Predictive Structure Improves Video Diffusion Dynamics

Mi Luo[1†] Yujia Chen[2] Alex Dimakis[3,4]
Kristen Grauman[1] Wen-Sheng Chu[2‡] Du Tran[2‡]

[1]The University of Texas at Austin [2]Google
[3]UC Berkeley [4]Bespoke Labs

Abstract. Video diffusion models have achieved remarkable visual fidelity, yet often struggle to adhere to basic physical constraints, such as support, motion consistency, and object permanence, leading to implausible dynamics even in short clips. We revisit this challenge through the lens of the Information Bottleneck principle, arguing that physically grounded video generation should prioritize time-predictive representations over high-entropy appearance variations. Building on this perspective, we propose *Latent Dynamics Optimization* (LDO), a post-training framework that leverages a predictive latent world model (V-JEPA 2) to guide a pretrained video diffusion model towards dynamics-consistent generation without altering its sampling process. LDO operates on two levels: (1) aligning Gram-matrix geometry of intermediate diffusion features with the latent world model's casually predicted dynamics, and (2) encouraging temporally consistent rollouts with Group-Relative Policy Optimization (GRPO) using the world model as a dynamics critic. Experiments on VideoPhy and PisaBench show that LDO substantially improves physical commonsense, object permanence, and trajectory fidelity while preserving visual quality, suggesting that predictive latent supervision offers a practical route to make video generators not only photorealistic but also physically legible.

Keywords: Diffusion Model · World Model · Video Generation

1 Introduction

Our world unfolds with structure. A cup slips, falls, and shatters; a ball rolls, strikes an edge, and changes course. We do not perceive these events as isolated images but as structured evolution: objects persist through time, motion is continuous, and physical constraints shape what can happen next. Humans use this intuitive physics [4, 13, 43] to anticipate outcomes and detect when something looks wrong, even as visual details shift with lighting, texture, or viewpoint. Yet today's most advanced video generators [1, 33, 45, 48, 60] often struggle precisely where such structure matters. Video diffusion models can generate strikingly realistic frames, but even within short clips, they may violate fundamental physical

[†]Work done during an internship at Google.
[‡]Joint last authors.

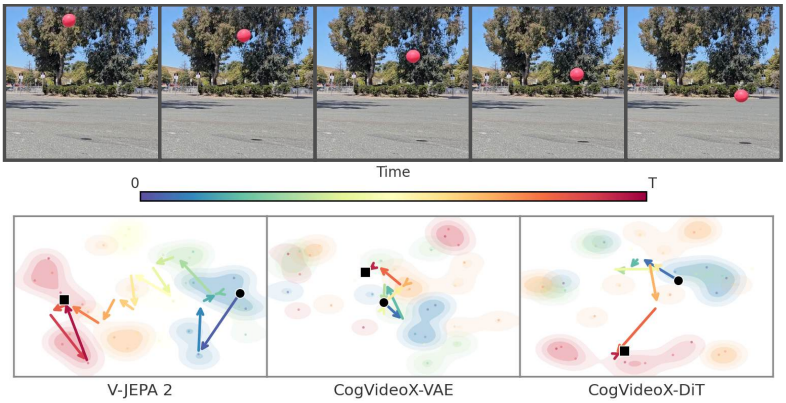

Fig. 1: t-SNE visualization of foreground token representations. We extract spatiotemporal patch tokens corresponding to a foreground object (ball undergoing parabolic motion) from three models: V-JEPA 2 encoder [2], CogVideoX-VAE [60], and CogVideoX-DiT [60]. Each point represents a single patch token, colored by its temporal position (blue→red). Contour lines outline regions of high token density at each time step. Arrows indicate the trajectory of per-timestep mean token positions. A smooth spectral gradient reflects coherent temporal encoding, while scattered colors indicate weak temporal structure. V-JEPA 2 exhibits well-separated temporal clusters with a coherent trajectory, demonstrating that its learned representations capture meaningful physical dynamics, motivating its use as a supervision signal for video generation models. In contrast, CogVideoX-DiT suffers from "temporal jitter" and lacks the global structural coherence of the World Model. (best view in color)

constraints such as support, contact, and motion consistency, or lose object permanence during brief occlusions. Despite each frame being visually sharp, these temporal inconsistencies can make the underlying dynamics implausible.

In this paper, we frame this gap through the lens of *Information Bottleneck* [46]. For conditional video generation, the model must translate a compact conditioning signal into a temporally extended sequence. A dynamics-aware solution should prioritize representation capacity for aspects that remain predictive across time, rather than focusing on incidental appearance variation. Existing work approaches this by incorporating physical inductive biases through language priors [56, 59], specialized vision encoders [66], or explicit, hardcoded physics modeling [26, 29, 53, 64]. We observe that latent predictive world models [2, 6] provide a concrete instantiation of this bias: by learning to predict masked temporal latent structure from visible context, they produce representations organized around underlying dynamical patterns. This suggests a route to strengthen diffusion generators without changing their sampling procedure: using a predictive latent model as a source of dynamics guidance.

We introduce **LDO** (*Latent Dynamics Optimization*), a post-training framework that uses a latent world model (instantiated as V-JEPA 2 [2]) to guide a pretrained video diffusion model toward dynamics-consistent generation. LDO provides guidance through two complementary perspectives. First, we distill temporal structure into the diffusion model by aligning the relational geometry of

intermediate diffusion features to V-JEPA 2 *causally predicted dynamics* computed over temporal bins, using Gram-matrix matching to transfer structure while remaining invariant to feature coordinates. Second, we optimize generation at the rollout level by treating the diffusion model as a stochastic policy and applying group-relative policy optimization (GRPO) with the V-JEPA 2 predictor as a dynamics critic. The critic scores groups of sampled videos by their long-horizon predictability under a causal predictor, and the resulting reward reinforces rollouts that remain temporally consistent.

Empirically, LDO delivers clear gains on established physics-oriented benchmarks. On VideoPhy [5], LDO improves Physical Commonsense by +13.6 absolute over the CogVideoX-5B base model while also increasing Semantic Adherence (+2.6), indicating that stronger dynamics need not come at the expense of prompt fidelity. On PisaBench [24], LDO reduces trajectory error and improves shape/object permanence metrics, reflecting more accurate gravity-driven motion and fewer structural failures under occlusion and contact. Taken together, these results suggest a simple but consequential message: world-structure is not a missing "dataset problem," but a missing "objective problem." By using a frozen predictive world model as a dynamics supervisor—without changing the diffusion sampler—LDO offers a practical path for the community to convert today's visually impressive video generators into models that are also physically legible, enabling more reliable simulation, editing, and embodied downstream use.

2 Related Work

Video Diffusion Models as World Simulators. Video generation has progressed from early frame prediction architectures to large-scale diffusion models capable of synthesizing diverse, open-domain dynamics [44, 55, 57, 60]. Recent works such as Sora [33] and COSMOS [1] have intensified interest in using generative models as "world simulators"—systems that implicitly learn the underlying causal laws of the physical world. However, emerging evidence suggests that scaling alone may not be sufficient for true physical grounding: PhyWorld [19] shows that even highly scaled diffusion models generalize poorly under out-of-distribution physical conditions, while PISA [24] demonstrates that post-training on specific dynamics does not yield robust physical generalization. These results suggest that current diffusion models act as *correlational world mimics* that reproduce pixel-level regularities without internalizing underlying causal mechanisms. We argue this stems from the training objective that prioritizes local denoising over long-range physical structure. Our work bridges this gap by shifting the optimization focus from pixel fidelity to latent predictability, using a predictive world model as a structural prior.

