CoDance for Physically Consistent Image-to-Video Generation via Coupled Reinforcement Learning

Anonymous authors
Paper under double-blind review

Abstract

Image-to-video models can produce visually realistic videos while violating the physical constraints governing motion and interactions, limiting their utility for embodied prediction and planning. Some recent attempts employ physical reasoning to refine the user prompt before feeding it into the video generation model, making the intended dynamics and interactions more explicit. However, reasoning over the prompt alone does not ensure that the generator can faithfully realize these inferred dynamics. This motivates jointly optimizing reasoning and generation to address the mismatch between inferred physical dynamics and their realization in generated videos. We introduce CoDance with improved physical quality by jointly optimizing a vision-language reasoner and a flow-based video generator with reinforcement learning. The resulting training pipeline, dubbed CoDanceRL, combines three parts: Coupled Policy Optimization, Physics-Aware Feedback, and Hierarchical Reward Normalization. CoDance obtains an IQ-Score of 45.4 on Physics-IQ Verified and an average embodied video generation score of 69.4 on RBench, compared with 34.8 and 58.1 for its Cosmos3-Nano base model. It also achieves an average closed-loop action-planning success rate of 41.5% across two WorldArena tasks, demonstrating its utility for downstream planning.

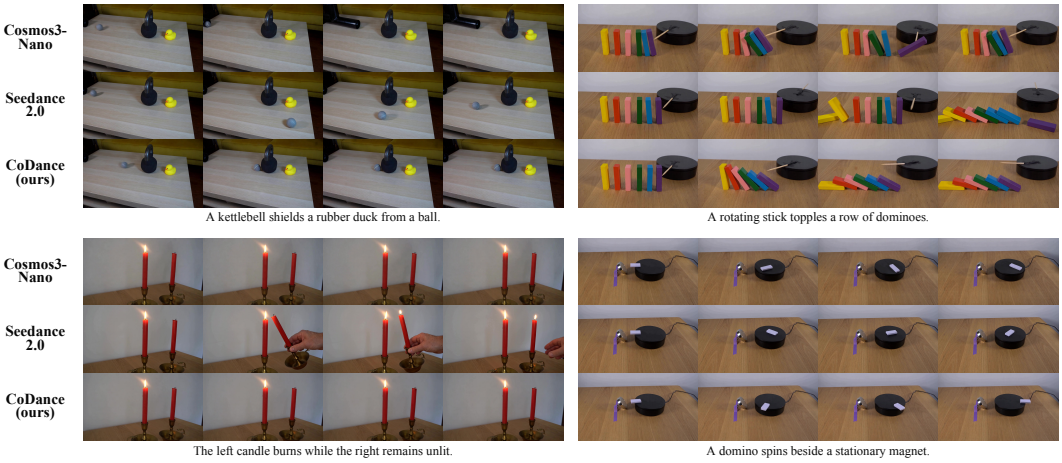


Figure 1: **Qualitative comparison on Physics-IQ Verified.** Within each example, columns show frames in temporal order, while rows compare Cosmos3-Nano (NVIDIA et al., 2026), Seedance 2.0 (Seedance et al., 2026), and CoDance. CoDance better preserves the intended physical interactions and scene constraints.

1 Introduction

Recent advances in image-to-video (I2V) generation have enabled increasingly realistic visual continuations of an observed scene (Seedance et al., 2026; NVIDIA et al., 2026). Beyond con-

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

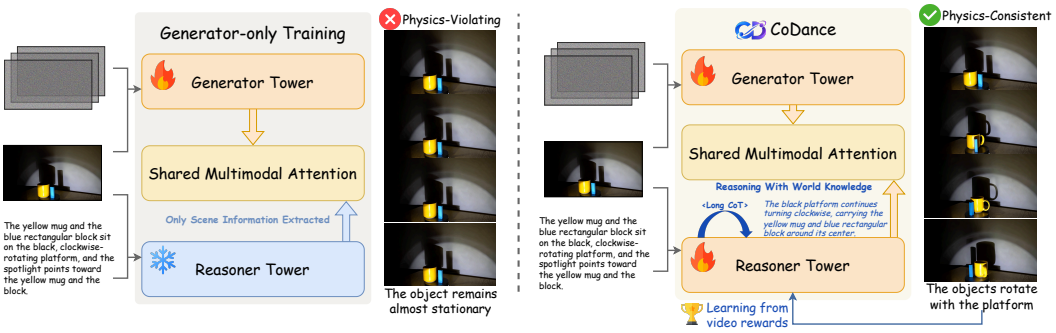


Figure 2: **Generator-only training versus CoDance.** In this example, generator-only training with a frozen reasoner leaves the objects nearly stationary on a rotating platform (left), whereas CoDance jointly trains the reasoner and generator with video-level feedback to produce the corresponding object motion (right). Flame and snowflake denote trainable and frozen towers.

tent creation, generated videos can represent candidate futures for embodied prediction and planning (Du et al., 2023; Liang et al., 2025). These applications require physically plausible motion and interactions, yet visually realistic videos can still violate physical constraints (Motamed et al., 2025), limiting their utility in embodied settings.

Existing work improves physical consistency through physics-informed supervision (Zhang et al., 2026) or reward- and preference-based adaptation of video generators (Cai et al., 2026; Le et al., 2025). Other approaches guide pretrained generators with physical reasoning. Training-free methods refine prompts or construct keyframes and trajectories at inference time (Liu et al., 2025; Zhuang et al., 2026), while PhyPrompt learns a prompt rewriter from video-level physical feedback with the generator frozen (Wu et al., 2026).

Despite these advances, generated videos can depict the intended event while violating the physical relationships that govern its evolution (Bansal et al., 2025). A physically plausible continuation must also capture the causal progression of interactions (Wang et al., 2026b). In I2V generation, the initial image and prompt specify the scene and intended event but may leave these interactions and their consequences implicit.

Physical reasoning can make these relationships explicit, but its usefulness depends on how the generator realizes the guidance. Prior work connects understanding and generation. For example, VANS coordinates a vision-language model and a video generator through sequential two-stage optimization for next-event prediction (Cheng et al., 2026), while PhysMaster learns image-conditioned physical representations through staged fine-tuning and preference optimization (Ji et al., 2025). We introduce a coupled RL paradigm for physically consistent I2V generation in which the language-based physical reasoner and video generator remain trainable throughout the same RL process and are updated from shared feedback on each nested rollout. The reasoner learns to provide guidance that benefits generation, while the generator learns to realize that guidance as physically consistent motion.

Building on this insight, we introduce **CoDance**, an I2V model built on Cosmos3-Nano (NVIDIA et al., 2026) for physically consistent video generation. Given the initial image and user prompt, its vision-language reasoner infers the scene's physical dynamics and rewrites the prompt accordingly, with the resulting representations conditioning the flow-based video generator through shared multimodal attention, as shown in Figure 2. To develop these capabilities, we train CoDance with **CoDanceRL**, a coupled reinforcement-learning framework with three components. *Coupled Policy Optimization* alternates updates to the reasoner and generator using shared feedback from generated videos. *Physics-Aware Feedback* compares generated videos with real references through complementary physical-fidelity and geometric trajectory rewards. *Hierarchical Reward Normalization* constructs prompt- and video-level advantages by separately normalizing these rewards within the nested rollout hierarchy.

Empirically, we evaluate CoDance on physical I2V generation, embodied video generation, and downstream closed-loop action planning. On Physics-IQ Verified (Rädsch et al., 2026), CoDance improves the IQ-Score of Cosmos3-Nano from 34.8 to 45.4. The held-out VLM-judge score, which

is excluded from reward optimization, also rises from 31.7 to 52.1. On RBench (Deng et al., 2026), CoDance achieves an average score of 69.4, compared with 58.1 for the base model and 63.1 for its SFT initialization. For closed-loop planning with frozen video-model features, the average success rate across two WorldArena (Shang et al., 2026b) tasks increases from 14.0% to 41.5%. Joint-training comparisons and inference-time component recombinations in Appendix A.6.2 further support the benefit of adapting both the reasoner and generator.

2 PRELIMINARIES

Flow-based image-to-video generation. Image-to-video models generate videos from reference images (Zhang et al., 2023; Blattmann et al., 2023). In this work, we consider a model conditioned on both a reference image I and a text prompt p, with $c = (I, p)$. Let z_0 denote a clean video latent and $\xi \sim \mathcal{N}(0, \mathbf{I})$ denote Gaussian noise. Under a linear flow-matching path (Lipman et al., 2023), the intermediate latent is defined as $z_t = (1-t)z_0 + t\xi, t \in [0, 1]$. A DiT (Peebles & Xie, 2023) parameterizes a conditional velocity field $v_\theta(z_t, t, c)$, which is trained to predict the target velocity $\xi - z_0$ by minimizing

$$\mathcal{L}_{\mathrm{FM}}(\theta) = \mathbb{E}_{z_0,c,t,\xi}\left[\|v_\theta(z_t, t, c) - (\xi - z_0)\|_2^2\right]. \tag{1}$$

Generation starts from Gaussian noise at $t = 1$ and integrates the learned velocity field toward $t = 0$, after which the resulting latent is decoded into a video (Lipman et al., 2023).

Response-Level GRPO. We use a response-level variant of Group Relative Policy Optimization (GRPO) (Shao et al., 2024), which estimates advantages from groups of sampled outputs. Given a condition c, a group of prompt rewrites $\{p_i\}_{i=1}^G$ is sampled from the old policy $\pi_{\phi_{\mathrm{old}}}(\cdot \mid c)$. For reward scores $\{R_i\}_{i=1}^G$, the group-relative advantage is $\hat{A}_i = (R_i - \mu_R)/(\sigma_R + \epsilon_{\mathrm{norm}})$, where μ_R and σ_R are the group mean and standard deviation, and $\epsilon_{\mathrm{norm}} > 0$ ensures numerical stability. The policy is optimized by maximizing

$$\mathcal{J}_{\mathrm{GRPO}}(\phi) = \mathbb{E}_{\{p_i\}_{i=1}^G \sim \pi_{\phi_{\mathrm{old}}}} \left[\frac{1}{G}\sum_{i=1}^{G} \min\Big(\rho_i(\phi)\hat{A}_i, \mathrm{clip}(\rho_i(\phi), 1-\epsilon, 1+\epsilon)\,\hat{A}_i\Big) - \beta D_{\mathrm{KL}}[\pi_\phi \,\|\, \pi_{\phi_{\mathrm{ref}}}]\right], \tag{2}$$

where $\rho_i(\phi) = \frac{\pi_\phi(p_i|c)}{\pi_{\phi_{\mathrm{old}}}(p_i|c)}$ is the probability ratio for the full response, and $\pi_{\phi_{\mathrm{ref}}}$ is the fixed reference policy. Here, ϵ is the clipping threshold and β is the KL coefficient.

DanceGRPO for Flow Matching Diffusion Models. DanceGRPO (Xue et al., 2025b) extends GRPO to visual generation by treating the stochastic sampling process as a sequence of policy transitions. For a flow-based video model, we discretize the reverse-time dynamics with a stochastic sampler and denote the transition density at sampling step n for the j-th generated video by $p_\theta(z_{j,n-1} \mid z_{j,n}, c)$, where c is the generation condition. Here, $n = N, \ldots, 1$ indexes the N discrete sampling steps from Gaussian noise $z_{j,N} = \xi_j$ to the final latent $z_{j,0}$. The corresponding probability ratio is $\rho_{j,n}(\theta) = \frac{p_\theta(z_{j,n-1}|z_{j,n},c)}{p_{\theta_{\mathrm{old}}}(z_{j,n-1}|z_{j,n},c)}$. Rewards on the final videos provide group-relative advantages for the sampled trajectories. DanceGRPO applies the clipped surrogate in Eq. 2 to these transitions, averaging over videos and sampling steps and omitting the KL term in its original formulation.

3 CODANCERL FOR PHYSICALLY CONSISTENT VIDEO GENERATION

This section presents CoDanceRL, the coupled post-training paradigm illustrated in Figure 3. Section 3.1 describes the rollout construction and coupled policy optimizations for the reasoner and generator. Section 3.2 introduces complementary physical-fidelity and trajectory rewards for evaluating generated dynamics. Finally, Section 3.3 explains how hierarchical reward normalization produces video-level advantages for the generator and prompt-level advantages for the reasoner.

3.1 COUPLED POLICY OPTIMIZATION

Before reinforcement learning, we perform supervised fine-tuning (SFT) to activate the reasoner's physical reasoning capabilities and the generator's understanding of physical dynamics, as detailed

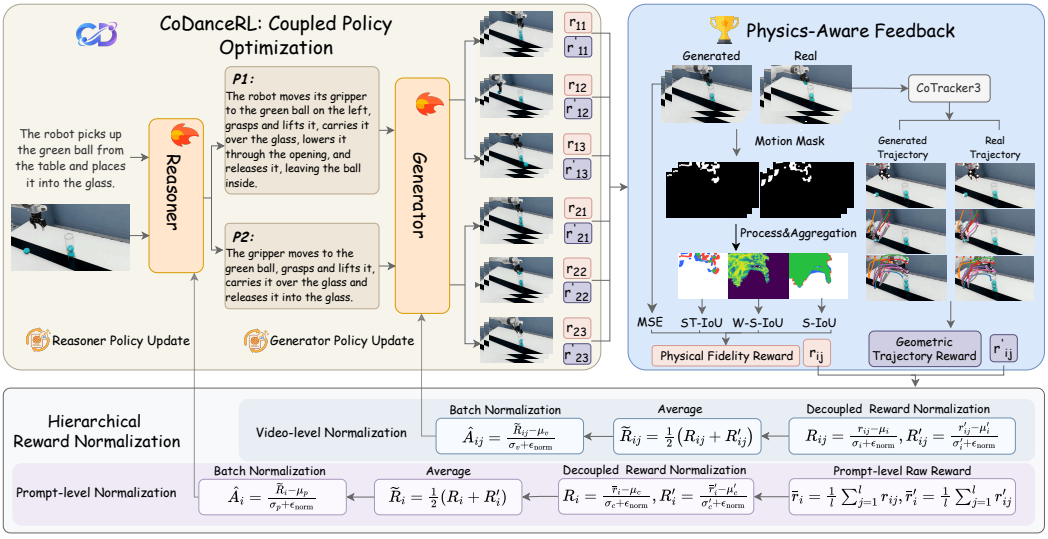


Figure 3: **Overview of CoDanceRL.** The reasoner samples prompt rewrites, and the generator produces multiple video rollouts for each rewrite. Complementary physical-fidelity and trajectory rewards are hierarchically normalized into video- and prompt-level advantages that update the generator and reasoner, respectively.

in Appendix A.5. Starting from this initialization, CoDanceRL jointly optimizes the two components using feedback on the physical consistency of generated videos. Unlike a sequential two-stage schedule that completes reasoner training before generator training, CoDanceRL keeps both policies trainable throughout a single RL run. Each nested rollout supplies advantages for both policies and is followed by an update to each.

