0.7185**	54.1
8	0.6612	0.4584	0.6881	48.2

B.1 Impact of model depth on Aligner

Table 7 investigates the effect of varying the number of RoPE attention blocks in the Aligner. We observe a significant performance jump when increasing the depth from 1 to 4 blocks, indicating that a single-layer Aligner lacks the capacity to effectively map VAE latents into the V-JEPA embedding space. However, further increasing the depth to 8 blocks degrades performance on VideoBench-test

across all four measures. These results suggest that a depth of 4 blocks provides a sufficient balance between model capacity and generalization for the alignment task.

B.2 Human evaluation protocol

We construct 30 test pairs, each consisting of a baseline generation and a LEAP generation from the same prompt. Each participant judges 10 pairs and selects, for each pair, the video that appears more physically plausible; no tie option is offered, and the presentation order and left/right placement are randomized per trial. We collect 220 judgments from 22 participants, yielding 5 to 12 ratings per pair. To prevent frequently rated pairs from dominating the aggregate, we first compute the preference rate within each pair and then average across pairs. Significance is assessed with a binomial test on the per-pair majority votes against a 50% null. As Table 2 shows, the preference for LEAP is at least 60% on 28 pairs and unanimous on 9 of them, while only 2 pairs favor the baseline. Because participants judge physical plausibility directly, these results indicate that the gains measured by the automatic evaluators correspond to perceptible improvements rather than evaluator artifacts. The study is used for evaluation only; human judgments never enter the training signal, which remains free of preference data by construction.

C Limitations and future work.

LEAP has two main practical limitations. First, the Aligner's quality is bottlenecked by the scale and diversity of the video data on which it is trained: rare physical phenomena under-represented in pre-training translate into noisier surprise gradients downstream, and scaling the Aligner to broader, higher-fidelity video corpora is a direct path to a stronger latent physics reward. Second, while the EMA velocity regularizer suppresses the most obvious reward-hacking failures, it is only one of several possible anti-hacking strategies, and combining it with complementary mechanisms is a promising direction for further stabilizing reward fine-tuning. Beyond these, our current results focus on a single T2V backbone and one world-model family; extending the aligner-mediated reward to other latent generators (e.g., autoregressive or rectified-flow models) and to other self-supervised priors (e.g., point-tracking or 3D-aware encoders) is a natural next step.