Physics Understanding in Foundation Models. Intuitive physics—the ability to make structured predictions about object motion and interaction [3, 7, 42]—is a central goal in both cognitive science and video foundation models. Both Simulation-based reasoning [14, 47] and learning-centric approaches [52] have been explored as pathways toward physical understanding. While benchmarks like Physion [8], VideoPhy [5] evaluate human-like judgments, they con-

sistently reveal a persistent physical gap in current systems. Recent evidence from LikePhys [62] and PhysWorld [19] suggest that large-scale models often rely on visual shortcuts, reproducing the appearance of motion without internalizing underlying causal constraints. This deficiency is further underscored by studies [10, 63] showing that state-of-the-art models struggle with principles like permanence and solidity, failing to capture the latent knowledge required for accurate physical evolution. Even when models do capture relevant physical properties, PhysVid [65] demonstrates that these latent representations are often not efficiently utilized during synthesis. To bridge this gap, we propose treating predictive stability as a proxy for physical grounding to steer diffusion toward causally consistent dynamics.

Alignment for Diffusion Model. Scaling laws [20] have improved visual fidelity [19], but ensuring diffusion models follow structural representations and human-centric values remains challenging. *Representation alignment* methods impose structural priors by emphasizing spatial similarity over classification accuracy (REPA [40, 61]) and extend to physics-aware video generation (VideoREPA [66]). To address motion and geometric inconsistencies, existing work either disentangles dynamics from appearance or enforces 3D coherence via geometric priors [9, 51]. In parallel, *preference-based* methods steer models toward subjective quality and adherence. For video, VideoDPO [28] and DenseDPO [54] mitigate motion bias and improve fine-grained detail via localized preference signals. Recently, GRPO-style training stabilizes RL training for diffusion and flow-matching models through stochastic exploration [25, 27, 58]. Different from these specialized directions—either hard-coding geometric constraints, disentangling dynamics from appearance, or relying on human/heuristic preference signals—LDO uses a frozen predictive world model as a dynamics teacher and critic, injecting a causal, time-predictive structural prior.

3 Why does Diffusion fail to learn world structures?

3.1 The Problem Statement

World structures refer to the patterned regularities in how the world appears and evolves over time. In this paper, we view **physical laws** as a subset of these structures and focus on the **physically-plausible conditional video generation** task, where we train a model with parameters θ that defines a distribution $p_\theta(Y \mid C)$ over a target video Y, conditioned on a signal C (*e.g.*, text, an image, or a partial video). The model is trained to sample $Y \sim p_\theta(\cdot \mid C)$ such that the generated video adheres to relevant physical constraints.

3.2 An Information Bottleneck (IB) Perspective

What does "physics" mean in videos? We view each observed frame x_t as a rendering of an underlying, low-dimensional physical state s_t (*e.g.*, positions, velocities, identities, contacts) and high-dimensional nuisance factors n_t (*e.g.*,

texture, lighting, camera noise). Then the physical state evolves over time from s_t to s_{t+1} via a transition function F according to the true world dynamics:

$$x_t = M(s_t, n_t), \qquad s_{t+1} = F(s_t, u_t), \tag{1}$$

where u_t denotes external inputs like actions or forces, and M denotes the rendering map. Intuitively, for $p_\theta(Y \mid C)$, rather than "modeling pixels," the model must extract the underlying state s_t while filtering out the non-causal noise n_t. This transforms the complex evolution of a video into a predictable trajectory within a *latent representation space*.

To formalize what makes a "good" representation space for physics, we turn to the IB principle [46]. Suppose we learn a representation $R = f(C)$ from some context C. The IB principle characterizes a useful representation as one that maximally compresses the input C while preserving only the information necessary to predict the target Y:

$$\min_f \; I(R;C) \; - \; \beta \, I(R;Y), \tag{2}$$

where $I(.;.)$ denotes the mutual information and $\beta > 0$ is the tradeoff hyperparameter. For physically plausible video generation, the structure of the target Y is governed by the underlying temporal evolution of the low-dimensional physical state sequence $S = \{s_t\}$, distinct from the sequence of high-dimensional nuisance factors $N = \{n_t\}$. Crucially, if the nuisance factors are uninformative about the future once the physical trajectory is known, we have the conditional independence:

$$Y \perp\!\!\!\perp N \mid S. \tag{3}$$

Under this assumption, encoding the nuisance N into R increases the compression cost $I(R;C)$ without improving the predictive term $I(R;Y)$. Consequently, the IB optimum discards the stochastic nuisance n_t to favor a minimal sufficient statistic aligned with the causal state s_t.

What Diffusion/Flow-matching optimizes? Latent generative models (*e.g.*, diffusion [60], flow-matching [18]) learn a conditional distribution $p_\theta(Z \mid C)$ over VAE [21] latents $Z = \mathrm{Enc}_{\mathrm{VAE}}(Y)$. Abstractly, they minimize conditional cross-entropy decomposable into an irreducible entropy and a KL divergence:

$$\mathbb{E}_{p(Z,C)}\big[-\log p_\theta(Z \mid C)\big] = H(Z \mid C) + \mathbb{E}_{p(C)}\big[\mathrm{KL}\big(p(Z \mid C) \,\|\, p_\theta(Z \mid C)\big)\big]. \tag{4}$$

Crucially, VAE latents Z typically entangle physical states s and nuisances n. Consequently, this objective acts as a standard conditional likelihood rather than an IB-style bottleneck. In IB terms, Z fails to be a minimal sufficient statistic; the high-entropy nuisance n inflates the irreducible term $H(Z \mid C)$, forcing the model to allocate capacity to stochastics $H(n \mid C)$ rather than the low-dimensional physical dynamics in s.

What Latent World Models (V-JEPA) Optimize? In contrast, latent predictive models like V-JEPA [2, 6] bypass pixel-level reconstruction and actively avoid modeling the full conditional entropy $H(Z \mid C)$. Given a context region x_c and a masked target x_t, V-JEPA employs an encoder E and a predictor P to minimize the prediction error directly in the representation space:

$$\min \mathbb{E}\Big[\big\|P(E(x_c)) - \mathrm{sg}(E(x_t))\big\|_2^2\Big], \tag{5}$$

where $\mathrm{sg}(\cdot)$ is the stop-gradient operation. Unlike likelihood-based models that are forced to capture the irreducible nuisance entropy $H(n \mid C)$, this objective only reinforces features that reduce the predictive uncertainty $H(Z_t \mid Z_c)$. Because stochastic nuisances n_t are unpredictable across the context-target gap, attempting to predict them under an L_2 loss merely yields their mean. To minimize variance, the encoder learns to map these noise dimensions to zero or a constant, systematically discarding them.

In this way, JEPA acts as an implicit Information Bottleneck. Without a pixel-level decoder, it is not driven to maximize mutual information with the full pixel space, instead learning a minimal sufficient statistic of the causal state:

$$z = E(x) \approx \psi(s). \tag{6}$$

This shifts the objective from matching a high-entropy rendered distribution $p_\theta(Z \mid C)$ to purely learning low-dimensional physical dynamics.

To qualitatively validate this representation gap, Figure 1 compares the latent spaces of V-JEPA2 and a well-known diffusion model (CogVideoX [60]) for a simple physical process (a ball undergoing parabolic motion trajectory). We extract foreground spatiotemporal tokens from V-JEPA2, alongside CogVideoX's VAE and denoising transformer spaces, projecting them via t-SNE [31] color-coded by time (blue→red) with density contours and mean trajectories.