During rollout collection, the reasoner and generator are fixed at parameter snapshots ϕ_{old} and θ_{old}, respectively. Given c, the reasoner samples G physically explicit prompt rewrites $\{p_i\}_{i=1}^{G}$ from $\pi_{\phi_{\text{old}}}(\cdot \mid c)$. Encoded by the language pathway, each rewrite conditions video generation through joint attention, defining $c_i = (I, p_i)$.

For each c_i, the generator samples l latent trajectories, with $z_{ij,n}$ denoting the latent of the j-th video at step n. Starting from Gaussian noise $z_{ij,N} = \xi_{ij}$, we sample $z_{ij,n-1} \sim p_{\theta_{\text{old}}}(\cdot \mid z_{ij,n}, c_i)$ for $n = N, \ldots, 1$. Decoding $z_{ij,0}$ produces the video V_{ij}, giving Gl videos in total.

We use the physical consistency of these videos to guide updates to both the generator and the reasoner. Each V_{ij} receives the physical-fidelity reward r_{ij} and geometric trajectory reward r'_{ij} defined in Section 3.2. The hierarchical normalization in Section 3.3 then converts these rewards into a video-level advantage $\hat{A}_{ij}$ for the generator and a prompt-level advantage $\hat{A}_i$ for the reasoner.

Both objectives take expectations over these nested rollouts for a given c. We maximize the reasoner's response-level GRPO objective

$$\mathcal{J}_{\text{R}}(\phi) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\min\Big(\rho_i(\phi)\hat{A}_i, \text{clip}(\rho_i(\phi), 1-\epsilon, 1+\epsilon)\,\hat{A}_i\Big) - \beta D_{\text{KL}}[\pi_\phi(\cdot \mid c) \,\|\, \pi_{\phi_{\text{ref}}}(\cdot \mid c)]\right]. \tag{3}$$

At the video level, we use $\rho_{ij,n}(\theta) = \frac{p_\theta(z_{ij,n-1}|z_{ij,n},c_i)}{p_{\theta_{\text{old}}}(z_{ij,n-1}|z_{ij,n},c_i)}$ at sampling step n. All steps for V_{ij} share the video-level advantage $\hat{A}_{ij}$, giving the generator objective

$$\mathcal{J}_{\text{G}}(\theta) = \mathbb{E}\left[\frac{1}{GlN}\sum_{i=1}^{G}\sum_{j=1}^{l}\sum_{n=1}^{N}\min\Big(\rho_{ij,n}(\theta)\hat{A}_{ij}, \text{clip}(\rho_{ij,n}(\theta), 1-\epsilon, 1+\epsilon)\,\hat{A}_{ij}\Big)\right]. \tag{4}$$

As in the original DanceGRPO formulation (Xue et al., 2025b), $\mathcal{J}_{\mathrm{G}}$ does not include a KL term. CoDanceRL alternates updates that maximize $\mathcal{J}_{\mathrm{R}}$ and $\mathcal{J}_{\mathrm{G}}$. Both updates use feedback from the same rollout. The prompt-level advantage $\hat{A}_i$ evaluates a rewrite through its generated videos, while $\hat{A}_{ij}$ evaluates an individual generation trajectory.

3.2 Physics-Aware Feedback

The usefulness of a prompt rewrite depends on whether the generator realizes its physical guidance in the sampled videos. The coupled updates above therefore need feedback on what each rollout actually depicts. For each generated video V_{ij} and real reference $V^\star$, the physical-fidelity reward r_{ij} measures where and when motion occurs and how well the video matches the reference, while the geometric trajectory reward r'_{ij} compares the relative positions, approach times, and final displacements of tracked points.

Physical-fidelity reward. We adapt the motion-overlap and reconstruction metrics of Physics-IQ Verified (Rädsch et al., 2026) to compare each generated video with a single real reference. The spatial (SP), spatiotemporal (ST), and weighted spatial (WS) IoU terms compare where motion occurs, whether it overlaps at the same time, and how frequently regions are active, respectively. MSE measures reconstruction error between videos with RGB values scaled to $[0, 1]$. For each metric M, let $m_{ij}^M = M(V_{ij}, V^\star)$. With larger IoU and smaller MSE indicating better agreement, we compute the raw physical-fidelity reward as

$$r_{ij} = \frac{1}{4}\left[1 - m_{ij}^{\mathrm{MSE}} + \sum_{M\in\{\mathrm{SP},\mathrm{ST},\mathrm{WS}\}} m_{ij}^M\right]. \tag{5}$$

The four terms lie in $[0, 1]$ and are averaged with equal weights. We compute all metrics on spatially and temporally aligned videos, excluding the conditioning frame.

Geometric trajectory reward. Video-level motion overlap does not explicitly compare how corresponding points move relative to one another. Two videos can activate similar image regions while tracked points approach one another at different times. We therefore use CoTracker3 (Karaev et al., 2024) to track the same grid from the conditioning image in V_{ij} and $V^\star$. The geometry score m_{ij}^{geom} compares pairwise distances and emphasizes differences when points approach. The timing score m_{ij}^{time} compares when pairs reach their closest separation, and the final score m_{ij}^{final} compares moving points' displacements after removing global translation. We average these scores to obtain

$$r'_{ij} = \frac{m_{ij}^{\mathrm{geom}} + m_{ij}^{\mathrm{time}} + m_{ij}^{\mathrm{final}}}{3}. \tag{6}$$

Appendix A.2 gives the point selection, visibility rules, normalization, and component formulas. Section 3.3 converts both rewards into policy-specific advantages.

3.3 Hierarchical Reward Normalization

The physical rewards evaluate individual videos, while coupled optimization also requires feedback on their prompt rewrites. Following the reward-decoupling principle of Group reward-Decoupled Normalization Policy Optimization (GDPO) (Liu et al., 2026), we normalize each reward family separately before aggregation, then normalize the aggregate over the batch. We apply this design at both levels of the nested rollout hierarchy. Video-level advantages compare videos under a fixed rewrite, while prompt-level advantages compare rewrites through their mean video rewards. Appendix A.6.3 provides a combinatorial motivation based on distinct advantage-group counts, while Section 4.3 evaluates the effect of normalization on IQ-Score.

Video-level advantage. We compare the l videos generated under rewrite p_i, holding their generation condition c_i fixed. Let (μ_i, σ_i) and (μ'_i, σ'_i) be the mean and standard deviation of $\{r_{ij}\}_{j=1}^{l}$ and $\{r'_{ij}\}_{j=1}^{l}$, respectively. We normalize each reward family within this group using the stabilizer $\epsilon_{\mathrm{norm}} > 0$:

$$R_{ij} = \frac{r_{ij} - \mu_i}{\sigma_i + \epsilon_{\mathrm{norm}}}, \qquad R'_{ij} = \frac{r'_{ij} - \mu'_i}{\sigma'_i + \epsilon_{\mathrm{norm}}}. \tag{7}$$

We average the two normalized rewards with equal weights, then normalize the combined signal over the full video batch to obtain advantages on a common scale

$$\widetilde{R}_{ij} = \tfrac{1}{2}(R_{ij} + R'_{ij}), \qquad \hat{A}_{ij} = \frac{\widetilde{R}_{ij} - \mu_v}{\sigma_v + \epsilon_{\text{norm}}}, \tag{8}$$

where μ_v and σ_v are the batch mean and standard deviation of $\widetilde{R}_{ij}$. The advantage $\hat{A}_{ij}$ measures each video's relative physical rewards within its rewrite group.

Prompt-level advantage. The reasoner evaluates each rewrite through the average physical rewards of its generated videos. These averages must be computed from the raw rewards because video-level normalization centers each reward family within a rewrite. Averaging the normalized values would therefore erase differences in mean reward between rewrites. For each rewrite p_i, we first compute $\bar{r}_i = \frac{1}{l}\sum_{j=1}^{l} r_{ij}$ and $\bar{r}'_i = \frac{1}{l}\sum_{j=1}^{l} r'_{ij}$. For the G rewrites of the same initial condition c, let (μ_c, σ_c) and (μ'_c, σ'_c) be the mean and standard deviation of $\{\bar{r}_i\}_{i=1}^{G}$ and $\{\bar{r}'_i\}_{i=1}^{G}$, respectively. We normalize these averages within the group of rewrites as

$$R_i = \frac{\bar{r}_i - \mu_c}{\sigma_c + \epsilon_{\text{norm}}}, \qquad R'_i = \frac{\bar{r}'_i - \mu'_c}{\sigma'_c + \epsilon_{\text{norm}}}. \tag{9}$$

We again combine the normalized reward families with equal weights and normalize over the full prompt batch to obtain the reasoner's advantage

$$\widetilde{R}_i = \tfrac{1}{2}(R_i + R'_i), \qquad \hat{A}_i = \frac{\widetilde{R}_i - \mu_p}{\sigma_p + \epsilon_{\text{norm}}}, \tag{10}$$

where μ_p and σ_p are the batch mean and standard deviation of $\widetilde{R}_i$. The resulting $\hat{A}_i$ evaluates p_i relative to the other rewrites of c using their average physical rewards.

4 Experiments

We evaluate CoDance on physically consistent video generation, robotic interaction video generation, and downstream closed-loop action planning using Physics-IQ Verified (Rädsch et al., 2026), RBench (Deng et al., 2026), and WorldArena (Shang et al., 2026b), respectively. We compare CoDance with commercial and open-source image-to-video models (Seedance et al., 2026; Wu et al., 2025a; NVIDIA et al., 2026). Ablations examine the physical rewards and hierarchical reward normalization, with additional training-strategy and component comparisons in Appendix A.6.2. Further experimental details and results are provided in Appendices A.4 and A.6.

4.1 Physically Plausible Image to Video Generation

We first evaluate physical consistency on Physics-IQ Verified (Table 1). CoDance achieves the highest aggregate IQ-Score of 45.4 among the compared models, exceeding PhiZero, the strongest external baseline, by 4.8 points. It improves over the Cosmos3-Nano base model by 10.6 points and the SFT initialization by 7.0 points, with gains in all three IoU metrics over both. The best individual metrics vary across models: Cosmos3-Super I2V has the highest ST-IoU, and PhiZero has the highest reverse-normalized MSE score.

Because our physical-fidelity reward incorporates the Physics-IQ Verified metrics, we additionally report a held-out VLM-judge score that is not used for reward optimization. CoDance achieves 52.1, compared with 31.7 for Cosmos3-Nano, 42.3 for its SFT initialization, and 41.9 for Seedance 2.0. These results show that the gains also extend to a metric excluded from the training reward. Evaluation details are provided in Appendix A.4.

Figure 1 illustrates how these differences appear in generated interactions and scene states. In the domino example, CoDance depicts the progression of falling dominoes as the rotating stick interacts with the row. In the candle example, it keeps the right candle unlit throughout the sequence, whereas Seedance 2.0 introduces a hand that lights it. These examples illustrate the evolution of an intended interaction and the preservation of scene constraints that should hold over time.

Table 1: Comparison on Physics-IQ Verified (Rädsch et al., 2026) and a held-out VLM judge. MSE is reported as a reverse-normalized score, so higher is better. Bold and underlined values mark the best and second-best results, respectively.

Model	Physics-IQ Verified					Held-out
	S-IoU	ST-IoU	W-S-IoU	MSE	IQ-Score	VLM-Judge
Commercial						
Sora 2 (OpenAI, 2025)	37.3	27.0	26.9	14.8	26.5	–
Grok-Imagine-Video (xAI, 2026)	52.7	21.4	35.7	29.6	34.8	–
MiniMax-H3 Max (Team fal, 2026)	54.9	38.2	24.5	27.2	36.2	-
Seedance 2.0 (Seedance et al., 2026)	57.3	33.0	<u>41.6</u>	24.6	39.1	41.9
Open-source						
Wan2.2-I2V-A14B (Wan et al., 2025)	51.1	20.5	28.5	28.9	32.2	33.2
HunyuanVideo 1.5 (Wu et al., 2025a)	47.1	26.9	29.7	30.0	33.4	26.70
Cosmos3-Super I2V (NVIDIA et al., 2026)	46.4	**49.3**	34.7	27.8	39.5	36.4
Physics-aware methods						
Magi-1 24B+GeoPhys (Internò et al., 2026)	–	–	–	–	33.7	–
PhiZero (Shang et al., 2026a)	<u>57.4</u>	36.3	26.8	**41.7**	<u>40.6</u>	33.9
Ours and base model						
Cosmos3-Nano (NVIDIA et al., 2026)	45.2	26.5	31.3	<u>36.3</u>	34.8	31.7
Cosmos3-Nano + SFT	51.4	33.3	36.1	32.9	38.4	<u>42.3</u>
CoDance	**59.0**	<u>44.2</u>	**45.2**	33.3	**45.4**	**52.1**

4.2 Embodied Video Generation and Planning

We next evaluate robotic interaction videos on RBench, covering five task categories and four embodiments. CoDance averages 69.4, compared with 58.1 for the base model and 63.1 for its SFT initialization (Table 2). The gains over SFT are concentrated in the embodiment categories: their unweighted mean rises by 13.20 points, compared with 0.74 points across the five task categories. All four embodiment scores improve, while task-level changes are mixed. Spatial and reasoning scores increase, whereas manipulation, multi-entity, and long-horizon scores decline.

Figure 4 illustrates the correspondence between actions and object states. CoDance shows the lamp turning off as the gripper presses its button, and the bottle returning toward an upright pose after tilting over the glass. These examples show how the depicted interactions progress over time.

To evaluate downstream utility, we use the WorldArena `adjust_bottle` and `click_bell` tasks with frozen video-model features and task-specific inverse dynamics models. Across 100 rollouts per task, CoDance increases success from 15.0% to 34.0% and from 13.0% to 49.0%, respectively, relative to the base model (Table 3). The mean success rate rises from 14.0% to 41.5%, also exceeding the SFT initialization's 22.5%. Figure 5 illustrates the corresponding improvement in task completion on `adjust_bottle`.

