WORLD MODELS NEED A PHYSICAL REGISTRY: LEARNING INTERNAL PREDICTIVE STATES FOR PHYSICALLY PLAUSIBLE WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Video world models must generate physically plausible futures, not just realistic frames. Existing approaches typically use semantic, physical, or predictive information to guide the visual generation stream, leaving a fundamental question: **where inside the generator should predictions of scene evolution live?** Standard video tokens jointly encode appearance and dynamics, but provide no distinct state for learning and reading predictive dynamics. We answer this question with the **Physical Registry**: temporally grounded registry slots inside a video DiT that are initialized from an observation-conditioned forecast, co-evolve with video tokens, and receive future-representation supervision, making predictive dynamics an explicit, readable state of the generator. In a controlled comparison of predictive-state interfaces on Physics-IQ Verified, Physical Registry improves vanilla fine-tuning by $9.9/8.7$ points on I2V/V2V, the largest gain among the interfaces we compare. Across all Cosmos and WAN backbones, Physical Registry improves Physics-IQ Verified scores. With only 0.45% additional parameters, it enables Cosmos 3 Nano 16B to surpass the published I2V score of Cosmos 3 Super 64B (43.56 vs. 42.7). On WorldModelBench, it also improves physical adherence and the total score for all four evaluated backbones. Because the registry becomes candidate-specific while remaining readable, the *Physical Registry Selector (PRS)* ranks emerging futures by how far their intermediate registry readouts depart from the shared forecast before full denoising is complete. Across four backbones, this selection improves I2V scores by 5.6–15.9% without completed-video external scoring. Together, these results support a broader principle: **dynamics prediction should live as an explicit model state that co-evolves with the video representation.**

1 INTRODUCTION

Video generators are increasingly used as world models, i.e., environments for training agents (Ha & Schmidhuber, 2018; Bruce et al., 2024; NVIDIA et al., 2025b). From this perspective, a generated video should contain not only high-quality visuals, but also physically plausible phenomena such as solid mechanics, fluid dynamics, optics, thermodynamics, and magnetism (Motamed et al., 2026). However, recent video models still often fail to follow them (Li et al., 2025a; Motamed et al., 2026).

To address this, exploiting physical priors has made some progress by supplying explicit cues about scene evolution, such as motion, force, or physical control signals (Yang et al., 2025a; Gillman et al., 2025; Wang et al., 2025a). However, it depends on availability of the controls and corresponding supervision. On the other hand, supervision for aligning the intermediate features in the video models to pre-trained representations is readily available (Yu et al., 2025; Zhang et al., 2025; Hwang et al., 2025; Bhowmik et al., 2026). However, their appearance and dynamics can remain entangled (Bhowmik et al., 2026). This context motivates us to ask a more basic question: **How and where can we internalize scene evolution using representation alignment?**

We introduce the *Physical Registry*, dedicated registry slots containing scene evolution inside video DiTs. They are encouraged to be aligned with representations of future frames to be generated, which are predicted from a source frame. In detail, the Physical Registry of the first DiT layer starts from JEPA features of the source and future frames via the JEPA encoder and encoder–

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

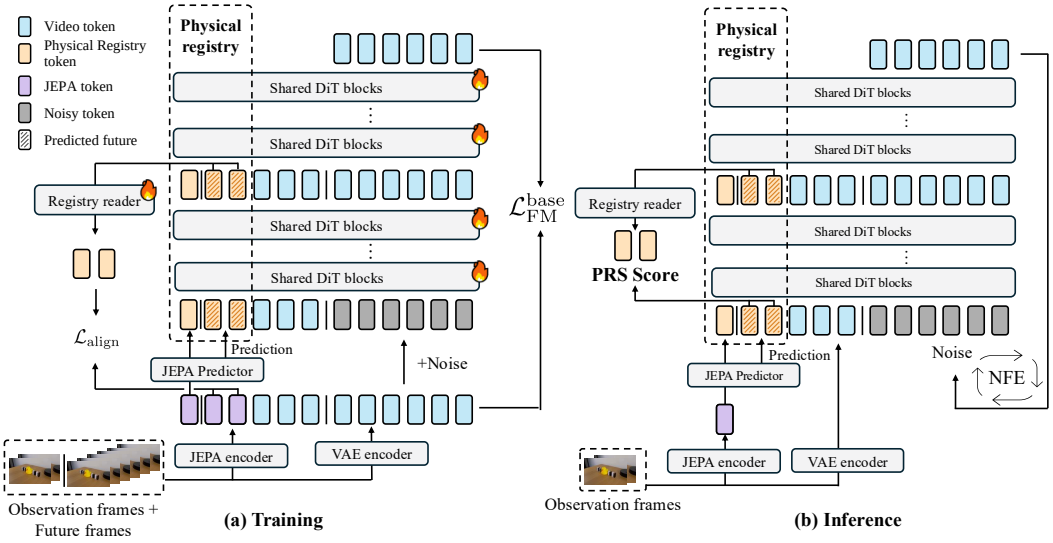


Figure 1: **Physical Registry during training and inference.** The Physical Registry gives predicted future dynamics a dedicated, readable state inside the video DiT. (a) Training: forecast-initialized registry tokens co-evolve with video tokens and receive predictive supervision. (b) Inference: the same registry provides readouts for PRS (Section 3.6).

predictor, respectively. Then, a readout layer aligns the registries for future frames to the original representations of the future frames from the JEPA encoder. As a whole, the registries and video tokens are designed to co-evolve through the transformer blocks such that the registries can estimate the original representations from the predicted representations. We choose JEPA as the foundation of the Physical Registry because its predictive representation learning paradigm naturally aligns with our goal of modeling future physical states. JEPA's predictive features provide a compact representation of how the scene is expected to evolve over time, making them particularly suitable for encoding physical dynamics and properties. Previous works have also demonstrated the effectiveness of JEPA-based representations for capturing physical properties (Garrido et al., 2025; Joseph et al., 2026).

We demonstrate that it is crucial to use all three components: the Physical Registry, alignment to the original encoded future frames, and the readout layer. Notably, the Physical Registry yields the largest gain in physical plausibility, and shows that this gain generalizes to Cosmos and WAN backbones. Furthermore, the readouts can be used to choose the best sample among denoising processes from different initial noises, namely *Physical Registry Selector (PRS)*. It works in the middle of denoising processes instead of completely generated videos (Yuan et al., 2026; Internò et al., 2026).