Under our established framework, if a model successfully discards the stochastic nuisance n_t to isolate the causal state s_t, its temporal evolution should form a smooth, predictable manifold ($Z_t \approx \psi(s_t)$). Conversely, representations burdened by irreducible nuisance entropy $H(n \mid C)$ will exhibit noisy temporal mixing. As shown, V-JEPA2 displays a highly coherent temporal gradient and smooth progression. In contrast, the CogVideoX components exhibit severe temporal entanglement and weaker global structure. This qualitatively reinforces that predictive JEPA objectives naturally induce dynamics-aligned, minimal sufficient statistics, whereas likelihood-driven models remain entangled with high-entropy appearance factors.

4 Latent Dynamics Optimization (LDO)

The Information Bottleneck analysis highlights a clear dichotomy: while diffusion models excel at modeling the full distribution of visual appearance, they lack the inherent predictive bottleneck that forces latent world models to capture structural dynamics. To bridge this gap, we introduce **LDO** (Latent Dynamics

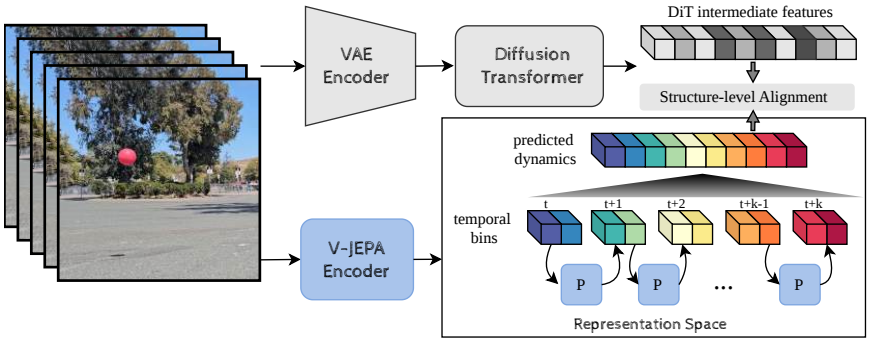

Fig. 2: Illustration of Predictive Dynamics Alignment (PDA). PDA distills predictive physical dynamics into the generative model by simultaneously processing input video frames through two parallel pathways. In the top generative path, the video passes through a VAE Encoder and a Diffusion Transformer (DiT) to produce intermediate spatiotemporal features. Concurrently, in the bottom teacher path, a frozen V-JEPA Encoder maps the frames into a latent representation space divided into discrete temporal bins. V-JEPA predictor network (P) iteratively forecasts future latent bins based on past causal context, generating features that encode a strong understanding of physical dynamics. These predicted dynamics are then distilled into the DiT's intermediate features via a structure-level alignment (Gram matrix matching), enforcing a causal dynamics bias within the diffusion model's internal representations.

Optimization), a post-training framework designed to retain the high-fidelity generative strengths of video diffusion while injecting a dynamics-centric bias derived from a frozen V-JEPA2 teacher. Rather than forcing the diffusion model to perfectly reconstruct the teacher's latent space—which risks degrading visual quality—LDO operates on the following two complementary levels.

4.1 Predictive Dynamics Alignment (PDA)

The first component, PDA, distills the causal prediction capability of a frozen latent world model (V-JEPA 2 [2]) directly into the intermediate representations of the diffusion transformer [35], as illustrated in Fig. 2.

Given a training video Y, the frozen V-JEPA 2 encoder $\mathcal{E}$ produces patch tokens $\mathbf{h} = \mathcal{E}(Y) \in \mathbb{R}^{N \times D}$, where $N = T_{\text{grid}} \times H_{\text{grid}} \times W_{\text{grid}}$ is the number of patch tokens resulted from encoding Y with V-JEPA 2 encoder. Concurrently, the diffusion transformer at a specific layer l produces alignment features $\mathbf{z}$, projected to dimension D via a learnable MLP and spatially downsampled.

The temporal axis is partitioned into K equal bins of $\Delta_t = T_{\text{grid}}/K$ steps each. For each target bin $k \in \{1, \ldots, K-1\}$, the frozen V-JEPA predictor $\mathcal{P}$ receives context tokens from bins 0 through $k-1$, a context mask $\text{ctx}^{(k)}$, and a target mask $\text{tgt}^{(k)}$ for bin k. Given $S = H_{\text{grid}} \times W_{\text{grid}}$, the predictor outputs target features representing its causal understanding of the physical dynamics:

$$\hat{\mathbf{h}}^{(k)} = \mathcal{P}\left(\mathbf{h}_{[0:k-1 \cdot \Delta_t \cdot S]}, \text{ctx}^{(k)}, \text{tgt}^{(k)}\right). \tag{7}$$

We use the relation distillation loss from [66] to align the Diffusion and V-JPEA2 features. Let $\tilde{\mathbf{z}}$ and $\hat{\tilde{\mathbf{h}}}$ denote the L2-normalized versions of the diffusion alignment features and the V-JEPA predictor outputs, respectively, restricted to valid temporal bins. We compute and match their Gram matrices (token-to-token cosine similarities): $\mathbf{G}_t^{\text{diff}} = \tilde{\mathbf{z}}_t \cdot \tilde{\mathbf{Z}}^\top, \qquad \mathbf{G}_t^{\text{pred}} = \hat{\tilde{\mathbf{h}}}_t \cdot \hat{\tilde{\mathbf{H}}}^\top$, where $\tilde{\mathbf{z}}_t \in \mathbb{R}^{S\times D}$ are spatial tokens at frame t; $\tilde{\mathbf{Z}} \in \mathbb{R}^{(T_{\text{valid}}\cdot S)\times D}$ the full spatiotemporal token set. The PDA loss enforces this structural difference within margin m:

$$\mathcal{L}_{\text{PDA}} = \frac{1}{T_{\text{valid}}} \sum_t \text{mean}\left(\text{ReLU}\left(\left|\mathbf{G}_t^{\text{diff}} - \mathbf{G}_t^{\text{pred}}\right| - m\right)\right). \tag{8}$$

4.2 GRPO with Structure-aware Reward

While PDA aligns internal feature representations, the second component directly optimizes the generated output distribution $p_\theta(Y \mid C)$ using Group Relative Policy Optimization (GRPO) [39], treating V-JEPA's predictive capability as a informative reward signal.