4.3 Ablation and Analysis

Table 4 evaluates the two physical rewards and the video- and prompt-level components of hierarchical reward normalization on Physics-IQ Verified.

Reward components. Without reward-decoupled normalization, the physical-fidelity and geometric trajectory rewards alone yield IQ-Scores of 42.8 and 40.1, respectively. Combining them reduces the score to 41.3 relative to the physical-fidelity reward alone. During joint optimization of the unnormalized objectives, we observe reward hacking as policy updates over-optimize the geometric trajectory reward, destabilizing training and reducing IQ-Score.

Hierarchical reward normalization. With both rewards enabled, video-level normalization raises IQ-Score from 41.3 to 43.9, while prompt-level normalization raises it to 43.3. Enabling both levels yields the highest score of 45.4, exceeding video-only and prompt-only normalization

Table 2: Comparison of embodied video generation on RBench (Deng et al., 2026) across task types and embodiments. Avg. denotes the unweighted mean across the five task and four embodiment categories. The best and second-best results in each column are shown in bold and underlined, respectively.

Model	Tasks					Embodiments				Avg.
	Manip.	Spatial	Multi-entity	Long-horizon	Reasoning	Single-arm	Dual-arm	Quadruped	Humanoid	
Open-source										
Wan2.2-TI2V-5B (Wan et al., 2025)	33.1	31.3	14.2	31.8	23.4	43.6	44.8	59.0	60.7	38.0
LTX-2 (HaCohen et al., 2026)	28.4	30.4	23.3	38.6	16.4	45.3	42.4	62.2	55.5	38.1
LongCat-Video (Team et al., 2025)	37.2	31.0	22.0	38.4	18.6	58.6	57.6	68.1	62.1	43.7
HunyuanVideo 1.5 (Wu et al., 2025a)	44.2	31.6	31.2	43.8	36.4	51.3	52.6	63.4	59.5	46.0
Wan2.2-I2V-A14B (Wan et al., 2025)	38.1	45.4	37.3	50.1	33.0	60.8	58.2	69.0	64.8	50.7
Commercial										
Sora 2 (OpenAI, 2025)	20.8	26.8	18.6	25.5	11.5	47.6	51.3	66.4	56.1	36.1
Kling 2.6 (Kuaishou Technology, 2025)	52.9	59.8	36.4	53.0	35.8	57.0	60.5	63.7	61.3	53.4
Hailuo v2 (MiniMax, 2025)	56.0	63.7	38.6	54.5	47.4	59.4	61.1	64.0	63.5	56.5
Seedance 1.5 (Seedance et al., 2025)	57.7	49.5	48.4	57.0	47.0	64.8	64.1	68.0	69.2	58.4
Veo 3.1 (Google DeepMind, 2025)	54.1	47.4	53.4	59.2	46.7	67.0	66.6	74.3	70.4	59.9
Wan2.6 (Alibaba Cloud, 2025)	54.6	65.6	47.9	51.4	53.1	66.6	68.1	72.3	66.7	60.7
Robotics-specific										
UnifoLM-WMA-0 (Unitree, 2025)	3.6	4.0	1.8	6.2	0.0	26.8	19.4	29.3	20.0	12.3
Vidar (Feng et al., 2025)	7.3	10.6	5.0	5.4	5.0	38.2	41.0	37.4	35.7	20.6
DreamGen (gr1) (Jang et al., 2025)	31.2	37.2	29.7	33.4	21.5	56.4	53.2	57.9	57.5	42.0
Abot-PhysWorld (Chen et al., 2026)	48.6	54.8	43.4	52.3	45.4	66.2	66.8	53.1	45.5	52.9
Cosmos3-Super (NVIDIA et al., 2026)	48.7	64.2	44.4	59.1	39.5	61.5	62.3	73.9	69.1	58.1
PF-Cosmos (Zhang et al., 2026)	58.9	**69.7**	51.3	60.8	53.0	69.3	70.0	72.2	69.0	63.8
Ours and base model										
Cosmos3-Nano (NVIDIA et al., 2026)	59.1	51.9	49.2	45.2	46.2	69.5	69.1	68.5	64.2	58.1
Cosmos3-Nano + SFT	**61.1**	50.9	**54.2**	**64.2**	55.2	68.5	68.1	77.5	68.2	63.1
CoDance	60.3	65.3	45.7	61.5	**56.5**	**87.2**	**80.5**	**87.1**	**80.3**	**69.4**

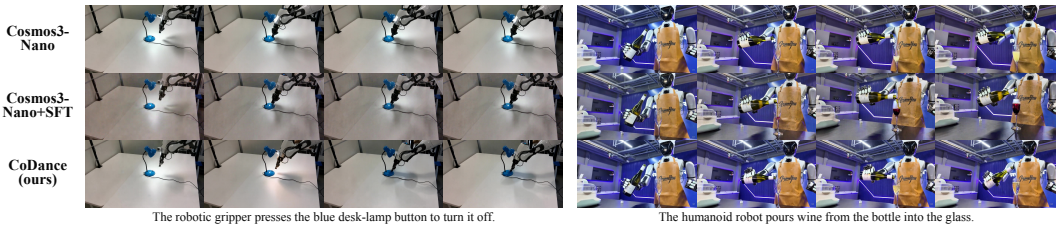


Figure 4: **Qualitative comparison on RBench.** Rows compare Cosmos3-Nano, Cosmos3-Nano + SFT, and CoDance on two robotic interaction tasks, with columns within each case showing temporally ordered frames.

by 1.5 and 2.1 points, respectively. These results support applying reward-decoupled normalization at both levels of the rollout hierarchy.

Joint training. The RL strategies compared in Appendix A.6.2 start from the same SFT checkpoints and use both physical rewards with reward-decoupled normalization. Joint training with CoDanceRL reaches an IQ-Score of 45.4, compared with 44.1 for sequential GRPO+DanceGRPO, 41.5 for reasoner-only GRPO, and 39.3 for generator-only DanceGRPO. The 1.3-point margin over the sequential baseline favors joint adaptation among the evaluated training configurations.

5 Related Work

Physically Consistent Video Generation. Two complementary directions in physical consistency research focus on constraining generated dynamics and improving the physical guidance supplied to the generator. The first grounds generation in simulation (Liu et al., 2024) or adapts generators through physical supervision (Xiong et al., 2026; Zhang et al., 2026; Lin et al., 2026) and physics-aware preferences or rewards (Cai et al., 2026; Le et al., 2025). The second makes interactions explicit in the conditioning through prompt refinement (Xue et al., 2025a; Liu et al., 2025) or struc-

Table 3: WorldArena action-planning closed-loop success rate (%).

Model	Adjust bottle	Click bell	Avg.
RoboMaster (Fu et al., 2026)	8.0	20.0	14.0
Wan2.2-TI2V-5B (Wan et al., 2025)	12.0	20.0	16.0
TesserAct (Zhen et al., 2025)	1.0	35.0	18.0
WoW (Chi et al., 2025)	20.0	21.0	20.5
PhysisForcing (Zhang et al., 2026)	22.0	26.0	24.0
Cosmos3-Nano (NVIDIA et al., 2026)	15.0	13.0	14.0
Cosmos3-Nano+SFT	21.0	24.0	22.5
CoDance	**34.0**	**49.0**	**41.5**

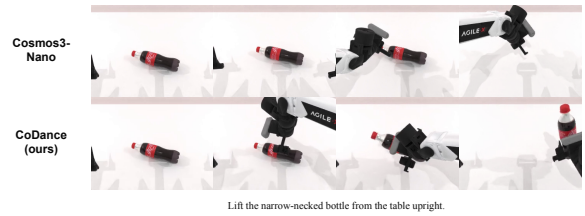


Figure 5: **Qualitative comparison on adjust_bottle.** CoDance grasps the bottle and holds it upright, whereas Cosmos3-Nano does not reach this pose.

Table 4: Ablation of physics-aware feedback and hierarchical reward normalization on Physics-IQ Verified. Checkmarks (✓) and crosses (✗) denote enabled and disabled components, respectively. Higher IQ-Score is better.

Configuration	Reward		Hierarchical Reward Normalization		IQ-Score↑
	Physical fidelity	Geometric trajectory	Video level	Prompt level	
Physical-fidelity reward	✓	✗	✗	✗	42.8
Geometric trajectory reward	✗	✓	✗	✗	40.1
Combined rewards	✓	✓	✗	✗	41.3
Video-level normalization	✓	✓	✓	✗	43.9
Prompt-level normalization	✓	✓	✗	✓	43.3
CoDance (full)	✓	✓	✓	✓	**45.4**

tured keyframes and trajectories (Zhuang et al., 2026), without updating the generator. Generation feedback connects these directions by using the physical consistency of generated videos to train the conditioning model. PhyPrompt uses this feedback to learn a rewriter with a frozen generator (Wu et al., 2026), whereas PhysMaster adapts both the physical encoder and generator in stages (Ji et al., 2025).

Reinforcement Learning for Video Generation. I2V post-training depends on both the update rule and the video feedback that drives it. DanceGRPO applies GRPO to I2V generators (Xue et al., 2025b), while TAGRPO refines the update through reward-ranked alignment of denoising trajectories (Wang et al., 2026a). The reward specifies the desired behavior. Scaled reward models support preference alignment (Wu et al., 2025b), while task-specific rewards target motion control (Zhao et al., 2026) or identity preservation (Shen et al., 2026). These methods adapt the generator under given conditioning. VANS also trains the VLM producing that conditioning for next-event prediction, but does so before adapting the video generator in a separate second stage (Cheng et al., 2026). CoDance instead co-trains the reasoner and generator throughout one coupled RL process using shared physical rewards, with separate advantages for prompt rewrites and video rollouts.

6 Conclusion

We presented **CoDance**, an image-to-video model that combines physical reasoning with video generation, and **CoDanceRL**, a framework that jointly trains both components using shared feedback. Coupled policy optimization, complementary physical rewards, and hierarchical reward normalization allow the reasoner and generator to improve together. CoDance achieves an IQ-Score of 45.4 on Physics-IQ Verified, an average score of 69.4 on RBench, and 41.5% average closed-loop planning success across two WorldArena tasks. Ablations support adapting both components and normalizing rewards at both levels of the rollout hierarchy, with joint training also outperforming the sequential baseline on Physics-IQ Verified. A held-out VLM judge also reports gains across all five criteria, with an overall score of 52.1, compared with 31.7 for the base model. Trajectory fidelity remains a weakness, motivating further work on matching reference motion.

AI USE STATEMENT

Generative AI tools were used in research execution and to generate synthetic training data in two specific roles. First, GPT-5.6 Sol generated physically informed prompt rewrites that served as supervised training targets for reasoner fine-tuning, as described in Appendix A.3. Second, GPT-5.6 Luna and GPT performed automated model evaluation as the VLM judges for Physics-IQ Verified and RBench, respectively, as described in Appendix A.4. These judges were excluded from reward optimization. LLM-based tools were also used to improve the grammar, clarity, and readability of author-written text and to help locate potentially relevant literature and source material. They were not used for research ideation, to originate scientific claims or technical content, to formulate or prove mathematical claims, or for other research tasks requiring disclosure beyond the roles stated above. All AI-assisted work was conducted under author supervision and reviewed before inclusion. The authors verified candidate references against their original sources and checked the AI-assisted experimental outputs and their aggregation. The authors take responsibility for the final content of this work, including its text, claims, citations, data, and reported results.

REFERENCES

Sand. ai, Hansi Teng, Hongyu Jia, Lei Sun, Lingzhi Li, Maolin Li, Mingqiu Tang, Shuai Han, Tianning Zhang, W. Q. Zhang, Weifeng Luo, Xiaoyang Kang, Yuchen Sun, Yue Cao, Yunpeng Huang, Yutong Lin, Yuxin Fang, Zewei Tao, Zheng Zhang, Zhongshu Wang, Zixun Liu, Dai Shi, Guoli Su, Hanwen Sun, Hong Pan, Jie Wang, Jiexin Sheng, Min Cui, Min Hu, Ming Yan, Shucheng Yin, Siran Zhang, Tingting Liu, Xianping Yin, Xiaoyu Yang, Xin Song, Xuan Hu, Yankai Zhang, and Yuqiao Li. Magi-1: Autoregressive video generation at scale, 2025. URL https://arxiv.org/abs/2505.13211.

Alibaba Cloud. Alibaba Unveils Wan2.6 Series Enabling Everyone to Star in Videos. Official announcement, 2025. URL https://www.alibabacloud.com/en/press-room/alibaba-unveils-wan2-6-series-enabling-everyone.

Hritik Bansal, Clark Peng, Yonatan Bitton, Roman Goldenberg, Aditya Grover, and Kai-Wei Chang. Videophy-2: A challenging action-centric physical commonsense evaluation in video generation, 2025. URL https://arxiv.org/abs/2503.06800.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets, 2023. URL https://arxiv.org/abs/2311.15127.

Yuanhao Cai, Kunpeng Li, Menglin Jia, Jialiang Wang, Junzhe Sun, Feng Liang, Weifeng Chen, Felix Juefei-Xu, Chu Wang, Ali Thabet, Xiaoliang Dai, Xuan Ju, Alan Yuille, and Ji Hou. Phygdpo: Physics-aware groupwise direct preference optimization for physically consistent text-to-video generation, 2026. URL https://arxiv.org/abs/2512.24551.

Yuzhi Chen, Ronghan Chen, Dongjie Huo, Yandan Yang, Dekang Qi, Haoyun Liu, Tong Lin, Shuang Zeng, Junjin Xiao, Xinyuan Chang, Feng Xiong, Xing Wei, Zhiheng Ma, and Mu Xu. Abot-physworld: Interactive world foundation model for robotic manipulation with physics alignment, 2026. URL https://arxiv.org/abs/2603.23376.

Junhao Cheng, Liang Hou, Xin Tao, and Jing Liao. Video-as-answer: Predict and generate next video event with joint-GRPO. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 38915–38925, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Cheng_Video-as-Answer_Predict_and_Generate_Next_Video_Event_with_Joint-GRPO_CVPR_2026_paper.html.