Our contributions are threefold:

- The Physical Registry conveys future dynamics from the initial frames in a video generator.
- The Physical Registry provides an internal selection signal across different initial noises before fully generating videos.
- The Physical Registry outperforms other interfaces for predictive representations.

2 Related Work

Physical plausibility in video generation. Prior work improves physical plausibility or controllability by conditioning video generation on motion, force, trajectory, or language-derived physical signals (Li et al., 2025b; Gillman et al., 2025; Wang et al., 2025a; Yang et al., 2025a; Pathak et al., 2026). These approaches express the intended evolution through an external control signal, whereas Physical Registry represents predicted future dynamics as an internal state of the video generator.

Predictive representations inside video generators. I-JEPA learns predictive representations from visible context, with V-JEPA and V-JEPA 2 extending this principle to video (Assran et al.,

2023; Bardes et al., 2024; Assran et al., 2025). JEPA representations capture dynamics-relevant structure, including sensitivity to intuitive-physics violations and information about motion and physical plausibility (Garrido et al., 2025; Joseph et al., 2026). Generative models typically use such pretrained representations in one of two ways. REPA and its video extensions align them with intermediate generative features (Yu et al., 2025; Zhang et al., 2025; Hwang et al., 2025; Bhowmik et al., 2026; Lian et al., 2026), while ReDi (Kouzelis et al., 2025) and PHANTOM (Shen et al., 2026) jointly generate visual and representation states. Register tokens offer a different primitive: Vision Transformer registers provide additional tokens for internal computation (Darcet et al., 2024), and HyperDiT (He et al., 2026) uses representation-aligned semantic registers for image generation. Physical Registry brings these lines together by holding forecast-initialized predictive representations in dedicated, temporally grounded registry slots within the video backbone and supervising them toward future representations.

Inference-time guidance and candidate selection. WMReward uses V-JEPA 2 as a physics reward for inference-time guidance and best-of-N selection (Yuan et al., 2026), while GeoPhys uses an external physical-plausibility verifier to rerank generated candidates (Internò et al., 2026). PRS instead ranks candidates from the generator's own intermediate registry states, without an external reward model or verifier.

3 METHOD

We augment flow-matching video DiTs with Physical Registry (Sec. 3.1) with forecast-initialized registry states (Sec. 3.2) that evolve with the video representations (Sec. 3.3), and align them toward the target future (Sec. 3.4). Then we introduce our training objective (Sec. 3.5) and our inference technique (Sec. 3.6).

3.1 CONDITIONAL FLOW MATCHING

Let X_0 denote a clean video latent, $\epsilon \sim \mathcal{N}(0, I)$, and $t \sim \mathcal{U}(0, 1)$. With condition c, rectified flow (Liu et al., 2023) uses linear interpolation path, velocity target, and objective:

$$X_t = (1-t)X_0 + t\epsilon, \qquad V^\star = \epsilon - X_0, \tag{1}$$

$$\mathcal{L}_{\mathrm{FM}}^{\mathrm{base}} = \mathbb{E}_{X_0,\epsilon,t}\left[\|V_\theta(X_t, t \mid c) - V^\star\|_2^2\right], \text{respectively.} \tag{2}$$

Below, we augment DiTs with the Physical Registry using the same formulation.

3.2 PHYSICAL REGISTRY

Physical Registry represents the forecast in a fixed set of dedicated registry slots that carry predictive features alongside the video latent rather than within it. Let $X_t \in \mathbb{R}^{B\times C\times T_v\times H\times W}$ denote the noisy video latent. We use the frozen JEPA encoder E and predictor P. The encoder receives a full video and produces JEPA features for individual frames. The predictor receives the JEPA feature of the first frame and produces future representations in the shared representation space(Bardes et al., 2024; Assran et al., 2025). Given an initial observed frame x, we retain the first encoder-derived feature as an observed context and use the predictor to forecast the remaining features:

$$\begin{aligned} H^{\mathrm{enc}} &= E(x), \\ J^{\mathrm{cond}} &= [H^{\mathrm{enc}}; P(H^{\mathrm{enc}}; n_{\mathrm{context}} = 1)] \in \mathcal{J}^{T_r}. \end{aligned} \tag{3}$$

We define J^{cond} to be the *forecast*, with $T_r = 8$: one encoder-derived feature for the observed context and seven predictor-generated features for the subsequent future temporal positions. The forecast occupies T_r positions, called *registry slots*, placed after the video tokens. Each slot is a spatial grid of *registry tokens* which are the units for attention, whereas slots correspond to frames.

A learned spatio-temporal adapter G_ϕ (Appendix A) maps the forecast into the video latent's shape and coordinates. At each denoising step t, the same forecast initializes the registry, which is concatenated with the noisy video latent along the temporal dimension:

$$R_t^{(0)} = G_\phi\big(J^{\mathrm{cond}}\big) \in \mathbb{R}^{B\times C\times T_r\times H\times W}, \qquad Z_t^{(0)} = [X_t; R_t^{(0)}] \in \mathbb{R}^{B\times C\times (T_v+T_r)\times H\times W}. \tag{4}$$

3.3 REGISTRY COUPLING AT EACH DENOISING STEP

Coupling places the registry on the video's time axis and within the same computation as the video tokens. Each registry slot takes the video RoPE coordinate of its aligned time, and after patchification the two token types remain distinct while being jointly processed through shared bidirectional self-attention. Let $u_s \in [0, 1)$ denote the normalized clip time associated with registry slot s in J^{cond}, and let ρ_v map this normalized time to the corresponding video-latent RoPE coordinate. Physical Registry assigns

$$\tau_s^{\mathrm{reg}} = \rho_v(u_s), \qquad s \in \{1, \dots, T_r\}. \tag{5}$$

For uniformly spaced registry slots, $u_s = (s-1)/T_r$. Writing $X_t^{(\ell)}$ and $R_t^{(\ell)}$ for the video and registry states at DiT layer ℓ, respectively,

$$\left(X_t^{(\ell+1)}, R_t^{(\ell+1)}\right) = \mathrm{DiT}_\ell\left(X_t^{(\ell)}, R_t^{(\ell)}; c_{\mathrm{txt}}, t\right). \tag{6}$$

We call $R_t^{(\ell)}$ the *registry state* at layer ℓ: initialized from the forecast, it becomes after the first layer a candidate-conditioned function of it.