Reward Function. We use V-JEPA 2 [2] as a physics critic [13, 63]. Given a generated video Y, a sliding window scheme is applied to extract $\mathcal{W}$, a set of overlapping clips. For each clip $w \in \mathcal{W}$, the frozen V-JEPA encoder $\mathcal{E}$ processes the first C context frames, and the predictor $\mathcal{P}$ forecasts the tokens for the remaining future frames. The reward $r(Y)$ is the mean cosine similarity between the predicted tokens and the actual target-encoded tokens:

$$r(Y) = \frac{1}{|\mathcal{W}|} \sum_{w\in\mathcal{W}} \frac{1}{N_{\text{pred}}} \sum_i \cos\left(\mathcal{P}\left(\mathbf{h}^w_{\text{pred}}\right)_i, \mathbf{h}^w_{\text{tgt},i}\right) \tag{9}$$

GRPO Optimization. For each conditioning prompt C, G videos are generated via stochastic DDIM [41] ($\eta > 0$). The group-relative advantage for sample i is calculated as: $A_i = \frac{r_i - \mu_G}{\sigma_G + \epsilon}$, where μ_G and σ_G are the batch reward mean and stddev. Advantages are clipped to $[-c_{\text{adv}}, c_{\text{adv}}]$. For a randomly selected fraction ρ of denoising steps, the policy gradient loss applies PPO-style clipping:

$$\mathcal{L}_{\text{GRPO}} = -\frac{1}{G \cdot T_{\text{train}}} \sum_{i,t} \max\left(A_i \cdot \rho_{i,t}, A_i \cdot \text{clip}(\rho_{i,t}, 1-\epsilon, 1+\epsilon)\right), \tag{10}$$

where the importance ratio $\rho_{i,t} = \exp\left(\log \pi_\theta(\mathbf{x}^{(i)}_{t-1} \mid \mathbf{x}^{(i)}_t) - \log \pi_{\text{old}}(\mathbf{x}^{(i)}_{t-1} \mid \mathbf{x}^{(i)}_t)\right)$ is derived from the Gaussian transition density of stochastic DDIM:

$$\log \pi(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = -\frac{|\mathbf{x}_{t-1} - \mu_\theta(\mathbf{x}_t, t)|^2}{2\sigma_t^2} - \log \sigma_t - \frac{d}{2}\log(2\pi), \tag{11}$$

with $\sigma_t = \eta\sqrt{\tilde{\beta}_t}$. Here, η controls DDIM noise injection ($\eta = 0$ deterministic), and $\tilde{\beta}_t$ is the schedule-derived effective variance at step t.

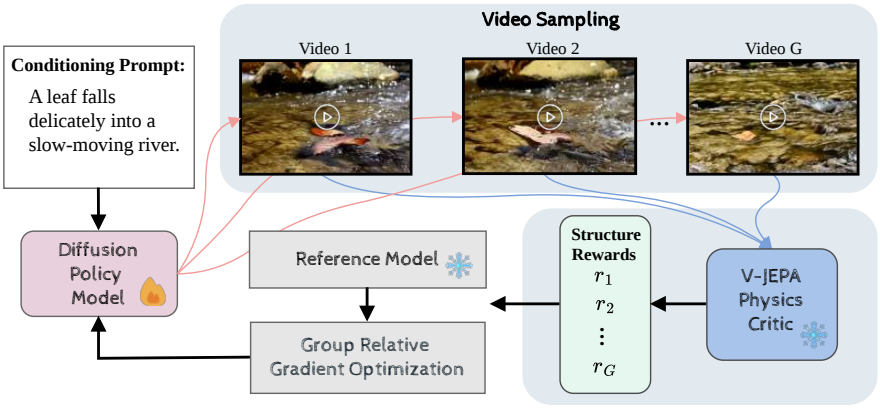

Fig. 3: Illustration of Group Relative Policy Optimization (GRPO) with V-JEPA Causal Reward. Given a conditioning prompt, the Diffusion Policy Model generates a group of G sampled videos. These generated videos are evaluated by a frozen V-JEPA Physics Critic, which computes structure rewards $(r_1, \ldots, r_G)$ based on the causal predictability of the generated dynamics over sliding windows. Using these rewards, Group Relative Gradient Optimization is applied—comparing the policy model against a Reference Model—to calculate group-relative advantages. This directly optimizes the generated output distribution, updating the diffusion policy to favor dynamically coherent trajectories.

4.3 Unified Training Objective

Optimizing the dense V-JEPA distillation graph and the GRPO generation rollout simultaneously is infeasible due to GPU memory constraints. Therefore, we design LDO as an interleaved training framework where the total loss at step n is time-dependent:

$$\mathcal{L}_{\text{total}}^{(n)} = \mathcal{L}_{\text{diffusion}} + \lambda \mathcal{L}_{\text{PDA}} + \gamma(n)\mathcal{L}_{\text{GRPO}} \tag{12}$$

During standard steps ($\gamma(n) = 0$), the model optimizes only the diffusion and PDA losses. Every m steps, $\gamma(n)$ activates to prioritize policy-gradient updates ($\mathcal{L}_{\text{GRPO}}$) for global dynamic coherence while staying within memory limits; full objectives and scheduling details are in the supplement.

5 Experiments

5.1 Evaluation Settings

To evaluate our method's ability to model world dynamics and physical commonsense, we conduct experiments on the VideoPhy [5] and PisaBench [24].

VideoPhy [5] assesses how well text-to-video models adhere to real-world physical laws and commonsense reasoning. The benchmark comprises 344 text prompts describing complex physical interactions across three categories: solid-solid, solid-fluid, and fluid-fluid. We employ the automatic VLM-based evaluator *videocon_physics* [5] provided by the authors, which uses templated queries to

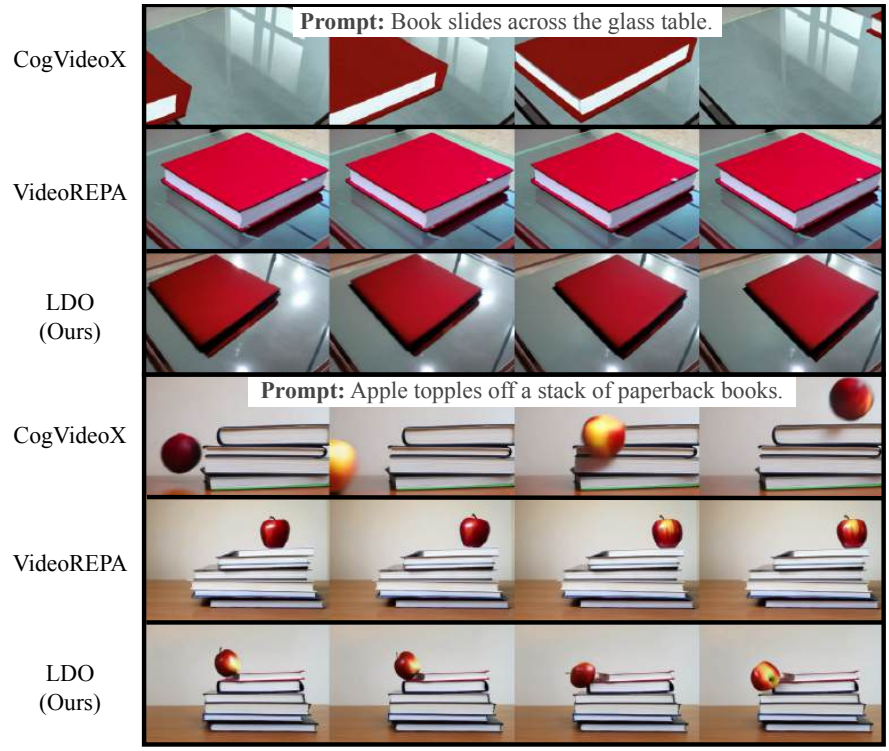

Fig. 4: Qualitative comparison on VideoPhy. LDO better captures physics, preserving state changes and support/contact dynamics compared to CogVideoX and VideoREPA, which produce physically implausible outcomes—*e.g.*, books sliding across a glass table but unrealistically floating above the surface (top), or an apple hovering mid-air instead of falling (bottom)—LDO maintains proper contact and support, keeping the books sliding on the tabletop and the apple yielding to gravity.

compute Semantic Adherence (SA) and Physical Commonsense (PC) scores, applying a binarization threshold of 0.5.