Xiaowei Chi, Peidong Jia, Chun-Kai Fan, Xiaozhu Ju, Weishi Mi, Kevin Zhang, Zhiyuan Qin, Wanxin Tian, Kuangzhi Ge, Hao Li, Zezhong Qian, Anthony Chen, Qiang Zhou, Yueru Jia, Jiaming Liu, Yong Dai, Qingpo Wuwu, Chengyu Bai, Yu-Kai Wang, Ying Li, Lizhang Chen, Yong Bao, Zhiyuan Jiang, Jiacheng Zhu, Kai Tang, Ruichuan An, Yulin Luo, Qiuxuan Feng, Siyuan Zhou, Chi min Chan, Chengkai Hou, Wei Xue, Sirui Han, Yike Guo, Shanghang Zhang,

and Jian Tang. Wow: Towards a world omniscient world model through embodied interaction, 2025. URL https://arxiv.org/abs/2509.22642.

Yufan Deng, Zilin Pan, Hongyu Zhang, Xiaojie Li, Ruoqing Hu, Yufei Ding, Yiming Zou, Yan Zeng, and Daquan Zhou. Rethinking video generation model for the embodied world, 2026. URL https://arxiv.org/abs/2601.15282.

Yilun Du, Mengjiao Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Joshua B. Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation, 2023. URL https://arxiv.org/abs/2302.00111.

Yao Feng, Hengkai Tan, Xinyi Mao, Chendong Xiang, Guodong Liu, Shuhe Huang, Hang Su, and Jun Zhu. Vidar: Embodied video diffusion model for generalist manipulation, 2025. URL https://arxiv.org/abs/2507.12898.

Xiao Fu, Xintao Wang, Xian Liu, Jianhong Bai, Runsen Xu, Pengfei Wan, Di Zhang, and Dahua Lin. Learning video generation for robotic manipulation with collaborative trajectory control, 2026. URL https://arxiv.org/abs/2506.01943.

Google DeepMind. Veo 3.1. Official model page, 2025. URL https://deepmind.google/models/veo/.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, Eitan Richardson, Guy Shiran, Itay Chachy, Jonathan Chetboun, Michael Finkelson, Michael Kupchick, Nir Zabari, Nitzan Guetta, Noa Kotler, Ofir Bibi, Ori Gordon, Poriya Panet, Roi Benita, Shahar Armon, Victor Kulikov, Yaron Inger, Yonatan Shiftan, Zeev Melumian, and Zeev Farbman. Ltx-2: Efficient joint audio-visual foundation model, 2026. URL https://arxiv.org/abs/2601.03233.

Christian Internò, Alexander Pondaven, Habon Issa, Fabio Pizzati, Francesco Pinto, Markus Olhofer, Ivan Laptev, Philip Torr, Eero P. Simoncelli, Barbara Hammer, and David Klindt. Geophys: The geometry of physical plausibility, 2026. URL https://arxiv.org/abs/2606.20707.

Joel Jang, Seonghyeon Ye, Zongyu Lin, Jiannan Xiang, Johan Bjorck, Yu Fang, Fengyuan Hu, Spencer Huang, Kaushil Kundalia, Yen-Chen Lin, Loic Magne, Ajay Mandlekar, Avnish Narayan, You Liang Tan, Guanzhi Wang, Jing Wang, Qi Wang, Yinzhen Xu, Xiaohui Zeng, Kaiyuan Zheng, Ruijie Zheng, Ming-Yu Liu, Luke Zettlemoyer, Dieter Fox, Jan Kautz, Scott Reed, Yuke Zhu, and Linxi Fan. Dreamgen: Unlocking generalization in robot learning through video world models, 2025. URL https://arxiv.org/abs/2505.12705.

Sihui Ji, Xi Chen, Xin Tao, Pengfei Wan, and Hengshuang Zhao. Physmaster: Mastering physical representation for video generation via reinforcement learning, 2025. URL https://arxiv.org/abs/2510.13809.

Nikita Karaev, Iurii Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker3: Simpler and better point tracking by pseudo-labelling real videos, 2024. URL https://arxiv.org/abs/2410.11831.

Kuaishou Technology. Kling AI Launches Video 2.6 Model with "Simultaneous Audio-Visual Generation" Capability, Redefining AI Video Creation Workflow. Official announcement, 2025. URL https://ir.kuaishou.com/news-releases/news-release-details/kling-ai-launches-video-26-model-simultaneous-audio-visual.

Minh-Quan Le, Yuanzhi Zhu, Vicky Kalogeiton, and Dimitris Samaras. What about gravity in video generation? post-training newton's laws with verifiable rewards, 2025. URL https://arxiv.org/abs/2512.00425.

Junbang Liang, Pavel Tokmakov, Ruoshi Liu, Sruthi Sudhakar, Paarth Shah, Rares Ambrus, and Carl Vondrick. Video generators are robot policies, 2025. URL https://arxiv.org/abs/2508.00795.

Shubo Lin, Xuanyang Zhang, Wei Cheng, Weiming Hu, Gang Yu, and Jin Gao. Mmphysvideo: Physically plausible video generation through joint rgb-perception modeling, 2026. URL https://arxiv.org/abs/2604.02817.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling, 2023. URL https://arxiv.org/abs/2210.02747.

Shaowei Liu, Zhongzheng Ren, Saurabh Gupta, and Shenlong Wang. Physgen: Rigid-body physics-grounded image-to-video generation, 2024. URL https://arxiv.org/abs/2409.18964.

Shih-Yang Liu, Xin Dong, Ximing Lu, Shizhe Diao, Peter Belcak, Mingjie Liu, Min-Hung Chen, Hongxu Yin, Yu-Chiang Frank Wang, Kwang-Ting Cheng, Yejin Choi, Jan Kautz, and Pavlo Molchanov. Gdpo: Group reward-decoupled normalization policy optimization for multi-reward rl optimization, 2026. URL https://arxiv.org/abs/2601.05242.

Yang Liu, Xilin Zhao, Peisong Wen, Siran Dai, and Qingming Huang. Bootstrapping physics-grounded video generation through vlm-guided iterative self-refinement, 2025. URL https://arxiv.org/abs/2511.20280.

MiniMax. MiniMax Hailuo 02, World-Class Quality, Record-Breaking Cost Efficiency. Official announcement, 2025. URL https://www.minimax.io/news/minimax-hailuo-02.

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles?, 2025. URL https://arxiv.org/abs/2501.09038.

Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation, 2025. URL https://arxiv.org/abs/2407.02371.

NVIDIA, Aditi, Niket Agarwal, Arslan Ali, Jon Allen, Martin Antolini, Adeline Aubame, Alisson Azzolini, Junjie Bai, Maciej Bala, Yogesh Balaji, Josh Bapst, Aarti Basant, Mukesh Beladiya, Mohammad Qazim Bhat, Zaid Pervaiz Bhat, Dan Blick, Vanni Brighella, Han Cai, Tiffany Cai, Eric Cameracci, Jiaxin Cao, Yulong Cao, Mark Carlson, Carlos Casanova, Ting-Yun Chang, Yan Chang, Yu-Wei Chao, Prithvijit Chattopadhyay, Roshan Chaudhari, Chieh-Yun Chen, Junyu Chen, Ke Chen, Qizhi Chen, Wenkai Chen, Xiaotong Chen, Yu Chen, An-Chieh Cheng, Click Cheng, Xiu Chia, Jeana Choi, Chaeyeon Chung, Wenyan Cong, Yin Cui, Magdalena Dadela, Nalin Dadhich, Wenliang Dai, Joyjit Daw, Alperen Degirmenci, Rodrigo Vieira Del Monte, Robert Denomme, Sameer Dharur, Marco Di Lucca, Ke Ding, Wenhao Ding, Yifan Ding, Yuzhu Dong, Nicole Drumheller, Yilun Du, Aigul Dzhumamuratova, Aleksandr Efitorov, Hamid Eghbalzadeh, Naomi Eigbe, Imad El Hanafi, Hassan Eslami, Benedikt Falk, Jiaojiao Fan, Jim Fan, Amol Fasale, Sergiy Fefilatyev, Liang Feng, Francesco Ferroni, Sanja Fidler, Xiao Fu, Vikram Fugro, Prashant Gaikwad, TJ Galda, Katelyn Gao, Yihuai Gao, Wenhang Ge, Sreyan Ghosh, Arushi Goel, Vivek Goel, Akash Gokul, Rama Govindaraju, Jinwei Gu, Miguel Guerrero, Elfie Guo, Aryaman Gupta, Siddharth Gururani, Hugo Hadfield, Song Han, Ankur Handa, Zekun Hao, Mohammad Harrim, Ali Hassani, Nathan Hayes-Roth, Yufan He, Chris Helvig, Cyrus Hogg, Madison Huang, Michael Huang, Sophia Huang, Yufan Huang, Jacob Huffman, DeLesley Hutchins, Suneel Indupuru, Boris Ivanovic, Arihant Jain, Joel Jang, Ryan Ji, Yanan Jian, Dongfu Jiang, Jingyi Jin, Atharva Joshi, Nikhilesh Joshi, Pranjali Joshi, Andy Ju, Jaehun Jung, Weiwei Kang, Scott Kassekert, Jan Kautz, Ashna Khetan, Julia Kiczka, Slawek Kierat, Gwanghyun Kim, Kuno Kim, Sunny Kim, Kezhi Kong, Xin Kong, Zhifeng Kong, Tomasz Kornuta, Egor Krivov, Hui Kuang, Saurav Kumar, Chia-Wen Kuo, George Kurian, Wojciech Kutak, JF Lafleche, Himangshu Lahkar, Omar Laymoun, Jayjun Lee, Sanggil Lee, Gabriele Leone, Boyi Li, Freya Li, Jiajun Li, Jinfeng Li, Ling Li, Pengcheng Li, Shangru Li, Tingle Li, Xiaolong Li, Xuan Li, Zhaoshuo Li, Zhiqi Li, Hao Liang, Maosheng Liao, Chen-Hsuan Lin, Tsung-Yi Lin, Ming-Yu Liu, Sifei Liu, Zihan Liu, Hai Loc Lu, Xiangyu Lu, Alice Luo, Ruipu Luo, Wenjie Luo, Jiangran Lyu, Martin Ding Ma, Nic Ma, Qianli Ma, Dawid Majchrowski, Louis Marcoux, Miguel Martin, Qing Miao, Ashkan Mirzaei, Shreyas Misra, Kaichun Mo, Durra Mohsin, Hyejin Moon, Pawel Morkisz, Saeid Motiian, Kirill Motkov, Seungjun Nah, Yashraj Narang, Deepak Narayanan, Thabang Ngazimbi, Julian Ouyang, Shubham Pachori, David Page, Yatian Pang, Sehwi Park, Mahesh Patekar, Mostofa Patwary, Marco Pavone, Trung Pham, Wei Ping, Soha Pouya, Shrimai

Prabhumoye, Varun Praveen, Delin Qu, Hesam Rabeti, Morteza Ramezanali, Marilyn Reeb, Xuanchi Ren, Kristen Rumley, Wojciech Rymer, Jun Saito, Yeongho Seol, John Shao, Piyush Shekdar, Tianwei Shen, Humphrey Shi, Min Shi, Stella Shi, Kevin Shih, Mohammad Shoeybi, Mateusz Sieniawski, Shuran Song, Alexander Sotelo, Amir Sotoodeh, Sunil Srinivasa, Vignesh Srinivasakumar, Bartosz Stefaniak, Rahul Heinrich Steiger, Shangkun Sun, Jiaxiang Tang, Shitao Tang, Yangyang Tang, Yue Tang, Tolou Tavakkoli, Kayley Ting, Krzysztof Tomala, Wei-Cheng Tseng, Jibin Varghese, Sergei Vasilev, Thomas Volk, Raju Wagwani, Roger Waleffe, Andrew Z. Wang, Boxiang Wang, Haoxiang Wang, Qiao Wang, Shihao Wang, Shijie Wang, Ting-Chun Wang, Yan Wang, Yu Wang, Rohit Watve, David Wehr, Fangyin Wei, Xinshuo Weng, Jay Zhangjie Wu, Kedi Wu, Hongchi Xia, Summer Xiao, Tianjun Xiao, Kevin Xie, Daguang Xu, Jiashu Xu, Mengyao Xu, Ruqing Xu, Xingqian Xu, Yao Xu, Dinghao Yang, Dong Yang, Hans Yang, Xiaodong Yang, Xuning Yang, Yichu Yang, Yurong You, Zhiding Yu, Hao Yuan, Simon Yuen, Xiaohui Zeng, Pengcuo Zeren, Cindy Zha, Haotian Zhang, Jenny Zhang, Jing Zhang, Liangkai Zhang, Paris Zhang, Shun Zhang, Xuanmeng Zhang, Zhizheng Zhang, Ann Zhao, Yilin Zhao, Yuliya Zhautouskaya, Charles Zhou, Fengzhe Zhou, Shilin Zhu, Yuke Zhu, Dima Zhylko, and Artur Zolkowski. Cosmos 3: Omnimodal world models for physical ai, 2026. URL https://arxiv.org/abs/2606.02800.

OpenAI. Sora 2 System Card. System card, 2025. URL https://openai.com/index/sora-2-system-card/.

OpenAI. GPT-5.6 Sol. Official model documentation, 2026. URL https://developers.openai.com/api/docs/models/gpt-5.6-sol. Accessed September 24, 2026.

William Peebles and Saining Xie. Scalable diffusion models with transformers, 2023. URL https://arxiv.org/abs/2212.09748.

Tim Rädsch, Yuki M Asano, Hilde Kuehne, Stefan Bauer, Priyank Jaini, Robert Geirhos, and Carsten T. Lüth. Physics-iq verified, 2026. URL https://arxiv.org/abs/2606.18943.