3.4 PREDICTIVE SUPERVISION OF THE REGISTRY

We supervise the registry toward representations of the realized future. Conditioning and target features play complementary roles: J^{cond} is an observation-conditioned forecast that initializes $R_t^{(0)}$, while Z^{tgt} is the target encoding of the corresponding full training video v:

$$Z^{\mathrm{tgt}} = E^{\mathrm{tgt}}(v) = \left(Z_1^{\mathrm{tgt}}, \dots, Z_{T_r}^{\mathrm{tgt}}\right), \qquad Z_s^{\mathrm{tgt}} \in \mathcal{J}. \tag{7}$$

Here Z^{tgt} is detached. Because J^{cond} depends only on the observation, the registry starts each denoising step from an observation-only prediction. As the registry exchanges information with the evolving video tokens, predictive supervision encourages this state to align with the representation of the future realized in the training video. We apply the supervision at an intermediate DiT block so that subsequent blocks can continue joint video–registry processing. Let $H_{\mathrm{reg},t,s}$ denote the hidden feature grid of registry slot s at the supervised layer and denoising step t. A readout q_ψ maps each registry state to a slot-level vector in the JEPA feature space:

$$\widehat{Z}_{t,s} = q_\psi(H_{\mathrm{reg},t,s}). \tag{8}$$

Writing $\overline{Z}_s^{\mathrm{tgt}}$ for the corresponding slot-level target, we optimize

$$\mathcal{L}_{\mathrm{align}} = -\frac{1}{T_r}\sum_s \cos\left(\widehat{Z}_{t,s}, \overline{Z}_s^{\mathrm{tgt}}\right). \tag{9}$$

Appendix A specifies the readout and target construction. After training, the same readout q_ψ is frozen and reused for selection in Sec. 3.6.

3.5 TRAINING OBJECTIVE

Given the augmented sequence $Z_t^{(0)} = [X_t; R_t^{(0)}]$, the DiT jointly processes the video and registry tokens. We denote by $V_\theta^X(Z_t^{(0)}, t, c_{\mathrm{txt}})$ the velocity prediction from the video tokens. The flow-matching objective is

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}_{X_0,\epsilon,t}\left[\left\| V_\theta^X\left(Z_t^{(0)}, t, c_{\mathrm{txt}}\right) - V^\star \right\|_2^2\right]. \tag{10}$$

Both terms optimize the shared DiT:

$$\mathcal{L} = \mathcal{L}_{\mathrm{FM}} + \lambda_{\mathrm{align}}\mathcal{L}_{\mathrm{align}}. \tag{11}$$

The video tokens receive the generation objective and the registry slots receive predictive supervision through the same DiT blocks, so the registry carries future-referenced information while the backbone continues to learn generation.

3.6 THE PHYSICAL REGISTRY SELECTOR

Because the registry remains readable during denoising, candidates can be compared on their emerging future dynamics before generation is complete. For a fixed observation, all candidates share the

same observation-conditioned forecast, while interaction with their evolving video tokens makes the registry states candidate-specific. Departure from this shared forecast therefore captures the trajectory-specific update incorporated into each registry. The *Physical Registry Selector (PRS)* ranks candidates by this departure and continues denoising only the selected trajectory for the remaining steps.

For input x, let the N candidate trajectories be $\mathcal{C}(x) = \{Y^{(n)}\}_{n=1}^{N}$, sharing the forecast J^{cond}. After K denoising steps, the PRS compares each candidate's registry readouts with this shared forecast (Figure 2) and continues only the highest-scoring candidate from its current latent for the remaining $D - K$ steps, where D is the total number of denoising steps.

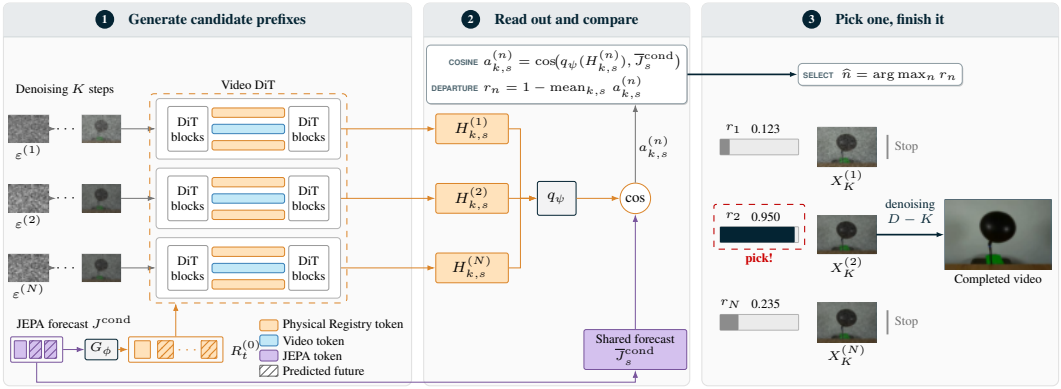


Figure 2: **Three-stage prefix selection with PRS.** Candidates share the same forecast initialization, are ranked from registry readouts over a K-step prefix, and only the selected trajectory continues for the remaining $D - K$ steps. The completed-video frame is illustrative.

The readout q_ψ maps the registry hidden state $H_{k,s}^{(n)}$ into the JEPA feature space. Let $\overline{J}_s^{\text{cond}}$ denote the corresponding slot-level shared-forecast reference. The *forecast-departure score* for candidate n, accumulated over K denoising steps and a registry subset $\mathcal{S}$, is

$$r_{\text{depart}}^{(n)} = 1 - \frac{1}{K|\mathcal{S}|} \sum_{k=1}^{K} \sum_{s \in \mathcal{S}} \cos\left(q_\psi\left(H_{k,s}^{(n)}\right), \overline{J}_s^{\text{cond}}\right),$$
$$\hat{n} = \arg\max_{n \in \{1,\dots,N\}} r_{\text{depart}}^{(n)}. \tag{12}$$

The score measures angular departure from the shared forecast. Appendix F.2 empirically supports this ranking direction and analyzes its robustness.

4 Experiments

4.1 Experimental Details

Benchmarks and tasks. We evaluate on Physics-IQ Verified (Rädsch et al., 2026), which scores physical plausibility against recorded futures, and on WorldModelBench (Li et al., 2025a), which scores instruction following, physical adherence, and common sense over its full 350-prompt set. Metric definitions are in Appendix A.