PisaBench [24] evaluates a model's understanding of fundamental physics—specifically gravity—in an image-to-video setting. Given an initial image of a suspended object, the model must generate a realistic sequence of the object falling and interacting with the ground. The dataset contains 361 real-world and 60 simulated dropping scenarios. Performance is measured against ground-truth videos using three primary metrics: Trajectory L2 evaluates motion accuracy by computing the average Euclidean distance between object centroids; Chamfer Distance (CD) [11] assesses shape fidelity by measuring the distance between generated and ground-truth object masks; and Intersection over Union (IoU) [37] tracks object permanence by calculating the overlap of segmentation masks across frames. Frame masks are extracted using SAM 2 [36].

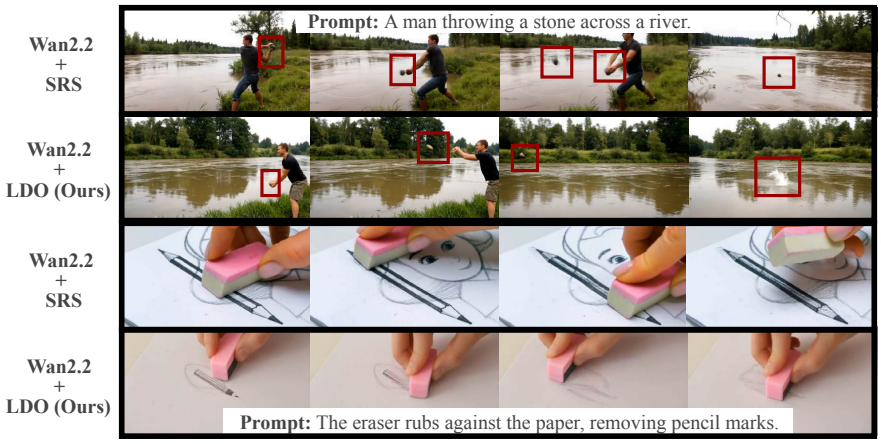

Fig. 5: Qualitative comparison on VideoPhy with base model Wan2.2-T2V-A14B. SRS denotes Self-Refining Sampling [16]. Our LDO better captures object permanence (the stone example), and contact dynamics (the eraser example).

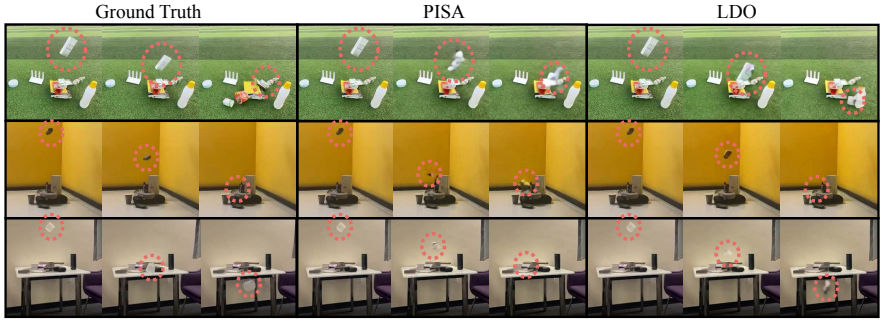

Fig. 6: Qualitative comparison on PisaBench. LDO improves shape consistency and physical realism over the baseline, eliminating artifacts like unnatural melting or disappearing objects. In the 3rd row, LDO correctly predicts the object falling past the table edge under gravity, whereas the baseline unnaturally melts the object.

5.2 Training Strategy

For VideoPhy, LDO is trained on OpenVid [32], a large-scale open-world video dataset, with videos resized to a resolution of $49 \times 480 \times 720$. For the PisaBench evaluation, we train LDO on their officially released training set of $5,000$ object-freefall videos, processed at a resolution of $32 \times 256 \times 256$. For the Predictive Dynamics Alignment (PDA) training, we use a batch size of 64 and set the number of temporal bins to $K = 4$. Following [66], the representation alignment is applied at the 18th layer of the diffusion transformer using a LoRA [15] rank of 128. For GRPO training, we use a batch size of 8, a group size of $G = 16$, and 8 gradient accumulation steps. Candidate videos are generated using stochastic DDIM [41]. Classifier-free guidance is applied with scale w=6.0. Text prompt embeddings are pre-computed via T5 and cached to disk, avoiding redundant encoding across the G rollouts per prompt. The V-JEPA 2 critic processes each generated video with a sliding-window scheme. Within each window, the first

K=8 frames serve as encoder context under causal masking, and the predictor reconstructs tokens for the remaining future frames. The reward for a video is derived from the predictor's reconstruction quality across all windows, serving as a proxy for how well the generated dynamics conform to learned physical priors. The reward model is frozen throughout training. At each training step, a random fraction ρ=0.6 of the denoising timesteps is selected for gradient computation; the remaining steps are skipped to reduce memory cost. We use a maximum gradient norm of 1.0. Gradient checkpointing is enabled to fit the generation and update stages within GPU memory. The V-JEPA critic utilizes the V-JEPA 2-ViT-G [2] architecture, computing reward scores with causal masking based on 8 context frames, a window size of 16, and a temporal stride of 8. The model is trained for 1 epoch. For evaluations on VideoPhy and PisaBench, we use CogVideoX-5B [60] and Open-Sora [67] as our base models, respectively. We also include Wan2.2-T2V-A14B [48] as the base model in Fig. 5. All training is conducted on 8 NVIDIA GH200 (120GB) GPUs.

5.3 Baselines

We compare LDO against a comprehensive suite of strong text-to-video and image-to-video models. For general-purpose video generation, our baselines include Wan2.2 [48], MAGI-1 [45], VideoCrafter2 [12], Sora [34], Kling V1.5 [23], Runway Gen3 [38], DynamiCrafter [55], Pyramid-Flow [18], Open-Sora [67], DreamMachine [30], LaVIE [50], Cosmos-7B [1], and HunyuanVideo [22].

We also evaluate against physics-specific and domain-aligned models. On VideoPhy, we compare against PhyT2V [56], WISA [49], PhysMaster [17], and VideoREPA [66] (a model specialized in aligning diffusion representations with video self-supervised video models). On PisaBench, we compare against PISA's Supervised Fine-Tuning (PISA-psft) and Object Reward Optimization (PISA-oro) baselines. Specifically, PISA-oro uses the segmentation reward [24]. To ensure fair comparisons, we reproduce the results for VideoREPA and PISA models using the exact same training settings as our method. We utilize the detailed long prompts from [66] to evaluate CogVideoX, VideoREPA, and our LDO.

5.4 Does LDO improve world structure understanding?

We test if LDO captures physical structure rather than memorizing appearances by evaluate its handling of physical laws, causal interactions, and object permanence over time using complex state changes and rigid body dynamics.

Physical Commonsense and State Changes. An analysis of the VideoPhy benchmark (Table 1) shows a measurable shift in physical accuracy. Compared to the CogVideoX-5B base model, LDO achieves a +13.6 absolute increase in the overall Physical Commonsense (PC) score. The data shows that the model particularly improves in categories requiring complex state modeling, yielding +20.0 PC in Fluid-Fluid and +14.7 PC in Solid-Solid interactions. Furthermore, this structural grounding does not come at the expense of text-to-video fidelity, as Semantic Adherence (SA) also increases by +2.6.

Table 1: Quantitative evaluation on the VideoPhy benchmark. We compare our model LDO-5B against recent state-of-the-art video generation models and physics-specific baselines across different physical interaction categories. SA measures video-text alignment, while PC evaluates adherence to real-world physical laws. Colored subscripts indicate absolute performance changes relative to the CogVideoX-5B base. LDO-5B consistently achieves the best results, yielding a substantial +13.6 improvement in overall Physical Commonsense without compromising Semantic Adherence.