Team Seedance, Heyi Chen, Siyan Chen, Xin Chen, Yanfei Chen, Ying Chen, Zhuo Chen, Feng Cheng, Tianheng Cheng, Xinqi Cheng, Xuyan Chi, Jian Cong, Jing Cui, Qinpeng Cui, Qide Dong, Junliang Fan, Jing Fang, Zetao Fang, Chengjian Feng, Han Feng, Mingyuan Gao, Yu Gao, Dong Guo, Qiushan Guo, Boyang Hao, Qingkai Hao, Bibo He, Qian He, Tuyen Hoang, Ruoqing Hu, Xi Hu, Weilin Huang, Zhaoyang Huang, Zhongyi Huang, Donglei Ji, Siqi Jiang, Wei Jiang, Yunpu Jiang, Zhuo Jiang, Ashley Kim, Jianan Kong, Zhichao Lai, Shanshan Lao, Yichong Leng, Ai Li, Feiya Li, Gen Li, Huixia Li, JiaShi Li, Liang Li, Ming Li, Shanshan Li, Tao Li, Xian Li, Xiaojie Li, Xiaoyang Li, Xingxing Li, Yameng Li, Yifu Li, Yiying Li, Chao Liang, Han Liang, Jianzhong Liang, Ying Liang, Zhiqiang Liang, Wang Liao, Yalin Liao, Heng Lin, Kengyu Lin, Shanchuan Lin, Xi Lin, Zhijie Lin, Feng Ling, Fangfang Liu, Gaohong Liu, Jiawei Liu, Jie Liu, Jihao Liu, Shouda Liu, Shu Liu, Sichao Liu, Songwei Liu, Xin Liu, Xue Liu, Yibo Liu, Zikun Liu, Zuxi Liu, Junlin Lyu, Lecheng Lyu, Qian Lyu, Han Mu, Xiaonan Nie, Jingzhe Ning, Xitong Pan, Yanghua Peng, Lianke Qin, Xueqiong Qu, Yuxi Ren, Kai Shen, Guang Shi, Lei Shi, Yan Song, Yinglong Song, Fan Sun, Li Sun, Renfei Sun, Yan Sun, Zeyu Sun, Wenjing Tang, Yaxue Tang, Zirui Tao, Feng Wang, Furui Wang, Jinran Wang, Junkai Wang, Ke Wang, Kexin Wang, Qingyi Wang, Rui Wang, Sen Wang, Shuai Wang, Tingru Wang, Weichen Wang, Xin Wang, Yanhui Wang, Yue Wang, Yuping Wang, Yuxuan Wang, Ziyu Wang, Guoqiang Wei, Wanru Wei, Di Wu, Guohong Wu, Hanjie Wu, Jian Wu, Jie Wu, Ruolan Wu, Xinglong Wu, Yonghui Wu, Ruiqi Xia, Liang Xiang, Fei Xiao, XueFeng Xiao, Pan Xie, Shuangyi Xie, Shuang Xu, Jinlan Xue, Shen Yan, Bangbang Yang, Ceyuan Yang, Jiaqi Yang, Runkai Yang, Tao Yang, Yang Yang, Yihang Yang, ZhiXian Yang, Ziyan Yang, Songting Yao, Yifan Yao, Zilyu Ye, Bowen Yu, Jian Yu, Chujie Yuan, Linxiao Yuan, Sichun Zeng, Weihong Zeng, Xuejiao Zeng, Yan Zeng, Chuntao Zhang, Heng Zhang, Jingjie Zhang, Kuo Zhang, Liang Zhang, Liying Zhang, Manlin Zhang, Ting Zhang, Weida Zhang, Xiaohe Zhang, Xinyan Zhang, Yan Zhang, Yuan Zhang, Zixiang Zhang, Fengxuan Zhao, Huating Zhao, Yang Zhao, Hao Zheng, Jianbin Zheng, Xiaozheng Zheng, Yangyang Zheng, Yijie Zheng, Jiexin Zhou, Jiahui Zhu, Kuan Zhu, Shenhan Zhu, Wenjia Zhu, Benhui Zou, and Feilong Zuo. Seedance 1.5 pro: A native audio-visual joint generation foundation model, 2025. URL https://arxiv.org/abs/2512.13507.

Team Seedance, De Chen, Liyang Chen, Xin Chen, Ying Chen, Zhuo Chen, Zhuowei Chen, Feng Cheng, Tianheng Cheng, Yufeng Cheng, Mojie Chi, Xuyan Chi, Jian Cong, Qinpeng Cui, Fei

Ding, Qide Dong, Yujiao Du, Haojie Duanmu, Junliang Fan, Jiarui Fang, Jing Fang, Zetao Fang, Chengjian Feng, Yu Gao, Diandian Gu, Dong Guo, Hanzhong Guo, Qiushan Guo, Boyang Hao, Hongxiang Hao, Haoxun He, Jiaao He, Qian He, Tuyen Hoang, Heng Hu, Ruoqing Hu, Yuxiang Hu, Jiancheng Huang, Weilin Huang, Zhaoyang Huang, Zhongyi Huang, Jishuo Jin, Ming Jing, Ashley Kim, Shanshan Lao, Yichong Leng, Bingchuan Li, Gen Li, Haifeng Li, Huixia Li, Jiashi Li, Ming Li, Xiaojie Li, Xingxing Li, Yameng Li, Yiying Li, Yu Li, Yueyan Li, Chao Liang, Han Liang, Jianzhong Liang, Ying Liang, Wang Liao, J. H. Lien, Shanchuan Lin, Xi Lin, Feng Ling, Yue Ling, Fangfang Liu, Jiawei Liu, Jihao Liu, Jingtuo Liu, Shu Liu, Sichao Liu, Wei Liu, Xue Liu, Zuxi Liu, Ruijie Lu, Lecheng Lyu, Jingting Ma, Tianxiang Ma, Xiaonan Nie, Jingzhe Ning, Junjie Pan, Xitong Pan, Ronggui Peng, Xueqiong Qu, Yuxi Ren, Yuchen Shen, Guang Shi, Lei Shi, Yinglong Song, Fan Sun, Li Sun, Renfei Sun, Wenjing Tang, Boyang Tao, Zirui Tao, Dongliang Wang, Feng Wang, Hulin Wang, Ke Wang, Qingyi Wang, Rui Wang, Shuai Wang, Shulei Wang, Weichen Wang, Xuanda Wang, Yanhui Wang, Yue Wang, Yuping Wang, Yuxuan Wang, Zijie Wang, Ziyu Wang, Guoqiang Wei, Meng Wei, Di Wu, Guohong Wu, Hanjie Wu, Huachao Wu, Jian Wu, Jie Wu, Ruolan Wu, Shaojin Wu, Xiaohu Wu, Xinglong Wu, Yonghui Wu, Ruiqi Xia, Xin Xia, Xuefeng Xiao, Shuang Xu, Bangbang Yang, Jiaqi Yang, Runkai Yang, Tao Yang, Yihang Yang, Zhixian Yang, Ziyan Yang, Fulong Ye, Bingqian Yi, Xing Yin, Yongbin You, Linxiao Yuan, Weihong Zeng, Xuejiao Zeng, Yan Zeng, Siyu Zhai, Zhonghua Zhai, Bowen Zhang, Chenlin Zhang, Heng Zhang, Jun Zhang, Manlin Zhang, Peiyuan Zhang, Shuo Zhang, Xiaohe Zhang, Xiaoying Zhang, Xinyan Zhang, Xinyi Zhang, Yichi Zhang, Zixiang Zhang, Haiyu Zhao, Huating Zhao, Liming Zhao, Yian Zhao, Guangcong Zheng, Jianbin Zheng, Xiaozheng Zheng, Zerong Zheng, Kuan Zhu, and Feilong Zuo. Seedance 2.0: Advancing video generation for world complexity, 2026. URL https://arxiv.org/abs/2604.14148.

Shuyao Shang, Yuqi Wang, Ruopeng Gao, Xu Chen, Tieniu Tan, Lue Fan, and Zhaoxiang Zhang. Phizero: A world model built around physical language, 2026a. URL https://arxiv.org/abs/2607.28624.

Yu Shang, Zhuohang Li, Yiding Ma, Weikang Su, Xin Jin, Ziyou Wang, Lei Jin, Xin Zhang, Yinzhou Tang, Haisheng Su, Chen Gao, Wei Wu, Xihui Liu, Dhruv Shah, Zhaoxiang Zhang, Zhibo Chen, Jun Zhu, Yonghong Tian, Tat-Seng Chua, Wenwu Zhu, and Yong Li. Worldarena: A unified benchmark for evaluating perception and functional utility of embodied world models, 2026b. URL https://arxiv.org/abs/2602.08971.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Liao Shen, Wentao Jiang, Yiran Zhu, Jiahe Li, Tiezheng Ge, Zhiguo Cao, and Bo Zheng. Identity-preserving image-to-video generation via reward-guided optimization, 2026. URL https://arxiv.org/abs/2510.14255.

Meituan LongCat Team, Xunliang Cai, Qilong Huang, Zhuoliang Kang, Hongyu Li, Shijun Liang, Liya Ma, Siyu Ren, Xiaoming Wei, Rixu Xie, and Tong Zhang. Longcat-video technical report, 2025. URL https://arxiv.org/abs/2510.22200.

Team fal. Introducing H3 Max by fal. Official announcement, 2026. URL https://blog.fal.ai/introducing-h3-max-by-fal/.

Unitree. Unifolm-wma-0: A world-model-action (wma) framework under unifolm family, 2025.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models, 2025. URL https://arxiv.org/abs/2503.20314.

Jin Wang, Jianxiang Lu, Guangzheng Xu, Comi Chen, Haoyu Yang, Linqing Wang, Peng Chen, Mingtao Chen, Zhichao Hu, Longhuang Wu, Shuai Shao, Qinglin Lu, and Ping Luo. Tagrpo: Boosting grpo on image-to-video generation with direct trajectory alignment, 2026a. URL https://arxiv.org/abs/2601.05729.

Zixuan Wang, Yixin Hu, Haolan Wang, Feng Chen, Yan Liu, Wen Li, and Yinjie Lei. Chain of event-centric causal thought for physically plausible video generation, 2026b. URL https://arxiv.org/abs/2603.09094.

Bing Wu, Chang Zou, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Jack Peng, Jianbing Wu, Jiangfeng Xiong, Jie Jiang, Linus, Patrol, Peizhen Zhang, Peng Chen, Penghao Zhao, Qi Tian, Songtao Liu, Weijie Kong, Weiyan Wang, Xiao He, Xin Li, Xinchi Deng, Xuefei Zhe, Yang Li, Yanxin Long, Yuanbo Peng, Yue Wu, Yuhong Liu, Zhenyu Wang, Zuozhuo Dai, Bo Peng, Coopers Li, Gu Gong, Guojian Xiao, Jiahe Tian, Jiaxin Lin, Jie Liu, Jihong Zhang, Jiesong Lian, Kaihang Pan, Lei Wang, Lin Niu, Mingtao Chen, Mingyang Chen, Mingzhe Zheng, Miles Yang, Qiangqiang Hu, Qi Yang, Qiuyong Xiao, Runzhou Wu, Ryan Xu, Rui Yuan, Shanshan Sang, Shisheng Huang, Siruis Gong, Shuo Huang, Weiting Guo, Xiang Yuan, Xiaojia Chen, Xiawei Hu, Wenzhi Sun, Xiele Wu, Xianshun Ren, Xiaoyan Yuan, Xiaoyue Mi, Yepeng Zhang, Yifu Sun, Yiting Lu, Yitong Li, You Huang, Yu Tang, Yixuan Li, Yuhang Deng, Yuan Zhou, Zhichao Hu, Zhiguang Liu, Zhihe Yang, Zilin Yang, Zhenzhi Lu, Zixiang Zhou, and Zhao Zhong. Hunyuanvideo 1.5 technical report, 2025a. URL https://arxiv.org/abs/2511.18870.

Jie Wu, Yu Gao, Zilyu Ye, Ming Li, Liang Li, Hanzhong Guo, Jie Liu, Zeyue Xue, Xiaoxia Hou, Wei Liu, Yan Zeng, and Weilin Huang. Rewarddance: Reward scaling in visual generation, 2025b. URL https://arxiv.org/abs/2509.08826.

Shang Wu, Chenwei Xu, Zhuofan Xia, Weijian Li, Lie Lu, Pranav Maneriker, Fan Du, Manling Li, and Han Liu. Phyprompt: Rl-based prompt refinement for physically plausible text-to-video generation, 2026. URL https://arxiv.org/abs/2603.03505.

xAI. Grok Imagine API. Official announcement, 2026. URL https://x.ai/news/grok-imagine-api.

Zhexiao Xiong, Yizhi Song, Liu He, Wei Xiong, Yu Yuan, Feng Qiao, and Nathan Jacobs. Physalign: Physics-coherent image-to-video generation through feature and 3d representation alignment, 2026. URL https://arxiv.org/abs/2603.13770.

Qiyao Xue, Xiangyu Yin, Boyuan Yang, and Wei Gao. Phyt2v: Llm-guided iterative self-refinement for physics-grounded text-to-video generation, 2025a. URL https://arxiv.org/abs/2412.00596.

Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, and Ping Luo. Dancegrpo: Unleashing grpo on visual generation, 2025b. URL https://arxiv.org/abs/2505.07818.

Peiwen Zhang, Yufan Deng, Shangkun Sun, Juncheng Ma, Duomin Wang, Jonas Du, Zilin Pan, Ye Huang, Hao Liang, Songyan Huang, Ruihua Zhang, Enze Xie, Ming-Yu Liu, and Daquan Zhou. Physisforcing: Physics reinforced world simulator for robotic manipulation, 2026. URL https://arxiv.org/abs/2606.28128.

Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models, 2023. URL https://arxiv.org/abs/2311.04145.

Kesen Zhao, Jiaxin Shi, Beier Zhu, Junbao Zhou, Xiaolong Shen, Yuan Zhou, Qianru Sun, and Hanwang Zhang. Real-time motion-controllable autoregressive video diffusion, 2026. URL https://arxiv.org/abs/2510.08131.

Haoyu Zhen, Qiao Sun, Hongxin Zhang, Junyan Li, Siyuan Zhou, Yilun Du, and Chuang Gan. Tesseract: Learning 4d embodied world models, 2025. URL https://arxiv.org/abs/2504.20995.

Sihan Zhuang, Xinyuan Chen, Tianfan Xue, and Yaohui Wang. Causalmotion: Structured physical reasoning as keyframe and trajectory guidance for training-free video generation, 2026. URL https://arxiv.org/abs/2606.14317.

A APPENDIX

A.1 CODANCERL ALGORITHM PIPELINE

Algorithm 1 summarizes CoDanceRL using the notation in Section 3. We omit the input-condition index for readability. The final batch normalization pools samples from all conditions in the rollout batch.