Multi-backbone evaluation. We use Cosmos and WAN 2.2 as two backbone families: Cosmos is a widely used world-model backbone that supports both I2V and V2V continuation, while WAN 2.2 is a widely adopted video-generation backbone in recent world-model research. Since WAN does not provide a V2V setting, we evaluate it on I2V only. Models are fine-tuned on the Enlarge dataset, which combines physics-oriented, web, simulation, and robot videos (Wang et al., 2025b; 2023; NVIDIA, 2026a;b) while excluding Physics-IQ videos (Appendix A). Each backbone–Physical Registry pair shares the task, conditioning input, sampling protocol, and evaluator.

Candidate selection. Each input uses the same ten-candidate seed pool. Plain averages the Physics-IQ Verified score over the pool, whereas WMReward (completed-video external scoring) and PRS

(intermediate registry readouts) each select one candidate. Where a published leaderboard baseline exists, Plain reports the published value, while selection still uses our ten-candidate pool. Additional selection details are provided in Appendix F.1.

4.2 Controlled Comparison of Predictive-State Interfaces

We first ask how predictive representations should be incorporated into the video generator: as external conditioning, as supervision on the video stream, through joint generation, or as a dedicated predictive state. For this controlled comparison, we use the 66 Physics-IQ scenarios, each observed from left, center, and right viewpoints. Each variant is adapted on the 66 center-view videos and evaluated on the 132 held-out left and right views using Physics-IQ Verified. All variants share the same pretrained video backbone, adaptation data, and training budget, and all JEPA-based variants use the same frozen JEPA source.

Table 1: **Controlled comparison of predictive-state interfaces.** Interface mechanisms and Physics-IQ Verified scores after center-view training and left/right evaluation, with std over three generation seeds and Δ relative to vanilla fine-tuning. Registry only sets $\lambda_{\text{align}} = 0$ in Eq. (11).

Variant	Mechanism	Registry	Alignment	Readout	I2V	Δ	V2V	Δ
Vanilla fine-tuning	–	×	×	×	41.3 ± 2.1	–	46.4 ± 1.0	–
REPA (Yu et al., 2025)	Feature alignment	×	✓	×	41.8 ± 1.7	+0.5	46.9 ± 0.9	+0.5
Cross-attention (Peebles & Xie, 2023)	External context	×	×	×	43.1 ± 1.9	+1.8	49.2 ± 1.2	+2.8
AdaLN (Peebles & Xie, 2023)	Feature modulation	×	×	×	45.1 ± 0.8	+3.8	48.3 ± 0.5	+1.9
Joint generation (Kouzelis et al., 2025)	Joint denoising	×	✓	×	45.1 ± 1.8	+3.8	48.0 ± 0.4	+1.6
Registry only	**Dedicated predictive state**	✓	×	✓	45.0 ± 1.3	+3.7	50.3 ± 0.9	+3.9
Physical Registry	**Dedicated predictive state**	✓	✓	✓	**51.2** ± 1.5	**+9.9**	**55.1** ± 1.1	**+8.7**

Table 1 summarizes the interfaces: cross-attention and AdaLN condition or modulate the generator, REPA aligns the video-token stream, and joint generation denoises video and future JEPA features together, whereas registry variants keep dedicated, readable predictive states (Appendix A).

Alignment is most effective on a dedicated registry state. Table 1 isolates the alignment term in the full objective (Eq. (11)). Removing it from the same registry (Registry only) lowers I2V/V2V from $51.2/55.1$ to $45.0/50.3$, whereas REPA gains only $+0.5/+0.5$ over its no-alignment control. Predictive supervision thus yields a larger gain on the forecast-initialized registry than on the video-token stream when each is measured against its own no-alignment control.

4.3 Fine-Tuning on the Enlarge Dataset across Video Backbones

Physical Registry improves Plain generation across backbones. To test whether these gains extend across video backbones, Table 2 compares each evaluated backbone with its Physical Registry counterpart, while also reporting additional reference models. Under the expanded-prompt configuration, which uses both prompt expansion and the corresponding generation settings described in Appendix A, Physical Registry improves Cosmos 3 Nano 16B from $37.3/46.45$ to $43.56/49.54$ on I2V/V2V, exceeding the Cosmos 3 Super 64B reference ($42.7/48.02$) while adding only 0.45% parameters (Table 6). In the WAN TI2V-5B matched-budget control, vanilla fine-tuning on the Enlarge dataset stays near pretrained performance while Physical Registry improves (Appendix B). Figure 3 shows matched examples across five physics categories in which Physical Registry better realizes the intended physical event.

Physical adherence improves across world-model benchmarks. On WorldModelBench, Physical Registry improves physical adherence and the total score for all four backbones, with common sense improved or maintained, although instruction following decreases on Cosmos v2.5 and WAN TI2V-5B (Table 3; full leaderboard in Table 10, examples in Appendix I.2). On EWMBench (Hu et al., 2025), which compares rollouts with recorded robot manipulation episodes, Physical Registry improves Cosmos 3 Nano's scene consistency and all three motion metrics, most in trajectory agreement (Appendix E).

Table 2: **Physics-IQ Verified.** Plain averages our candidate pool, or gives the published value where marked (*). Δ is the gain over the backbone above, green parentheses the PRS gain over Plain. Bold and underline mark the best and second-best Plain in each task. Gray marks unavailable entries, dashes unreported ones.

Model	I2V			V2V		
	Plain	Δ	PRS	Plain	Δ	PRS
Cosmos 3 Nano 16B (NVIDIA et al., 2026)	30.18 ± 0.84	–		38.78 ± 0.78	–	
+ Physical Registry	36.01 ± 0.95	+5.83	40.27 (+4.26)	45.12 ± 0.94	+6.34	46.78 (+1.66)
Cosmos 3 Nano 16B† (NVIDIA et al., 2026)	37.3 ± 0.9*	–		46.45 ± 0.76	–	
+ Physical Registry†	**43.56** ± 0.95	+6.26	45.81 (+2.25)	**49.54** ± 0.63	+3.09	52.06 (+2.52)
Cosmos v2.5 2B (NVIDIA et al., 2025b)	30.65 ± 0.37	–		37.40 ± 0.53	–	
+ Physical Registry	33.71 ± 1.11	+3.06	39.08 (+5.37)	42.33 ± 1.90	+4.93	47.67 (+5.34)
WAN 2.2 I2V-A14B (Wan Team, 2025)	32.2 ± 0.6*	–				
+ Physical Registry	38.44 ± 1.06	+6.24	40.58 (+2.14)			
WAN 2.2 TI2V-5B (Wan Team, 2025)	27.7 ± 0.9*	–				
+ Physical Registry	29.24 ± 1.30	+1.54	32.03 (+2.79)			
Cosmos 3 Super 64B† (NVIDIA et al., 2026)	<u>42.7</u> ± 0.8*	–		48.02 ± 0.75	–	
MiniMax H3 (MiniMax, 2026)	39.8 ± 0.3*	–		37.71 ± 1.75	–	
Grok Imagine Video (Anates Labs, 2026)	34.8 ± 0.6*	–				
HunyuanVideo 1.5 (Rädsch et al., 2026)	33.4 ± 0.8*	–				
Cosmos 3 Edge† (NVIDIA et al., 2026)	32.7 ± 1.1*	–		33.29 ± 0.78	–	
CogVideoX-5B (Yang et al., 2025b)	31.8 ± 1.5*	–				
Kandinsky-WM 1.0 (Anates Labs, 2026)	30.8 ± 0.9*	–				
Magi-1 (op) (Physics-IQ Benchmark, 2026)	30.2 ± 1.1*	–		<u>48.4</u> ± 1.1*	–	
Sora 2 (Rädsch et al., 2026)	26.5 ± 0.8*	–				
P-Video (Rädsch et al., 2026)	25.3 ± 1.8*	–				