Methods	Solid–Solid		Solid–Fluid		Fluid–Fluid		Overall	
	SA ↑	PC ↑	SA ↑	PC ↑	SA ↑	PC ↑	SA ↑	PC ↑
Wan2.2 [48]	52.7	13.9	71.9	17.8	67.5	30.2	63.5	18.1
MAGI-1 [45]	42.2	19.0	67.2	27.7	51.9	33.5	54.4	25.0
VideoCrafter2 [12]	50.4	32.2	50.7	27.4	48.1	29.1	50.3	29.7
DreamMachine [30]	55.1	21.7	59.6	23.3	58.2	18.2	57.5	21.8
LaVIE [50]	40.8	18.3	48.6	37.0	69.1	50.9	48.7	31.5
Cosmos-7B [1]	-	-	-	-	-	-	57.0	18.0
HunyuanVideo [22]	55.2	16.1	67.1	30.1	54.5	54.5	60.2	28.2
PhyT2V [56]	47.0	32.0	61.0	30.0	67.0	62.0	59.0	42.0
WISA [49]	-	-	-	-	-	-	67.0	38.0
PhysMaster [17]	-	-	-	-	-	-	67.0	40.0
CogVideoX-5B [60]	62.2	19.6	75.3	33.6	70.9	60.0	69.2	32.0
VideoREPA-5B [66]	60.1	26.6	78.1	37.7	76.4	67.3	70.3	37.8
LDO-5B (ours)	**62.9** (+0.7)	**34.3** (+14.7)	**78.8** (+3.5)	**43.8** (+10.2)	**76.4** (+5.5)	**80.0** (+20.0)	**71.8** (+2.6)	**45.6** (+13.6)

Table 2: Quantitative evaluation on PisaBench. We compare our LDO against leading commercial models and physics-specific baselines on both Real and Simulated video sets. Performance is measured using L2 distance, Chamfer Distance (CD), and Intersection over Union (IoU). Colored percentages denote relative improvements over the PISA (psft) baseline. Our LDO achieves superior structural consistency and motion accuracy compared to both foundational and physics-tailored models.

Method	Real Videos			Simulated Videos		
	L2 ↓	CD ↓	IoU ↑	L2 ↓	CD ↓	IoU ↑
Sora [33]	0.174	0.488	0.065	0.145	0.433	0.036
Kling-V1.5 [23]	0.155	0.424	0.058	0.135	0.401	0.031
Runway Gen3 [38]	0.187	0.526	0.042	0.160	0.485	0.039
CogVideoX-5B-I2V [60]	0.138	0.366	0.080	0.107	0.303	0.020
DynamiCrafter [55]	0.187	0.504	0.021	0.147	0.458	0.036
Pyramid-Flow [18]	0.175	0.485	0.062	0.128	0.367	0.054
Open-Sora [67]	0.175	0.502	0.069	0.135	0.389	0.035
PISA-psft [24]	0.079	0.194	0.138	0.035	0.076	**0.155**
PISA-oro [24]	0.078	0.189	0.136	0.033	0.071	0.153
LDO (ours)	**0.070** (+11.39%)	**0.170** (+12.37%)	**0.145** (+5.07%)	**0.030** (+14.29%)	**0.064** (+15.79%)	0.154 (-0.65%)

Qualitatively (Figure 4 and 5), these gains correspond to more accurate modeling of causal interactions and support dynamics. In scenarios where baselines produce physically impossible outcomes—such as books sliding across a glass table while floating above the surface, or objects hovering mid-air without support—LDO maintains contact and gravity-consistent motion, keeping the books grounded on the tabletop and the apple yielding to gravity.

Trajectory, Shape Consistency, and Gravity. We further assess rigid body dynamics and gravitational physics using PisaBench (Table 2). The results demonstrate a reduction in trajectory and shape errors when LDO is compared with both baselines. Against the strong PISA (psft) baseline, LDO reduces Trajectory L2 error by 11.39% on real videos and 14.29% on simulated videos. Additionally, it improves CD by up to 15.79%, indicating that the generated objects better maintain their geometric boundaries throughout the sequence.

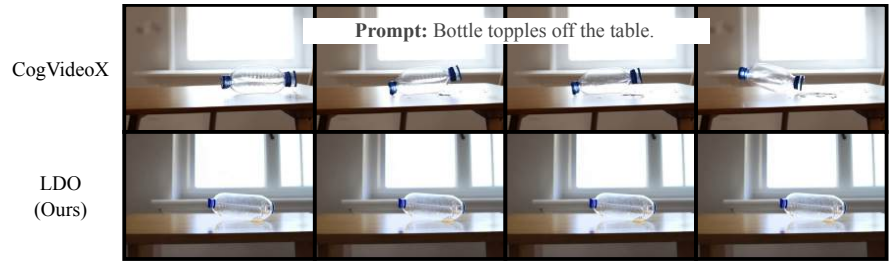

Fig. 7: Failure Case Analysis. Videos generated by LDO sometimes trade prompt adherence for physical realism. While CogVideoX animates a spontaneously toppling bottle (following the text but violating physics), LDO strictly enforces stability, generating a stationary bottle that avoids unphysical motion but ignores the prompt.

As shown in Figure 6, visually, these quantitative reductions in error correspond to stricter object permanence. While baseline models frequently struggle with structural degradation—causing falling objects to abruptly disappear or melt unnaturally into the environment—LDO preserves shape fidelity. When an object falls past a tabletop edge, LDO correctly anticipates the continued downward trajectory driven by gravity, whereas baselines incorrectly depict the object decelerating and landing on the table.

By combining these quantitative metrics and qualitative results, we find that LDO goes beyond surface-level pixel consistency. By distilling the causal predictive priors of the world model, LDO successfully enforces structural world dynamics and physical commonsense in the resulting video generation.

5.5 Ablation Studies

To validate the key design choices in LDO, we conduct ablation studies on the VideoPhy benchmark, focusing on the individual contributions of our proposed loss components.

Table 3: Ablation study on LDO's loss.

Loss Type	SA	PC
Vanilla diffusion	69.2	32.0
only PDA	**72.4**	39.2
only GRPO	70.6	40.4
All losses	71.8	**45.6**

Contribution of Loss Components. LDO integrates world model priors via two complementary pathways: structural feature alignment (PDA) and generation rollout optimization (GRPO). As shown in Table 3, each alone yields substantial improvements in physical commonsense over the vanilla diffusion baseline. PDA uniquely achieves the highest semantic adherence, while GRPO provides a slightly stronger physics prior. Together, their synergy results in the highest overall physical commonsense score with only a negligible trade-off in text alignment. This confirms that simultaneously aligning internal representations and optimizing generative rollouts is the most robust mechanism for enforcing physical laws.

5.6 Failure Case Analysis

While LDO significantly improves physics dynamics understanding, its strong physics prior can occasionally introduce a tension between text-prompt adher-

ence and physical realism. As illustrated in Figure 7, when given a prompt instructing a bottle to topple, the baseline CogVideoX model animates the bottle falling spontaneously without any external force, thereby aligning with the text but violating fundamental laws of inertia. In contrast, LDO recognizes that a stable object on a flat surface should not topple without an applied causal force. Consequently, it may over-optimize for structural stability and generate a completely stationary bottle. This failure mode highlights that while LDO successfully prevents unphysical, spontaneous events, its strict adherence to causal dynamics can result in overly conservative generations that fail to execute user instructions when explicit causal agents are missing from the text prompt.