Algorithm 1 CoDanceRL

Input
  1) Dataset $\mathcal{D}$ containing conditions $c = (I, p)$ and reference videos $V^\star$
  2) Reasoner π_ϕ, generator p_θ, and frozen reference policy $\pi_{\phi_{\text{ref}}}$
  3) Rewrites per input G, videos per rewrite l, and sampling steps N
  4) Clipping threshold ϵ, stabilizer $\epsilon_{\text{norm}} > 0$, KL weight β, and learning rates η_ϕ, η_θ
Initialize parameters ϕ, θ from supervised fine-tuning
for each training iteration **do**
  Freeze old policies by setting $\phi_{\text{old}} \leftarrow \phi$, $\theta_{\text{old}} \leftarrow \theta$
  Sample a batch $\mathcal{C} \subset \mathcal{D}$ and initialize rollout buffer $\mathcal{B} \leftarrow \emptyset$
  Nested rollout collection
  for each $(c, V^\star) \in \mathcal{C}$ **do**
    Sample rewrites $\{p_i\}_{i=1}^{G} \sim \pi_{\phi_{\text{old}}}(\cdot \mid c)$
    for $i = 1$ to G **do**
      Set the generation condition $c_i \leftarrow (I, p_i)$
      for $j = 1$ to l **do**
        Sample $\xi_{ij} \sim \mathcal{N}(0, \mathbf{I})$ and set $z_{ij,N} \leftarrow \xi_{ij}$
        Sample $z_{ij,n-1} \sim p_{\theta_{\text{old}}}(\cdot \mid z_{ij,n}, c_i)$ for $n = N, \ldots, 1$
        Decode $z_{ij,0}$ into video V_{ij}
        Compute rewards r_{ij}, r'_{ij} using Eqs. 5 and 6
      end for
      Video-level reward normalization
      Compute means/stds (μ_i, σ_i) of $\{r_{ij}\}_{j=1}^{l}$ and (μ'_i, σ'_i) of $\{r'_{ij}\}_{j=1}^{l}$
      $R_{ij} \leftarrow \frac{r_{ij} - \mu_i}{\sigma_i + \epsilon_{\text{norm}}}, \quad R'_{ij} \leftarrow \frac{r'_{ij} - \mu'_i}{\sigma'_i + \epsilon_{\text{norm}}}$ for all j
      $\widetilde{R}_{ij} \leftarrow \frac{1}{2}(R_{ij} + R'_{ij})$ for all j
      Compute raw reward means $\bar{r}_i \leftarrow \frac{1}{l}\sum_{j=1}^{l} r_{ij}, \quad \bar{r}'_i \leftarrow \frac{1}{l}\sum_{j=1}^{l} r'_{ij}$
    end for
    Prompt-level reward normalization
    Compute means/stds (μ_c, σ_c) of $\{\bar{r}_i\}_{i=1}^{G}$ and (μ'_c, σ'_c) of $\{\bar{r}'_i\}_{i=1}^{G}$
    $R_i \leftarrow \frac{\bar{r}_i - \mu_c}{\sigma_c + \epsilon_{\text{norm}}}, \quad R'_i \leftarrow \frac{\bar{r}'_i - \mu'_c}{\sigma'_c + \epsilon_{\text{norm}}}$ for all i
    $\widetilde{R}_i \leftarrow \frac{1}{2}(R_i + R'_i)$ for all i
    Store conditions, rewrites, latent trajectories, and $\widetilde{R}_i$, $\widetilde{R}_{ij}$ in $\mathcal{B}$
  end for
  Batch normalization
  Compute batch means/stds (μ_v, σ_v) of $\widetilde{R}_{ij}$ and (μ_p, σ_p) of $\widetilde{R}_i$
  Assign advantages $\hat{A}_{ij} \leftarrow \frac{\widetilde{R}_{ij} - \mu_v}{\sigma_v + \epsilon_{\text{norm}}}, \quad \hat{A}_i \leftarrow \frac{\widetilde{R}_i - \mu_p}{\sigma_p + \epsilon_{\text{norm}}}$ to all stored rollouts
  Coupled policy optimization
  for each optimization step on $\mathcal{B}$ **do**
    Compute response-level reasoner ratio $\rho_i(\phi) \leftarrow \frac{\pi_\phi(p_i|c)}{\pi_{\phi_{\text{old}}}(p_i|c)}$
    Update reasoner as $\phi \leftarrow \phi + \eta_\phi \nabla_\phi \mathcal{J}_{\text{R}}(\phi)$ using Eq. 3 *// θ fixed*
    Compute generator ratio $\rho_{ij,n}(\theta) \leftarrow \frac{p_\theta(z_{ij,n-1}|z_{ij,n},c_i)}{p_{\theta_{\text{old}}}(z_{ij,n-1}|z_{ij,n},c_i)}$
    Update generator as $\theta \leftarrow \theta + \eta_\theta \nabla_\theta \mathcal{J}_{\text{G}}(\theta)$ using Eq. 4 *// ϕ fixed*
  end for
end for
Return the trained reasoner π_ϕ and video generator p_θ

A.2 GEOMETRIC TRAJECTORY REWARD COMPUTATION

CoTracker3 tracks a common grid of points from the conditioning image in V_{ij} and $V^\star$. Spatial quantities, including grid spacing s, are normalized by the frame diagonal. A point is moving if its

maximum reference displacement after removing median global translation exceeds s. We use only observations visible in both videos after the conditioning frame. Unstarred and starred quantities below refer to V_{ij} and $V^\star$, respectively. Here, $\mathbb{E}$ is the uniform empirical average over the valid observations for each score.

For geometry, we consider pairs containing at least one moving point. Let $d, d^\star \in [0,1]$ be their normalized distances at jointly visible frames. The soft-contact value $e^{-d/s}$ increases as points approach. Averaging over valid pairs and frames gives

$$m_{ij}^{\text{geom}} = 1 - \frac{1}{2}\mathbb{E}\Big[|d - d^\star| + \Big|e^{-d/s} - e^{-d^\star/s}\Big|\Big]. \tag{11}$$

For timing, we retain these pairs when their reference distance changes by more than s. Let $\tau, \tau^\star \in [0,1]$ be their times of closest approach, normalized by video duration. Averaging over the retained pairs gives

$$m_{ij}^{\text{time}} = 1 - \mathbb{E}[|\tau - \tau^\star|]. \tag{12}$$

For final displacement, let $u, u^\star$ be a moving point's initial-to-final displacements after subtracting each video's median point displacement. We cap the displacement error at one and average over moving points visible in both final frames to obtain

$$m_{ij}^{\text{final}} = 1 - \mathbb{E}[\min(\|u - u^\star\|_2, 1)]. \tag{13}$$

The three scores are averaged in Eq. 6.

A.3 Dataset Construction Details

RL training data. Our RL training dataset combines videos of physical phenomena with embodied interaction videos. To support physically consistent video generation, we select approximately 1,000 videos from OpenVid-1M (Nan et al., 2025) depicting physical phenomena captured with a static camera, and supplement them with 254 real-world videos that we record to cover a range of physical phenomena. To support embodied video generation, we select approximately 3,000 high-quality videos from RoVid-X (Deng et al., 2026) featuring rich physical interactions. All training data are disjoint from the test data used in our evaluations.

SFT training data. The generator is fine-tuned separately on the full RL video dataset. Each training example uses the video's first frame I and original prompt p as the condition, and the corresponding ground-truth video $V^\star$ as the target. For the reasoner, we use GPT-5.6 Sol (OpenAI, 2026) with the `max` reasoning setting to produce a physically informed rewrite of the original prompt for every video in the same dataset. The reasoner takes the first frame I and original prompt p as input and is trained with teacher forcing to predict the teacher-generated rewrite.

A.4 Experimental Setup Details

Compared models. Tables 1, 2, and 3 compare CoDance with general-purpose video models, physics-aware methods, and embodied world models. The general-purpose video baselines include Sora 2 (OpenAI, 2025), Grok-Imagine-Video (xAI, 2026), Seedance 1.5 (Seedance et al., 2025) and 2.0 (Seedance et al., 2026), Kling 2.6 (Kuaishou Technology, 2025), Hailuo v2 (MiniMax, 2025), Veo 3.1 (Google DeepMind, 2025), Wan2.6 (Alibaba Cloud, 2025), Wan2.2-I2V-A14B (Wan et al., 2025), Wan2.2-TI2V-5B (Wan et al., 2025), HunyuanVideo 1.5 (Wu et al., 2025a), MiniMax-H3 Max (Team fal, 2026), LTX-2 (HaCohen et al., 2026), and LongCat-Video (Team et al., 2025). The physics-aware and robotics-oriented baselines include Magi-1 24B+GeoPhys (ai et al., 2025; Internò et al., 2026), PhiZero (Shang et al., 2026a), UnifoLM-WMA-0 (Unitree, 2025), Vidar (Feng et al., 2025), DreamGen (gr1) (Jang et al., 2025), Abot-PhysWorld (Chen et al., 2026), Cosmos3-Super I2V (NVIDIA et al., 2026), Cosmos3-Super (NVIDIA et al., 2026), and PF-Cosmos (Zhang et al., 2026). For closed-loop planning, we compare with RoboMaster (Fu et al., 2026), Wan2.2-TI2V-5B (Wan et al., 2025), TesserAct (Zhen et al., 2025), WoW (Chi et al., 2025), and PhysisForcing (Zhang et al., 2026). We also include Cosmos3-Nano (NVIDIA et al., 2026) and our Cosmos3-Nano + SFT variant to measure the gains from supervised initialization and subsequent reinforcement learning. In Table 8, GRPO (Shao et al., 2024) and DanceGRPO (Xue et al., 2025b) provide reasoner-only and generator-only training baselines, respectively, while component recombinations evaluate the contributions of the trained reasoner and generator.

Result sources. For Physics-IQ Verified (Table 1), we evaluate all models in the *Open-source* and *Ours and base model* groups, as well as MiniMax-H3 Max, Seedance 2.0, and PhiZero. The remaining results, for Sora 2, Grok-Imagine-Video, and Magi-1 24B+GeoPhys, are quoted from the Physics-IQ Verified paper (Rädsch et al., 2026) and released GitHub results.[1]

For RBench (Table 2), we evaluate Seedance 1.5, Cosmos3-Nano, Cosmos3-Nano + SFT, and CoDance. Category scores for all entries in the *Open-source* and *Robotics-specific* groups, and for the remaining *Commercial* entries, are quoted from Table 1 of PhysisForcing (Zhang et al., 2026).

For WorldArena (Table 3), we evaluate Cosmos3-Nano, Cosmos3-Nano + SFT, and CoDance. Results for the other planners are quoted from Table 3 of PhysisForcing (Zhang et al., 2026).

Ablation training protocols. All RL configurations in Table 8(a) start from the same component-wise SFT checkpoints. For GRPO+DanceGRPO, we first optimize the reasoner with GRPO while keeping the generator frozen. We then freeze the resulting reasoner and optimize the generator with DanceGRPO. Thus, no training iteration updates both components. This ordering follows the staged VLM-then-generator pattern of VANS (Cheng et al., 2026). For CoDanceRL, both components remain trainable throughout the same RL run. Each iteration collects one shared nested rollout and then takes one optimizer step for the reasoner and one for the generator from that rollout, as detailed in Appendix A.5.2.

VLM-as-judge evaluations. We use GPT-5.6 Luna with `low` reasoning effort as a held-out judge on Physics-IQ Verified (Rädsch et al., 2026), excluded from reward optimization. For each five-second continuation, the judge receives the best-practice prompt, conditioning image, and four 3×5 comparison grids. These grids align 20 uniformly sampled frames from the generated video and two real takes of the same scene and viewpoint.

Using category-specific criteria, the judge assigns integer scores from 1 to 5 for reference event fidelity e_i, trajectory fidelity t_i, physical plausibility p_i, temporal coherence c_i, and entity/scene consistency s_i. The real takes define the expected event progression while allowing natural variation. We retain one valid judgment per video, with reasons and frame evidence for each rating, and compute

$$q_i = 25\left[\min\left(\frac{e_i + t_i}{2}, \frac{2p_i + c_i + s_i}{4}\right) - 1\right], \qquad J_{\mathrm{VLM}} = \frac{1}{198}\sum_{i=1}^{198} q_i. \tag{14}$$

We report $J_{\mathrm{VLM}} \in [0, 100]$ in Table 1, averaging over 198 videos from 66 scenes and three viewpoints.

For our evaluations on RBench (Deng et al., 2026), we use GPT as the VLM judge and follow the official evaluation protocol and scoring criteria. We evaluate all 650 released inputs, comprising 50 examples in each of the five task categories and 100 in each of the four embodiment categories. Frame preparation follows the released evaluation code. Task evaluation first samples 16 frames uniformly over the video, then selects three frames for spatial relationships and six for each of the other four task categories. Embodiment stability uses the first frame and the frame at approximately 75% of the video, while physical plausibility and task adherence use six-frame grids. We report scores on a 0–100 scale and compute Avg. as the unweighted mean of the five task and four embodiment scores.

Evaluation settings. Tables 5 and 6 report generation settings only for models evaluated by us on Physics-IQ Verified and RBench, respectively. FPS denotes the video frame rate, independently of the number of frames sampled by the evaluator. Cosmos3-Nano, Cosmos3-Nano + SFT, and CoDance use identical video-generation settings within each benchmark, generating five-second videos at 1280×720 resolution and 24 FPS, with classifier-free guidance of 6.0 and a flow shift of 10.0. Each video contains 121 frames, including the initial conditioning frame. For these three configurations, we use 35 denoising steps when evaluating on Physics-IQ Verified and 50 on RBench.

For Physics-IQ Verified, we evaluate all 66 scenarios from three viewpoints, yielding 198 image–prompt inputs. Each input consists of the released switch frame and best-practice prompt. For the

[1] `https://github.com/google-deepmind/physics-IQ-benchmark`.

three Cosmos3-Nano-based configurations, we remove the conditioning frame from each generated video and score the remaining 120 frames, corresponding to five seconds. The official evaluator uses both real takes to estimate physical variance and reports the Verified IQ-Score together with S-IoU, ST-IoU, W-S-IoU, and the normalized MSE score. The MSE score is inverted during normalization, so higher values indicate better performance for all reported Verified metrics.

Table 5: Video-generation settings for models evaluated by us on Physics-IQ Verified. Duration denotes the five-second evaluated continuation.