† Expanded prompt: Claude Opus 5 converts the image-conditioned OP description to NVIDIA's official I2V JSON format; 50 steps, CFG 3.0, shift 4.75 (Appendix A). * Published Physics-IQ Verified reference; unmarked estimates are our runs.



Figure 3: **Qualitative comparisons across five Physics-IQ categories.** Top: Cosmos 3 Nano 16B backbone; bottom: Physical Registry, with matched conditioning and eight equally spaced frames. Physical Registry better realizes the intended physical event: ball separation after collision, liquid motion in the inverted jar, persistence of the ball and its reflection, preservation of the paper under rising smoke, and wrench attraction toward the magnet. More examples: Figures 16 to 18.

Table 3: **WorldModelBench results.** Parentheses show Physical Registry-minus-backbone changes computed before rounding. * marks published leaderboard scores under their original generation settings (Li et al., 2025a). Full leaderboard results are in Table 10.

Model	Instruction following (0–3)	Physical adherence (0–5)	Common sense (0–2)	Total (0–10)
Cosmos 3 Nano 16B (NVIDIA et al., 2026)	2.20 ± 0.02	4.50 ± 0.08	1.43 ± 0.06	8.13 ± 0.13
+ Physical Registry	2.20 ± 0.00 (+0.001)	**4.82** ± 0.03 (+0.32)	1.81 ± 0.02 (+0.38)	8.83 ± 0.04 (+0.69)
Cosmos v2.5 2B (NVIDIA et al., 2025b)	2.14 ± 0.03	4.43 ± 0.07	1.55 ± 0.05	8.12 ± 0.14
+ Physical Registry	2.07 ± 0.01 (-0.07)	4.54 ± 0.03 (+0.11)	1.58 ± 0.04 (+0.03)	8.19 ± 0.01 (+0.07)
WAN 2.2 I2V-A14B (Wan Team, 2025)	2.23 ± 0.03	4.45 ± 0.04	1.60 ± 0.03	8.28 ± 0.02
+ Physical Registry	2.30 ± 0.04 (+0.07)	4.53 ± 0.04 (+0.07)	1.60 ± 0.02 (+0.004)	8.43 ± 0.05 (+0.15)
WAN 2.2 TI2V-5B (Wan Team, 2025)	**2.33** ± 0.01	4.35 ± 0.10	1.57 ± 0.04	8.24 ± 0.14
+ Physical Registry	2.24 ± 0.05 (-0.09)	4.50 ± 0.05 (+0.14)	1.68 ± 0.06 (+0.11)	8.41 ± 0.11 (+0.17)
KLING (Kuaishou Technology, 2024)	2.32*	4.81*	**1.95***	**9.10***
Luma (Luma AI, 2024)	1.98*	4.49*	1.76*	8.24*
OpenSoraPlan-T2V (Lin et al., 2024)	1.72*	4.64*	1.68*	8.04*
Mochi 1 (Genmo Team, 2024)	2.06*	4.39*	1.45*	7.91*
CogVideoX-I2V (Yang et al., 2025b)	1.78*	4.17*	1.12*	7.08*

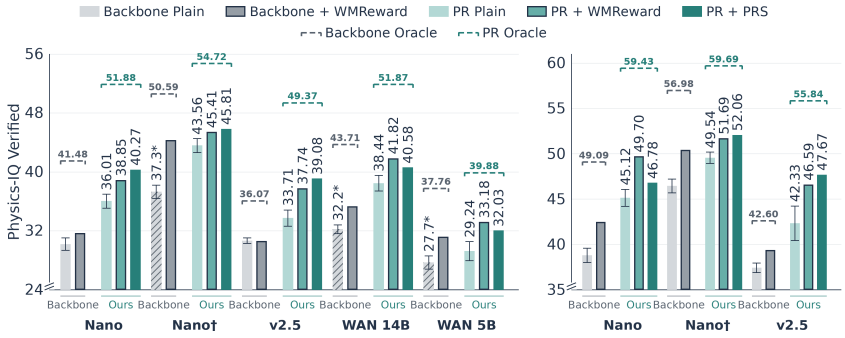


Figure 4: **Selection comparison on Physics-IQ Verified.** (left) I2V and (right) V2V. Dashed caps mark Oracle scores and hatched bars denote published scores.

4.4 Selection from Registry Readouts

Figure 4 compares PRS with Plain generation and completed-video WMReward. Physical Registry improves over its backbone under the same WMReward selector in all eight configurations, and PRS on Physical Registry exceeds WMReward on the corresponding backbone, improving I2V scores over Plain by 5.6%–15.9% across the four default backbone configurations. Unlike WMReward, PRS selects from intermediate registry readouts without completed-video external scoring.

On Cosmos v2.5, PRS after 10 of 35 denoising steps reaches 38.36/46.68 on I2V/V2V, compared with WMReward's 37.74/46.59. Continuing only the selected trajectory reduces the serial denoising budget from 350 to 125 candidate-steps, a 2.8× reduction. Appendices F.1 and F.3 give the full selection results and the early-selection analysis.

4.5 Ablation of the Physical Registry

To separate the effects of forecast content and predictive supervision, Table 4(a) removes or perturbs the forecast with the registry interface fixed, and Table 4(b) compares learnable registers, joint generation, forecast-initialized states, and the full registry in the controlled held-view setting.