6 Conclusion and Limitations

Bridging the gap between photorealistic video generation and *structurally correct* world dynamics remains a central challenge: modern diffusion models can produce sharp frames yet still violate support, motion continuity, and object permanence, yielding physically inconsistent videos. We introduced LDO, a post-training framework that uses a frozen predictive world model (V-JEPA2) as a *dynamics teacher and critic* to inject causal structure into pretrained video diffusion models *without changing their sampling procedure*. Across VideoPhy and PisaBench, LDO improves physical commonsense, trajectory fidelity, and object permanence while preserving generative quality—moving video generators closer to being not only visually realistic, but *physically grounded over time*.

Our LDO relies on a frozen predictive world model (V-JEPA 2) to provide both structural supervision and reward signals. As a result, the quality and coverage of the learned dynamics are ultimately bounded by the capabilities of the world model itself. If the world model fails to capture certain physical phenomena or domain-specific dynamics, these limitations may propagate to the diffusion model. However, our results show that even with a frozen teacher, the predictive structure encoded in the world model provides a useful supervisory signal that significantly improves the physical consistency of generated videos without degrading visual fidelity. LDO may sometimes prioritize physical plausibility over strictly following user instructions. Addressing this tension between causal plausibility and instruction following remains an interesting direction for future work, potentially through integrating language-aware reasoning or explicit causal signals into the generation process.

Acknowledgment Portion of this research has been supported by Vista GPU Cluster through the Center for Generative AI (CGAI) and the Texas Advanced Computing Center (TACC) at the University of Texas at Austin. UT Austin is supported by NSF Grants 2019844, 2112471, AF 1901292, CNS 2148141, Tripods CCF 1934932, Tripods 2217069, NSF AI Institute for Foundations of Machine Learning (IFML) 2019844, the Texas Advanced Computing Center (TACC) and research gifts by Western Digital, Wireless Networking and Communications Group (WNCG) Industrial Affiliates Program, UT Austin Machine Learning Lab (MLL), Cisco and the Stanly P. Finch Centennial Professorship in Engineering. We thank the anonymous reviewers for their constructive feedback.

References

1. Agarwal, N., Ali, A., Bala, M., Balaji, Y., Barker, E., Cai, T., Chattopadhyay, P., Chen, Y., Cui, Y., Ding, Y., et al.: Cosmos world foundation model platform for physical ai. arXiv preprint arXiv:2501.03575 (2025)
2. Assran, M., Bardes, A., Fan, D., Garrido, Q., Howes, R., Muckley, M., Rizvi, A., Roberts, C., Sinha, K., Zholus, A., et al.: V-jepa 2: Self-supervised video models enable understanding, prediction and planning. arXiv preprint arXiv:2506.09985 (2025)
3. Baillargeon, R.: Infants' physical world. Current Directions in Psychological Science **13**(3), 89–94 (2004)
4. Baillargeon, R.: Innate ideas revisited: For a principle of persistence in infants' physical reasoning. Perspectives on Psychological Science **3**(1), 2–13 (2008)
5. Bansal, H., Lin, Z., Xie, T., Zong, Z., Yarom, M., Bitton, Y., Jiang, C., Sun, Y., Chang, K.W., Grover, A.: Videophy: Evaluating physical commonsense for video generation. arXiv preprint arXiv:2406.03520 (2024)
6. Bardes, A., Garrido, Q., Ponce, J., Rabbat, M., LeCun, Y., Assran, M., Ballas, N.: Revisiting feature prediction for learning visual representations from video. arXiv:2404.08471 (2024)
7. Battaglia, P.W., Hamrick, J.B., Tenenbaum, J.B.: Simulation as an engine of physical scene understanding. Proceedings of the National Academy of Sciences **110**(45), 18327–18332 (2013)
8. Bear, D.M., Wang, E., Mrowca, D., Binder, F.J., Tung, H.Y.F., Pramod, R., Holdaway, C., Tao, S., Smith, K., Sun, F.Y., et al.: Physion: Evaluating physical prediction from vision in humans and machines. arXiv preprint arXiv:2106.08261 (2021)
9. Bhowmik, A., Korzhenkov, D., Snoek, C.G., Habibian, A., Ghafoorian, M.: Moalign: Motion-centric representation alignment for video diffusion models. arXiv preprint arXiv:2510.19022 (2025)
10. Bordes, F., Garrido, Q., Kao, J.T., Williams, A., Rabbat, M., Dupoux, E.: Intphys 2: Benchmarking intuitive physics understanding in complex synthetic environments. arXiv preprint arXiv:2506.09849 (2025)
11. Butt, M.A., Maragos, P.: Optimum design of chamfer distance transforms. IEEE Transactions on Image Processing **7**(10), 1477–1484 (1998)
12. Chen, H., Zhang, Y., Cun, X., Xia, M., Wang, X., Weng, C., Shan, Y.: Videocrafter2: Overcoming data limitations for high-quality video diffusion models (2024)
13. Garrido, Q., Ballas, N., Assran, M., Bardes, A., Najman, L., Rabbat, M., Dupoux, E., LeCun, Y.: Intuitive physics understanding emerges from self-supervised pretraining on natural videos. arXiv preprint arXiv:2502.11831 (2025)
14. Hamrick, J.B., Battaglia, P.W., Griffiths, T.L., Tenenbaum, J.B.: Inferring mass in complex scenes by mental simulation. Cognition **157**, 61–76 (2016)
15. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. Iclr **1**(2), 3 (2022)
16. Jang, S., Ki, T., Jo, J., Xie, S., Yoon, J., Hwang, S.J.: Self-refining video sampling. arXiv preprint arXiv:2601.18577 (2026)
17. Ji, S., Chen, X., Tao, X., Wan, P., Zhao, H.: Physmaster: Mastering physical representation for video generation via reinforcement learning. arXiv preprint arXiv:2510.13809 (2025)
18. Jin, Y., Sun, Z., Li, N., Xu, K., Jiang, H., Zhuang, N., Huang, Q., Song, Y., Mu, Y., Lin, Z.: Pyramidal flow matching for efficient video generative modeling. arXiv preprint arXiv:2410.05954 (2024)