Model	Resolution	FPS	Duration	Steps	Guidance	Shift
Commercial						
MiniMax-H3 Max	1344×768	24	5 s	–	–	–
Seedance 2.0	1280×720	24	5 s	–	–	–
Open-source						
Wan2.2-I2V-A14B	1280×720	16	5 s	35	3.5	5.0
HunyuanVideo 1.5	848×480	24	5 s	35	6.0	5.0
Cosmos3-Super I2V	832×480	24	5 s	35	6.0	10.0
Physics-aware methods						
PhiZero	896×512	8	5 s	50	–	5.0
Ours and base model						
Cosmos3-Nano	1280×720	24	5 s	35	6.0	10.0
Cosmos3-Nano + SFT	1280×720	24	5 s	35	6.0	10.0
CoDance	1280×720	24	5 s	35	6.0	10.0

Resolution is given as width $\times$ height. Here, s denotes seconds and – indicates an unreported or inapplicable value. Wan2.2-I2V-A14B uses 16 FPS and PhiZero uses 8 FPS, while the other listed models use 24 FPS. PhiZero pads 33 native frames to 40 scoring frames. The three Cosmos3-Nano-based configurations generate 121 frames and yield 120 scoring frames after removing the conditioning frame.

Table 6: Video-generation settings for models evaluated by us on RBench.

Model	Resolution	FPS	Duration	Steps	Guidance	Shift
Commercial						
Seedance 1.5	1280×720	24	5 s	–	–	–
Ours and base model						
Cosmos3-Nano	1280×720	24	5 s	50	6.0	10.0
Cosmos3-Nano + SFT	1280×720	24	5 s	50	6.0	10.0
CoDance	1280×720	24	5 s	50	6.0	10.0

Resolution is given as width $\times$ height. Here, s denotes seconds and – indicates an unreported or inapplicable value. The three Cosmos3-Nano-based configurations generate five-second videos at 24 FPS with 121 frames, including the initial conditioning frame.

For WorldArena, we follow the official action-planning protocol on the RoboTwin 2.0 Clean-50 data, using 40 training episodes and ten held-out test episodes per task. We keep the video model frozen and train a task-specific inverse dynamics model with the official architecture and training procedure. At each replanning step, the inverse dynamics model predicts an action chunk from the video-model features, and the planner refreshes its observation after executing the chunk. We evaluate `adjust_bottle` and `click_bell`, repeating each of the ten test initial conditions ten times for 100 rollouts per task. We use the official task-success criteria and report the success rate for each task and their arithmetic mean.

A.5 Training Details

A.5.1 Supervised Fine-Tuning

We fine-tune the reasoner and generator separately on the dataset in Appendix A.3, then use both checkpoints to initialize coupled reinforcement learning.

Reasoner SFT. The reasoner receives the first frame I and original prompt p and learns to predict the teacher-generated physical rewrite. We minimize the token-level negative log-likelihood with teacher forcing, updating LoRA adapters in the attention and feed-forward projections while keeping the base weights and vision encoder frozen.

Generator SFT. We condition the generator on the first frame I and original prompt p, using the corresponding reference video $V^\star$ as the target. The reference video is encoded into latent space for the flow-matching objective in Section 2. Training updates the video-generation parameters while keeping the language-conditioning tower, vision encoder, and video autoencoder frozen.

A.5.2 Reinforcement Learning

We report the parameter configuration for the full nested rollout protocol in Section 3.1, with $G = 8$ rewrites per input and $l = 8$ videos per rewrite.

Initialization and trainable parameters. Both policies start from the supervised checkpoints described in Appendix A.5.1. The reasoner is the Qwen3-VL 8B component of Cosmos3-Nano. During RL, we continue optimizing its attention and feed-forward LoRA adapters, using rank 16 and scaling parameter $\alpha = 32$. The generator updates its full video-generation parameter set, including the generation blocks, input and output projections, time embeddings, and final normalization. The reasoner's base weights, vision encoder, and video autoencoder remain frozen.

Rollout configuration. With the default batch size of 16 input conditions, each iteration contains 128 prompt rewrites and 1,024 generated videos. The reasoner samples rewrites with temperature 0.7, top-$p = 0.8$, top-$k = 20$, a repetition penalty of 1.0, and a presence penalty of 1.5. The maximum completion length is 8,192 tokens. Each of the eight rewrites conditions a separate group of eight videos. The physical-fidelity and geometric trajectory rewards are converted into video-level and prompt-level advantages using the hierarchical normalization in Section 3.3.

Video sampling. Training videos are generated at 1280×720 resolution and 24 FPS. Their frame counts follow the reference video's temporal profile and satisfy the $4k + 1$ constraint. For RL rollouts, we use 50 denoising steps, classifier-free guidance of 6.0, and a flow-schedule shift of 10.0. Dance-SDE uses a noise coefficient of $\eta = 0.3$. The generator's maximum text-conditioning length is 4,096 tokens.

Policy optimization. The reasoner and generator use separate AdamW optimizers with constant learning rates of 3×10^{-6} and 1×10^{-5}, and weight decay of 10^{-2} and 10^{-4}, respectively. Both optimizers use an Adam numerical-stability constant of 10^{-8} and a maximum gradient norm of 1.0. The PPO clipping thresholds are 0.2 for the reasoner and 10^{-4} for the generator. The reasoner additionally uses a dual-clipping coefficient of 3.0. Its KL penalty is computed against the frozen SFT adapter using the k_3 estimator with coefficient $\beta = 0.03$. The generator follows the objective in Eq. 4. Gradients are accumulated over the complete rollout batch before taking one optimizer step for each policy.

A.5.3 Training Dynamics

Figure 6 presents the mean joint reward and its two components over epochs 0–23, where each epoch comprises one complete nested rollout followed by one optimizer step for each policy. The joint and geometric trajectory rewards rise mainly during the initial epochs and then fluctuate around higher levels. The physical-fidelity reward fluctuates throughout training but ends above its initial value.

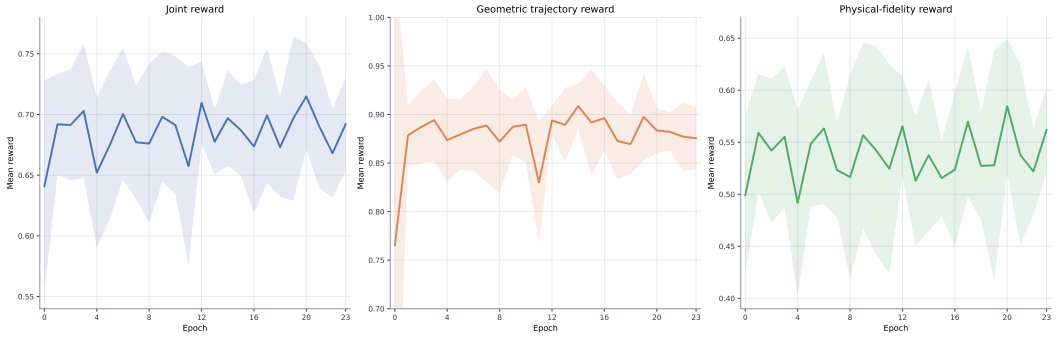


Figure 6: **Training dynamics of CoDanceRL.** Mean joint reward (left), geometric trajectory reward (middle), and physical-fidelity reward (right) over training epochs 0–23. Shaded bands indicate standard deviations.

A.6 Additional Experiments

A.6.1 Fine-Grained VLM-as-Judge Results

Table 7 breaks down the VLM as Judge score on Physics-IQ Verified into five criteria. The overall score follows the per-video aggregation in Equation 14. The judge is excluded from reward optimization; Appendix A.4 describes the evaluation protocol.

Table 7: Fine-grained and overall VLM-as-judge scores on Physics-IQ Verified, sorted by ascending VLM as Judge score. Scores are on a 0–100 scale (higher is better), with the best score in each column in bold.

Model	Reference event fidelity	Trajectory fidelity	Physical plausibility	Temporal coherence	Entity/scene consistency	VLM as Judge
HunyuanVideo-1.5	34.34	23.48	30.68	41.79	45.33	26.70
Cosmos3-Nano	37.25	27.53	40.40	45.45	64.27	31.69
Wan2.2	40.40	28.91	37.25	45.45	59.72	33.18
PhiZero	38.51	30.30	46.97	49.62	73.11	33.90
Cosmos3-Super I2V	42.68	32.70	42.30	51.89	63.64	36.40
Seedance 2 Pro	47.22	37.25	51.14	55.68	73.61	41.89
CoDance	**59.22**	**45.96**	**57.32**	**66.92**	**76.26**	**52.11**

CoDance leads all five components, with an overall score of 52.11 versus 31.69 for Cosmos3-Nano and 41.89 for Seedance 2 Pro. Its largest improvements over both models are in reference event fidelity and temporal coherence. Compared with Seedance 2 Pro, these gains are 12.00 and 11.24 points, respectively, versus 2.65 points for entity/scene consistency. The judge therefore distinguishes these models most strongly through event progression and temporal behavior.

Trajectory fidelity remains the lowest-scoring component for all seven models. CoDance scores 45.96 on this criterion, compared with 76.26 for entity/scene consistency. Under the judge's criteria, matching reference motion remains a common weakness despite the improvements in overall scores.

A.6.2 Training Strategies and Component Contributions

We examine the role of the reasoner and generator through two complementary comparisons. Table 8(a) evaluates training strategies that adapt either or both components, while Table 8(b) recombines trained components at inference time to assess their individual contributions.

Table 8: Comparison of training strategies and inference-time component compositions on Physics-IQ Verified (Rädsch et al., 2026). (a) Independent training configurations for the reasoner and video generator. (b) Base, SFT, and CoDance components are recombined at inference time to assess their individual contributions.

(a) Training configurations

Model	IQ-Score
Cosmos3-Nano (base)	34.8
SFT	38.4
GRPO (w/ SFT) (Shao et al., 2024)	41.5
DanceGRPO (w/ SFT) (Xue et al., 2025b)	39.3
GRPO+DanceGRPO (w/ SFT)	44.1
CoDance (joint training)	**45.4**

(b) Inference-time composition

Reasoner	Generator	IQ-Score
Base	Base	34.8
CoDance	Base	42.4
Base	CoDance	40.5
CoDance	SFT	43.0
SFT	CoDance	43.9
CoDance	**CoDance**	**45.4**

Joint training. All RL configurations in Table 8(a) start from the same component-wise SFT checkpoints and use both the physical-fidelity reward and the geometric trajectory reward introduced in Section 3.2. For both GRPO and DanceGRPO, including their respective stages in the sequential baseline, we use GDPO's reward-decoupled normalization. Each reward is normalized separately before aggregation, and the combined signal is then normalized over the batch. CoDanceRL applies this principle at both the prompt and video levels through the hierarchical construction in Section 3.3.

Reasoner-only GRPO updates the reasoner while keeping the SFT generator frozen, reaching an IQ-Score of 41.5. Generator-only DanceGRPO updates the generator while keeping the SFT reasoner frozen, reaching 39.3. These configurations improve over the SFT initialization (38.4) by 3.1 and 0.9 points, respectively. The sequential GRPO+DanceGRPO baseline first trains the reasoner with the generator frozen, then freezes the resulting reasoner and trains the generator, following the staged pattern of VANS (Cheng et al., 2026). It reaches 44.1, exceeding both single-component configurations, although the reasoner receives no further updates as the generator adapts in the second stage.

CoDanceRL keeps both components trainable throughout the same RL run. Each iteration collects a shared nested rollout and updates both policies using their respective advantages derived from the same generated videos. The reasoner therefore continues to receive feedback from the adapting generator throughout training. CoDanceRL reaches 45.4, exceeding reasoner-only GRPO by 3.9 points, generator-only DanceGRPO by 6.1 points, and the sequential baseline by 1.3 points. These results support the benefit of coupled adaptation when all RL configurations use the same two reward signals and reward-decoupled normalization.

Component contributions. Table 8(b) recombines base, SFT, and CoDance components at inference time without additional training. With the base generator fixed, replacing the base reasoner with the CoDance reasoner raises IQ-Score from 34.8 to 42.4. With the base reasoner fixed, replacing the base generator with the CoDance generator raises the score to 40.5. Thus, each trained component retains a benefit when paired with its unadapted counterpart.

Pairing the CoDance reasoner with the SFT generator yields 43.0, while pairing the SFT reasoner with the CoDance generator yields 43.9. Combining both CoDance components reaches 45.4, improving on these two partial SFT compositions by 2.4 and 1.5 points, respectively. These comparisons hold one CoDance component fixed while varying the other, supporting contributions from both the reasoner's physical guidance and the generator's ability to realize it. The strongest result comes from using the two jointly trained components together.

A.6.3 Combinatorial Motivation for Reward Normalization

Figure 7 compares distinct advantage-group counts in a binary-reward enumeration under three rules. These rules sum rewards before normalization at both levels, apply normalization only at the video level, or apply hierarchical reward normalization. The comparison increases the numbers of prompts and videos per prompt together from 1×1 to 16×16. Beyond the trivial 1×1 case, hierarchical reward normalization yields the largest count among the three rules in both the full-

space and sparse-subspace enumerations shown. The full-space enumeration extends through 4×4, while counts for larger groups are lower bounds obtained from the sparse subspace. A larger count can be interpreted as retaining more reward distinctions within this enumeration, so the comparison serves only as motivation. It does not establish that the resulting gradients are more reliable or that learned video quality improves. The training ablations in Section 4.3 evaluate the effect of reward normalization on IQ-Score.