Generation depends on the forecast content. In Table 4(a), zeroing the projected forecast drops Nano from 36.01 to 16.78 on I2V, whereas Gaussian perturbation ($\sigma = 0.5$) changes it only slightly, and WAN 14B shows the same trend. The registry is thus highly sensitive to removing the forecast yet robust to moderate perturbation. In a motion-transfer experiment (Appendix G), replacing the forecast with a source video's JEPA features, with the initial frame, prompt, and noise fixed, shifts the generated motion toward the source's, with less post-contact bouncing and slower pre-contact descent.

Table 4: **Registry ablations on Physics-IQ Verified.** (a) Forecast-content controls on Cosmos 3 Nano 16B and WAN 2.2 I2V-A14B. (b) Comparison of learnable registers, joint generation, forecast-initialized registry states, and full Physical Registry on Cosmos v2.5 2B.

Registry content	I2V	V2V
Physical Registry (Cosmos 3 Nano 16B)	**36.01** ± 0.95	**45.12** ± 0.94
Zero mask	16.78 ± 0.01	24.15 ± 0.32
Perturbed mask (Noise $\sigma = 0.5$)	35.84 ± 0.11	44.35 ± 0.75
Physical Registry (WAN 2.2 I2V-A14B)	**38.44** ± 1.06	
Zero mask	35.88 ± 0.22	
Perturbed mask (Noise $\sigma = 0.5$)	36.40 ± 0.39	

Variant	I2V	Δ	V2V	Δ
Vanilla fine-tuning	41.3 ± 2.1	–	46.4 ± 1.0	–
ViT Needs Registers (Darcet et al., 2024)	39.0 ± 1.1	−2.3	44.6 ± 0.9	−1.8
Joint generation	45.1 ± 1.8	+3.8	48.0 ± 0.4	+1.6
Registry only	45.0 ± 1.3	+3.7	50.3 ± 0.9	+3.9
Physical Registry	**51.2** ± 1.5	**+9.9**	**55.1** ± 1.1	**+8.7**

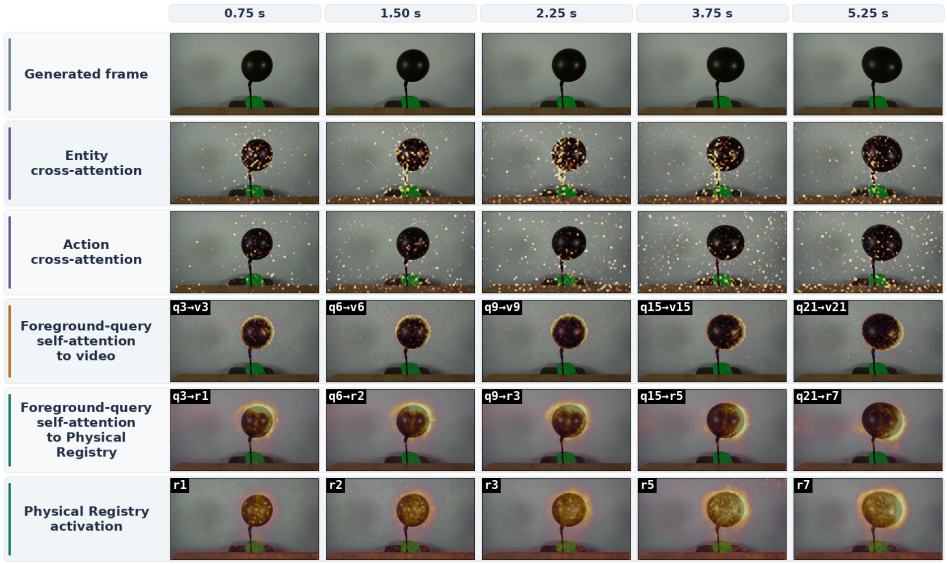


Figure 5: **Registry attention and activation during balloon inflation.** Foreground queries are video tokens inside the Physics-IQ motion mask of the generated video. Labels give query→key pairs (e.g., q3→r1: queries in latent frame 3 attending to its time-aligned registry slot r1); r1 alone marks that slot's activation. Protocol: Appendix A; more examples: Appendices H.1 and I.3.

Learnable registers alone are not sufficient. In Table 4(b), the ViT Needs Registers baseline, which prepends $K{=}4$ input-independent trainable tokens learned only through the generation objective, underperforms vanilla fine-tuning, whereas forecast initialization improves over it and predictive supervision improves further. The gains therefore do not come from additional register capacity alone: both forecast content and predictive supervision contribute.

4.6 Registry Attention and Activation

Figure 5 visualizes where registry states respond during balloon inflation, whose largest visible change occurs along the right boundary. Text cross-attention localizes the balloon or covers the interaction region, and foreground-query attention to video tokens follows the balloon surface. In contrast, attention to the aligned registry slots and registry activation concentrate around the expanding boundary, indicating that registry responses are spatially aligned with the region undergoing physical change.

5 Discussion and Conclusion

Physical Registry gives predictive representations a temporally grounded state within the video backbone. Controlled interface comparisons show that this dedicated predictive state provides larger gains than conditioning or video-stream alignment, while readout-based selection demonstrates its utility during generation. Physical Registry improves Physics-IQ Verified scores and WorldModelBench physical adherence and total score for all four backbones. Together, these results support explicit predictive state as a useful **interface for shaping and reading physically plausible futures**.

AI USE STATEMENT

Writing and language polishing. Generative AI tools were used to refine wording, shorten figure captions, and improve consistency in terminology and presentation.

Retrieval and discovery. Generative AI tools were used to assist in identifying relevant prior methods and related literature. Papers considered or cited in this work were subsequently checked against their original sources by the authors.

Research execution. Generative AI tools were used for routine operational assistance during experimental execution, including preparing and submitting experiment jobs to cloud computing environments, formatting execution commands and configuration files, and assisting with minor scripting or runtime issues.

Manuscript drafting. Generative AI tools were used to help convert author-provided technical notes and descriptions into preliminary prose, reorganize existing text for clarity, and suggest alternative wording. All AI-assisted material was reviewed by the authors.

REFERENCES

Anates Labs. Physics-IQ verified leaderboard, 2026. URL https://physics-iq-verified.anates.ai/. Accessed September 23, 2026.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15619–15629, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Assran_Self-Supervised_Learning_From_Images_With_a_Joint-Embedding_Predictive_Architecture_CVPR_2023_paper.html.

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

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mido Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=QaCCuDfBk2.