19. Kang, B., Yue, Y., Lu, R., Lin, Z., Zhao, Y., Wang, K., Huang, G., Feng, J.: How far is video generation from world model: A physical law perspective. arXiv preprint arXiv:2411.02385 (2024)
20. Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., Amodei, D.: Scaling laws for neural language models. arXiv preprint arXiv:2001.08361 (2020)
21. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
22. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al.: Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603 (2024)
23. Kuaishou: Kling. https://kling.kuaishou.com (2024), accessed: 2024
24. Li, C., Michel, O., Pan, X., Liu, S., Roberts, M., Xie, S.: Pisa experiments: Exploring physics post-training for video diffusion models by watching stuff drop. arXiv preprint arXiv:2503.09595 (2025)
25. Li, Y., Wang, Y., Zhu, Y., Zhao, Z., Lu, M., She, Q., Zhang, S.: Branchgrpo: Stable and efficient grpo with structured branching in diffusion models. arXiv preprint arXiv:2509.06040 (2025)
26. Li, Z., Tucker, R., Snavely, N., Holynski, A.: Generative image dynamics. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 24142–24153 (2024)
27. Liu, J., Liu, G., Liang, J., Li, Y., Liu, J., Wang, X., Wan, P., Zhang, D., Ouyang, W.: Flow-grpo: Training flow matching models via online rl. arXiv preprint arXiv:2505.05470 (2025)
28. Liu, R., Wu, H., Zheng, Z., Wei, C., He, Y., Pi, R., Chen, Q.: Videodpo: Omni-preference alignment for video diffusion generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 8009–8019 (2025)
29. Liu, S., Ren, Z., Gupta, S., Wang, S.: Physgen: Rigid-body physics-grounded image-to-video generation. In: European Conference on Computer Vision. pp. 360–378. Springer (2024)
30. Luma: Dream machine. https://lumalabs.ai/dream-machine (2024), accessed: 2024
31. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008)
32. Nan, K., Xie, R., Zhou, P., Fan, T., Yang, Z., Chen, Z., Li, X., Yang, J., Tai, Y.: Openvid-1m: A large-scale high-quality dataset for text-to-video generation. arXiv preprint arXiv:2407.02371 (2024)
33. OpenAI: Sora. https://sora.com (2024), accessed: 2024
34. OpenAI: Sora 2 system card. https://cdn.openai.com/pdf/50d5973c-c4ff-4c2d-986f-c72b5d0ff069/sora_2_system_card.pdf (2025), accessed: January 21, 2026
35. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
36. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K.V., Carion, N., Wu, C.Y., Girshick, R., Dollár, P., Feichtenhofer, C.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
37. Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union: A metric and a loss for bounding box regression. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 658–666 (2019)

38. Runway: Gen-3 Alpha (2024)
39. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024)
40. Singh, J., Leng, X., Wu, Z., Zheng, L., Zhang, R., Shechtman, E., Xie, S.: What matters for representation alignment: Global information or spatial structure? arXiv preprint arXiv:2512.10794 (2025)
41. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
42. Spelke, E.S., Breinlinger, K., Macomber, J., Jacobson, K.: Origins of knowledge. Psychological Review **99**(4), 605 (1992)
43. Spelke, E.S., Kinzler, K.D.: Core knowledge. Developmental science **10**(1), 89–96 (2007)
44. Srivastava, N., Mansimov, E., Salakhudinov, R.: Unsupervised learning of video representations using lstms. In: International Conference on Machine Learning. pp. 843–852. PMLR (2015)
45. Teng, H., Jia, H., Sun, L., Li, L., Li, M., Tang, M., Han, S., Zhang, T., Zhang, W., Luo, W., et al.: Magi-1: Autoregressive video generation at scale. arXiv preprint arXiv:2505.13211 (2025)
46. Tishby, N., Pereira, F.C., Bialek, W.: The information bottleneck method. arXiv preprint physics/0004057 (2000)
47. Ullman, T.D., Spelke, E., Battaglia, P., Tenenbaum, J.B.: Mind games: Game engines as an architecture for intuitive physics. Trends in Cognitive Sciences **21**(9), 649–665 (2017)
48. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
49. Wang, J., Ma, A., Cao, K., Zheng, J., Zhang, Z., Feng, J., Liu, S., Ma, Y., Cheng, B., Leng, D., et al.: Wisa: World simulator assistant for physics-aware text-to-video generation. arXiv preprint arXiv:2503.08153 (2025)
50. Wang, Y., Chen, X., Ma, X., Zhou, S., Huang, Z., Wang, Y., Yang, C., He, Y., Yu, J., Yang, P., et al.: Lavie: High-quality video generation with cascaded latent diffusion models. International Journal of Computer Vision **133**(5), 3059–3078 (2025)
51. Wu, H., Wu, D., He, T., Guo, J., Ye, Y., Duan, Y., Bian, J.: Geometry forcing: Marrying video diffusion and 3d representation for consistent world modeling. arXiv preprint arXiv:2507.07982 (2025)
52. Wu, J., Yildirim, I., Lim, J.J., Freeman, B., Tenenbaum, J.: Galileo: Perceiving physical object properties by integrating a physics engine with deep learning. Advances in Neural Information Processing Systems **28** (2015)
53. Wu, J., Lu, E., Kohli, P., Freeman, B., Tenenbaum, J.: Learning to see physics via visual de-animation. Advances in neural information processing systems **30** (2017)
54. Wu, Z., Kag, A., Skorokhodov, I., Menapace, W., Mirzaei, A., Gilitschenski, I., Tulyakov, S., Siarohin, A.: Densedpo: Fine-grained temporal preference optimization for video diffusion models. arXiv preprint arXiv:2506.03517 (2025)
55. Xing, J., Xia, M., Zhang, Y., Chen, H., Yu, W., Liu, H., Wang, X., Wong, T.T., Shan, Y.: Dynamicrafter: Animating open-domain images with video diffusion priors. arXiv preprint arXiv:2310.12190 (2023)
56. Xue, Q., Yin, X., Yang, B., Gao, W.: Phyt2v: Llm-guided iterative self-refinement for physics-grounded text-to-video generation. arXiv preprint arXiv:2412.00596 (2024)

57. Xue, T., Wu, J., Bouman, K., Freeman, B.: Visual dynamics: Probabilistic future frame synthesis via cross convolutional networks. Advances in Neural Information Processing Systems **29** (2016)
58. Xue, Z., Wu, J., Gao, Y., Kong, F., Zhu, L., Chen, M., Liu, Z., Liu, W., Guo, Q., Huang, W., et al.: Dancegrpo: Unleashing grpo on visual generation. arXiv preprint arXiv:2505.07818 (2025)
59. Yang, X., Li, B., Zhang, Y., Yin, Z., Bai, L., Ma, L., Wang, Z., Cai, J., Wong, T.T., Lu, H., et al.: Vlipp: Towards physically plausible video generation with vision and language informed physical prior. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12360–12370 (2025)
60. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072 (2024)
61. Yu, S., Kwak, S., Jang, H., Jeong, J., Huang, J., Shin, J., Xie, S.: Representation alignment for generation: Training diffusion transformers is easier than you think. arXiv preprint arXiv:2410.06940 (2024)
62. Yuan, J., Pizzati, F., Pinto, F., Kunze, L., Laptev, I., Newman, P., Torr, P., De Martini, D.: Likephys: Evaluating intuitive physics understanding in video diffusion models via likelihood preference. arXiv preprint arXiv:2510.11512 (2025)
63. Yuan, J., Zhang, X., Friedrich, F., Beltran-Velez, N., Hall, M., Askari-Hemmat, R., Han, X., Ballas, N., Drozdzal, M., Romero-Soriano, A.: Inference-time physics alignment of video generative models with latent world models. arXiv preprint arXiv:2601.10553 (2026)
64. Yuan, Y., Wang, X., Wickremasinghe, T., Nadir, Z., Ma, B., Chan, S.H.: Newtongen: Physics-consistent and controllable text-to-video generation via neural newtonian dynamics. arXiv preprint arXiv:2509.21309 (2025)
65. Zhan, G., Ma, X., Xie, W., Zisserman, A.: Inferring dynamic physical properties from video foundation models. arXiv preprint arXiv:2510.02311 (2025)
66. Zhang, X., Liao, J., Zhang, S., Meng, F., Wan, X., Yan, J., Cheng, Y.: Videorepa: Learning physics for video generation through relational alignment with foundation models. arXiv preprint arXiv:2505.23656 (2025)
67. Zheng, Z., Peng, X., Yang, T., Shen, C., Li, S., Liu, H., Zhou, Y., Li, T., You, Y.: Open-sora: Democratizing efficient video production for all. arXiv preprint arXiv:2412.20404 (2024)