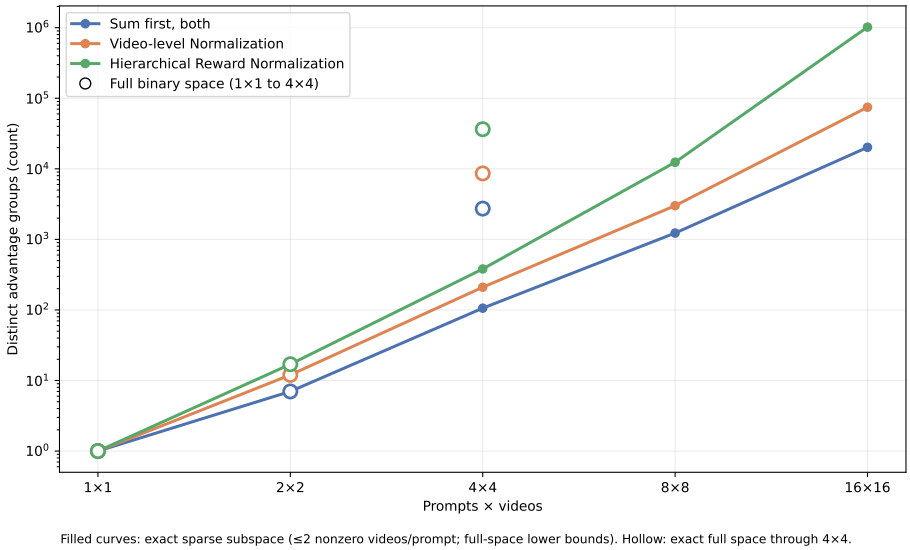
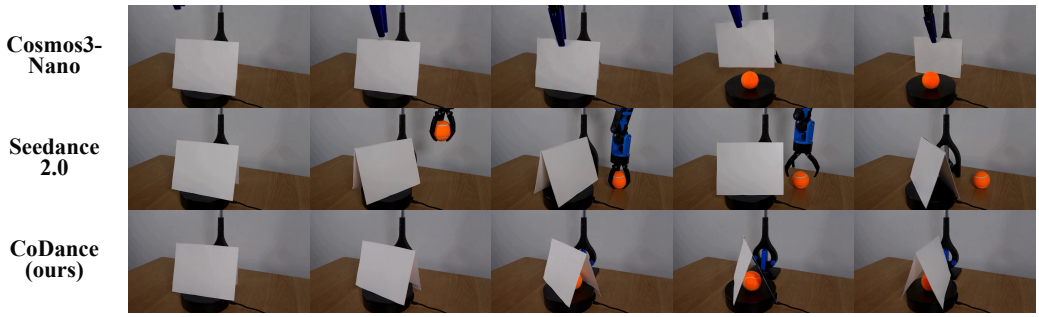


Figure 7: **Combinatorial motivation through advantage-group counts.** The horizontal axis gives the number of prompts and videos per prompt. Filled markers give exact counts in a sparse binary-reward subspace with at most two nonzero videos per prompt, providing lower bounds on full-space counts. Hollow markers give exact full-space counts through 4×4. These counts characterize distinctions among normalized advantage groups and do not measure gradient reliability or generated-video quality.

A.7 More Visualization Results

A.7.1 Physics-IQ Verified

Figures 8–13 provide six additional qualitative comparisons on Physics-IQ Verified. In each figure, columns show frames in temporal order, rows compare Cosmos3-Nano, Seedance 2.0, and CoDance, and the evaluated prompt appears below the frames.



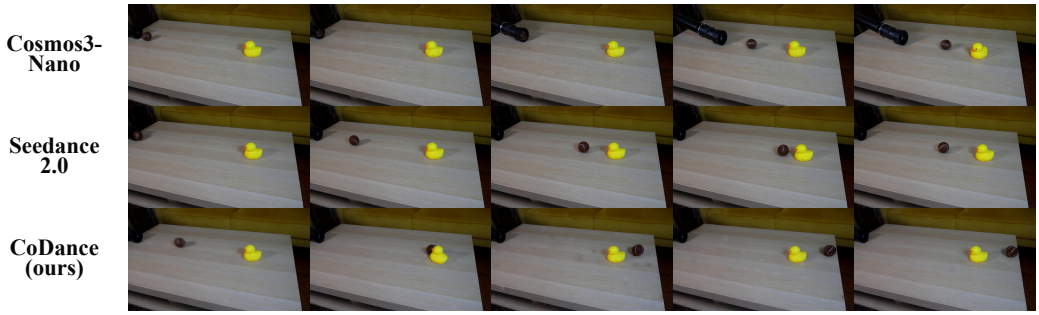
A ball is placed behind rotating cardstock.

Figure 8: **Additional Physics-IQ Verified comparison of a ball behind rotating cardstock.**



A ball rolls into a rubber duck (center view).

Figure 9: **Additional Physics-IQ Verified comparison of a ball colliding with a rubber duck (center view).**



A ball rolls into a rubber duck (right view).

Figure 10: **Additional Physics-IQ Verified comparison of a ball colliding with a rubber duck (right view).**

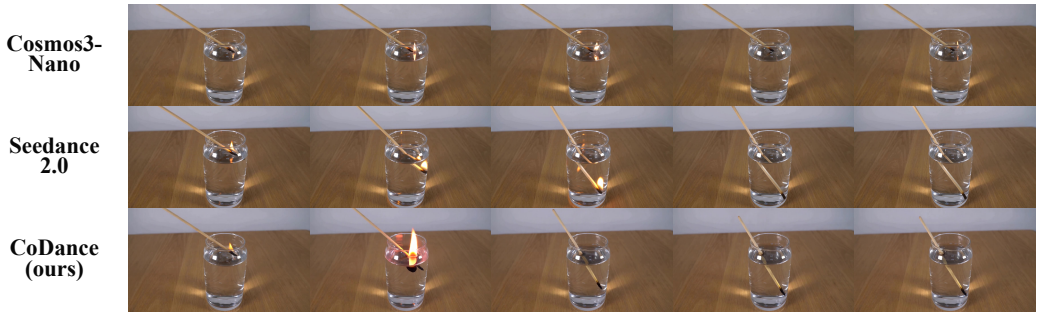


Figure 11: **Additional Physics-IQ Verified comparison of a lit match lowered into water.**

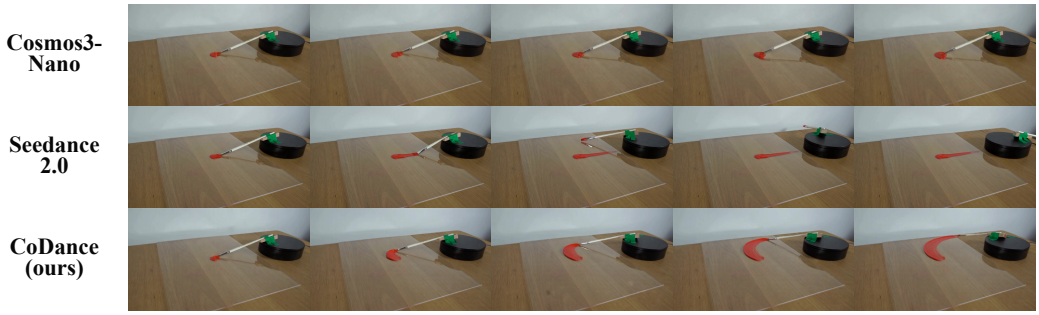


Figure 12: **Additional Physics-IQ Verified comparison of a rotating brush dragging paint across acrylic.**

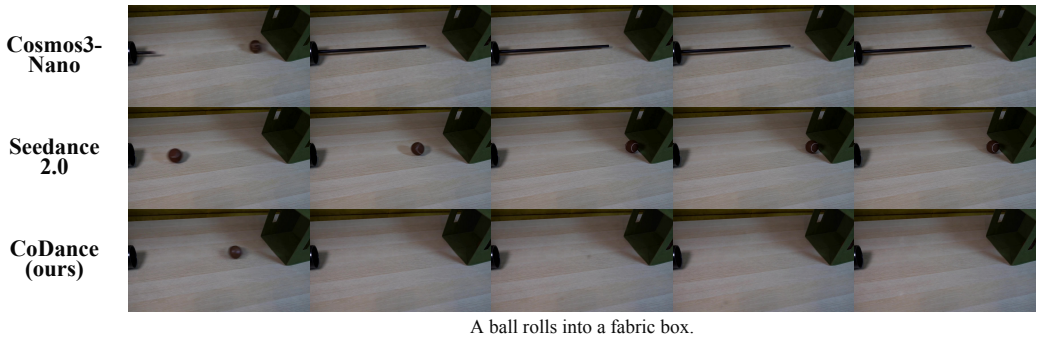
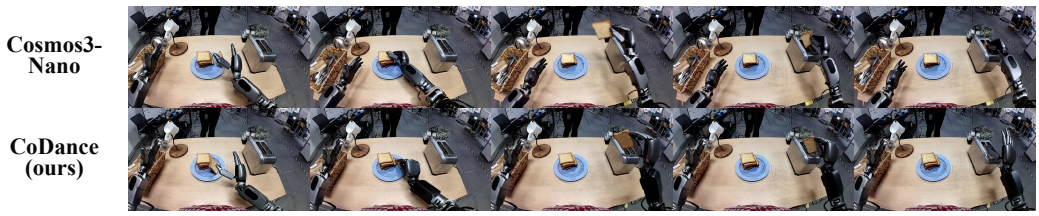


Figure 13: **Additional Physics-IQ Verified comparison of a ball rolling into a fabric box.**

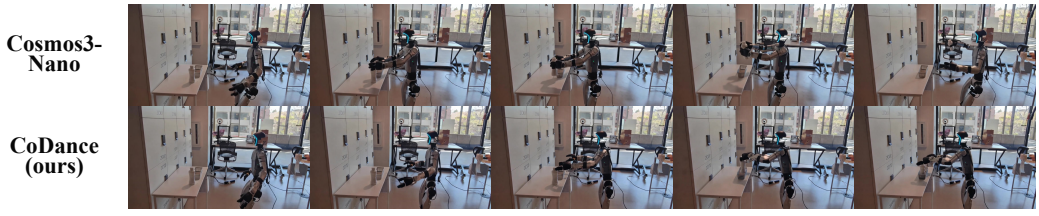
A.7.2 RBENCH

Figures 14–21 provide eight additional qualitative comparisons on RBench. In each figure, columns show frames in temporal order, rows compare Cosmos3-Nano and CoDance, and the evaluated prompt appears below the frames.



The humanoid robot inserts a slice of bread into the toaster.

Figure 14: **Additional RBench comparison for inserting a slice of bread into a toaster.**



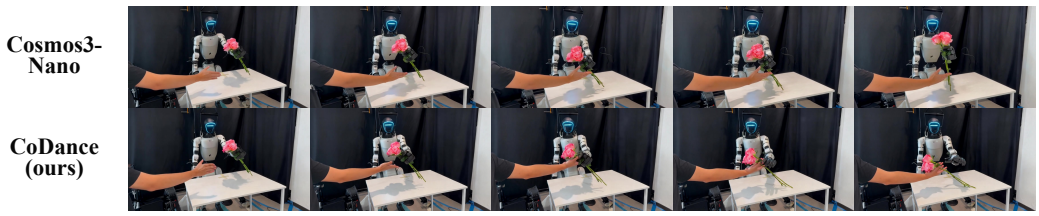
The humanoid robot places the bottle into the mesh basket.

Figure 15: **Additional RBench comparison for placing a bottle into a mesh basket.**



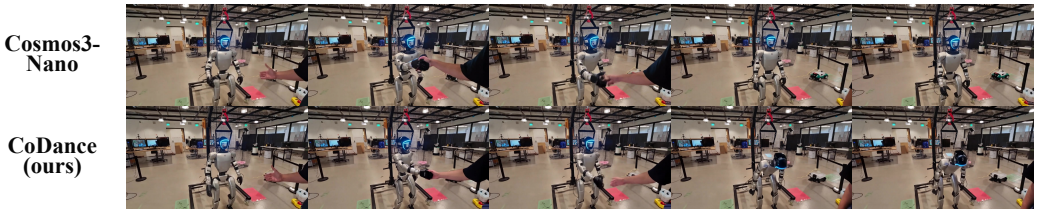
The robotic hand passes the red coffee cup to the woman.

Figure 16: **Additional RBench comparison for passing a red coffee cup to a person.**



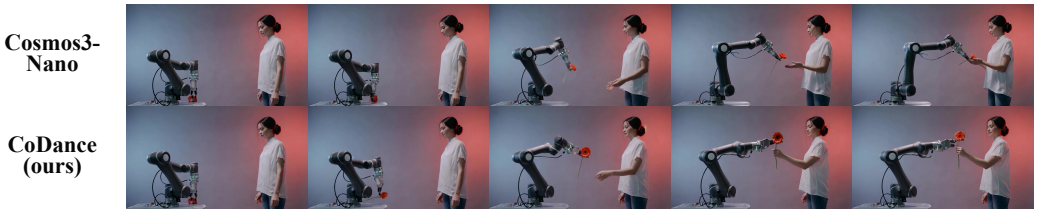
The humanoid robot hands the flower bouquet to the person.

Figure 17: **Additional RBench comparison for handing a flower bouquet to a person.**



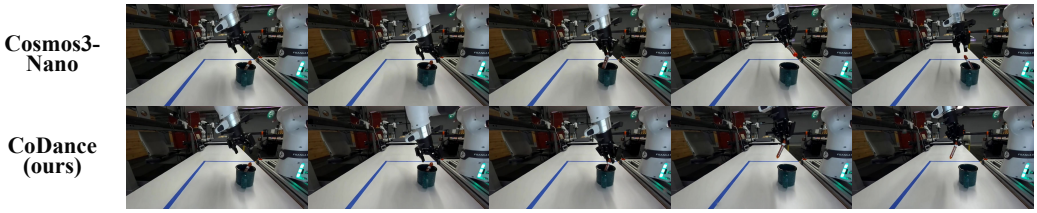
The humanoid robot shakes hands, then bows politely.

Figure 18: **Additional RBench comparison for shaking hands and then bowing.**



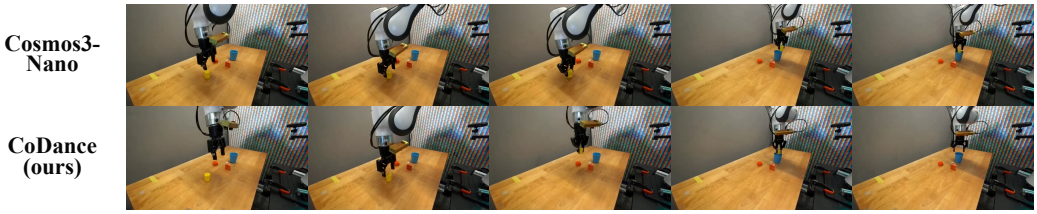
The gripper hands the red-orange flower to the person.

Figure 19: **Additional RBench comparison for handing a red-orange flower to a person.**



The gripper removes the red marker from the green mug.

Figure 20: **Additional RBench comparison for removing a red marker from a green mug.**



The gripper places the yellow toy block into the blue cup.

Figure 21: **Additional RBench comparison for placing a yellow toy block into a blue cup.**