Aritra Bhowmik, Denis Korzhenkov, Cees G. M. Snoek, Amirhossein Habibian, and Mohsen Ghafoorian. MoAlign: Motion-centric representation alignment for video diffusion models. In *International Conference on Learning Representations*, 2026. URL https://iclr.cc/virtual/2026/poster/10009784.

Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Maria Elisabeth Bechtle, Feryal Behbahani, Stephanie C.Y. Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando De Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative interactive environments. In *International Conference on Machine Learning*, volume 235, pp. 4603–4623, 2024. URL https://arxiv.org/abs/2402.15391.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International conference on learning representations*, volume 2024, pp. 2632–2652, 2024.

Quentin Garrido, Nicolas Ballas, Mahmoud Assran, Adrien Bardes, Laurent Najman, Michael Rabbat, Emmanuel Dupoux, and Yann LeCun. Intuitive physics understanding emerges from self-supervised pretraining on natural videos, 2025. URL https://arxiv.org/abs/2502.11831.

Genmo Team. Mochi 1. GitHub repository, 2024. URL https://github.com/genmoai/mochi.

Nate Gillman, Charles Herrmann, Michael Freeman, Daksh Aggarwal, Evan Luo, Deqing Sun, and Chen Sun. Force prompting: Video generation models can learn and generalize physics-based control signals. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2505.19386.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://worldmodels.github.io/.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, Poriya Panet, Sapir Weissbuch, Victor Kulikov, Yaki Bitterman, Zeev Melumian, and Ofir Bibi. LTX-Video: Realtime video latent diffusion, 2024. URL https://arxiv.org/abs/2501.00103.

Yu He, Lichen Ma, Zipeng Guo, Xinyuan Shan, Jingling Fu, Dong Chen, Junshi Huang, and Yan Li. HyperDiT: Hyper-connected transformers for high-fidelity pixel-space diffusion, 2026. URL https://arxiv.org/abs/2605.15741.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In *Advances in Neural Information Processing Systems*, pp. 8633–8646, 2022. URL https://arxiv.org/abs/2204.03458.

Yue Hu, Siyuan Huang, Yue Liao, Shengcong Chen, Pengfei Zhou, Liliang Chen, Guanghui Ren, and Maoqing Yao. EWMBench: Evaluating scene, motion, and semantic quality in embodied world models. In *British Machine Vision Conference (BMVC)*, 2025. URL https://arxiv.org/abs/2505.09694.

Siyuan Huang, Liliang Chen, Pengfei Zhou, Shengcong Chen, Yue Liao, Zhengkai Jiang, Yue Hu, Peng Gao, Hongsheng Li, Maoqing Yao, and Guanghui Ren. EnerVerse: Envisioning embodied future space for robotics manipulation. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2501.01895.

Sungwon Hwang, Hyojin Jang, Kinam Kim, Minho Park, and Jaegul Choo. Cross-frame representation alignment for fine-tuning video diffusion models, 2025. URL https://arxiv.org/abs/2506.09229.

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.

Sonia Joseph, Quentin Garrido, Randall Balestriero, Matthew Kowal, Thomas Fel, Shahab Bakhtiari, Blake Richards, and Mike Rabbat. Interpreting physics in video world models. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.07050.

Theodoros Kouzelis, Efstathios Karypidis, Ioannis Kakogeorgiou, Spyridon Gidaris, and Nikos Komodakis. Boosting generative image modeling via joint image-feature synthesis. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2504.16064.

Kuaishou Technology. KLING: Video generation model, 2024. URL https://kling.kuaishou.com/.

Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph E. Gonzalez, Ion Stoica, Song Han, and Yao Lu. WorldModelBench: Judging video generation models as world models. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2025a. URL https://arxiv.org/abs/2502.20694.

Jiachen Li, Weixi Feng, Tsu-Jui Fu, Xinyi Wang, Sugato Basu, Wenhu Chen, and William Yang Wang. T2V-Turbo: Breaking the quality bottleneck of video consistency model with mixed reward feedback. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2405.18750.

Quanhao Li, Zhen Xing, Rui Wang, Hui Zhang, Qi Dai, and Zuxuan Wu. MagicMotion: Controllable video generation with dense-to-sparse trajectory guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12112–12123, 2025b. URL https://arxiv.org/abs/2503.16421.

Jiesong Lian, Zixiang Zhou, Ruizhe Zhong, Yuan Zhou, Qinglin Lu, Rui Wang, Long Hu, Yixue Hao, and Baoru Huang. Sara: Semantically adaptive relational alignment for video diffusion models. *arXiv preprint arXiv:2605.07800*, 2026.

Bin Lin, Yunyang Ge, Xinhua Cheng, Zongjian Li, Bin Zhu, Shaodong Wang, Xianyi He, Yang Ye, Shenghai Yuan, Liuhan Chen, Tanghui Jia, Junwu Zhang, Zhenyu Tang, Yatian Pang, Bin She, Cen Yan, Zhiheng Hu, Xiaoyi Dong, Lin Chen, Zhang Pan, Xing Zhou, Shaoling Dong, Yonghong Tian, and Li Yuan. Open-Sora Plan: Open-source large video generation model, 2024. URL https://arxiv.org/abs/2412.00131.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2209.03003.

Luma AI. Dream Machine, 2024. URL https://lumalabs.ai/dream-machine.

MiniMax. Hailuo AI video generator, 2024. URL https://hailuoai.video/.

MiniMax. MiniMax H3, 2026. URL https://huggingface.co/MiniMaxAI/MiniMax-H3.

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles? In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 948–958, 2026. URL https://arxiv.org/abs/2501.09038.

Lorenzo Mur-Labadia, Matthew Muckley, Amir Bar, Mido Assran, Koustuv Sinha, Mike Rabbat, Yann LeCun, Nicolas Ballas, and Adrien Bardes. V-JEPA 2.1: Unlocking dense features in video self-supervised learning, 2026. URL https://arxiv.org/abs/2603.14482.

NVIDIA. PhysicalAI-WorldModel-Synthetic-Physical-Interaction-Scenes. Hugging Face, 2026a. URL https://huggingface.co/datasets/nvidia/PhysicalAI-WorldModel-Synthetic-Physical-Interaction-Scenes.

NVIDIA. PhysicalAI WorldModel Synthetic Embodied Robot Scenes. Hugging Face, 2026b. URL https://huggingface.co/datasets/nvidia/PhysicalAI-WorldModel-Synthetic-Embodied-Robot-Scenes.

NVIDIA et al. Cosmos world foundation model platform for physical AI, 2025a. URL https://arxiv.org/abs/2501.03575.

NVIDIA et al. World simulation with video foundation models for physical AI, 2025b. URL https://arxiv.org/abs/2511.00062.

NVIDIA et al. Cosmos 3: Omnimodal world models for physical AI, 2026. URL https://arxiv.org/abs/2606.02800.

Saurabh Pathak, Elahe Arani, Mykola Pechenizkiy, and Bahram Zonooz. PhysVid: Physics aware local conditioning for generative video models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 41847–41858, 2026. URL https://arxiv.org/abs/2603.26285.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023. URL https://arxiv.org/abs/2212.09748.

Physics-IQ Benchmark. Physics-IQ verified leaderboard. GitHub repository README, 2026. URL https://github.com/google-deepmind/physics-IQ-benchmark.

Tim Rädsch, Yuki M. Asano, Hilde Kuehne, Stefan Bauer, Priyank Jaini, Robert Geirhos, and Carsten T. Lüth. Physics-IQ Verified. *arXiv preprint arXiv:2606.18943*, 2026.

Runway. Introducing Gen-3 Alpha: A new frontier for video generation, 2024. URL https://runwayml.com/research/introducing-gen-3-alpha.

Ying Shen, Jerry Xiong, Tianjiao Yu, and Ismini Lourentzou. PHANTOM: Physics-infused video generation via joint modeling of visual and latent physical dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11185–11194, 2026. URL https://arxiv.org/abs/2604.08503.

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3D convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4489–4497, 2015. URL https://arxiv.org/abs/1412.0767.

Wan Team. Wan2.2, 2025. URL https://github.com/Wan-Video/Wan2.2.

Chen Wang, Chuhao Chen, Yiming Huang, Zhiyang Dou, Yuan Liu, Jiatao Gu, and Lingjie Liu. PhysCtrl: Generative physics for controllable and physics-grounded video generation. In *Advances in Neural Information Processing Systems*, 2025a. URL https://arxiv.org/abs/2509.20358.

Jing Wang, Ao Ma, Ke Cao, Jun Zheng, Jiasong Feng, Zhanjie Zhang, Wanyuan Pang, and Xiaodan Liang. WISA: World simulator assistant for physics-aware text-to-video generation. In *Advances in Neural Information Processing Systems*, 2025b. URL https://arxiv.org/abs/2503.08153.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, Conghui He, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Yu Qiao. InternVid: A large-scale video-text dataset for multimodal understanding and generation, 2023. URL https://arxiv.org/abs/2307.06942.

Jiannan Xiang, Guangyi Liu, Yi Gu, Qiyue Gao, Yuting Ning, Yuheng Zha, Zeyu Feng, Tianhua Tao, Shibo Hao, Yemin Shi, Zhengzhong Liu, Eric P. Xing, and Zhiting Hu. Pandora: Towards general world model with natural language actions and video states, 2024. URL https://arxiv.org/abs/2406.09455.

Xindi Yang, Baolu Li, Yiming Zhang, Zhenfei Yin, Lei Bai, Liqian Ma, Zhiyong Wang, Jianfei Cai, Tien-Tsin Wong, Huchuan Lu, and Xu Jia. 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, 2025a. URL https://openaccess.thecvf.com/content/ICCV2025/html/Yang_VLIPP_Towards_Physically_Plausible_Video_Generation_with_Vision_and_Language_ICCV_2025_paper.html.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. CogVideoX: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, 2025b. URL https://arxiv.org/abs/2408.06072.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.06940.

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

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16118–16129, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Yuan_Inference-time_Physics_Alignment_of_Video_Generative_Models_with_Latent_World_CVPR_2026_paper.html.

Xiangdong Zhang, Jiaqi Liao, Shaofeng Zhang, Fanqing Meng, Xiangpeng Wan, Junchi Yan, and Yu Cheng. VideoREPA: Learning physics for video generation through relational alignment with foundation models. In *Advances in Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=oHjLfABsK4.

Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-Sora: Democratizing efficient video production for all, 2024. URL https://arxiv.org/abs/2412.20404.

Zangwei Zheng, Xiangyu Peng, Yuxuan Lou, Chenhui Shen, Tom Young, Xinying Guo, Binluo Wang, Hang Xu, Hongxin Liu, Mingyan Jiang, Wenjun Li, Yuhui Wang, Anbang Ye, Gang Ren, Qianran Ma, Wanying Liang, Xiang Lian, Xiwen Wu, Yuting Zhong, Zhuangyan Li, Chaoyu Gong, Guojun Lei, Leijun Cheng, Limin Zhang, Minghao Li, Ruijie Zhang, Silan Hu, Shijie Huang, Xiaokang Wang, Yuanheng Zhao, Yuqi Wang, Ziang Wei, and Yang You. Open-sora 2.0: Training a commercial-level video generation model in $200k, 2025. URL https://arxiv.org/abs/2503.09642.

APPENDIX

Appendix A gives experimental details. Appendices B to E report additional benchmark results: separating fine-tuning on the Enlarge dataset from the registry interface, Physics-IQ components, WorldModelBench references, and EWMBench. Appendix F details the selector, Appendix G the source-transfer probe, and Appendix H the mechanistic diagnostics. Appendix I collects qualitative comparisons.

A IMPLEMENTATION AND EXPERIMENTAL DETAILS

Backbone adaptation. The frozen JEPA encoder is V-JEPA 2.1 (Mur-Labadia et al., 2026) ViT-g at 384 px input, loaded from the pretrained checkpoint's target-encoder branch. The same encoder also extracts the full-video features for the alignment target. The registry input has eight slots (one context feature and seven predictor rollout steps), each with 196 spatial tokens of 5,632 channels, and the alignment target is the mean over the 196 spatial tokens. Video and registry tokens share dense self-attention, while text remains a separate cross-attention condition.

Alignment weight. For the full Physical Registry objective in Equation (11), we use $\lambda_{\text{align}} = 0.5$, following the default projection-loss coefficient in the official REPA implementation (Yu et al., 2025). Registry only retains the forecast-initialized registry tokens and removes the alignment term.

Alignment locations. As in prior alignment work, which aligns depth 8 of the 24-block SiT-L/2 (depth 6 or 8 performing best in its ablation) (Yu et al., 2025) and depth 18 of CogVideoX (Zhang et al., 2025), we supervise an early-to-middle generation block. The table lists block indices, not denoising steps.

Mean-pooled implementation	Generation blocks	Alignment block index
Cosmos 3 Nano 16B	36	10
Cosmos v2.5 2B	28	14
WAN 2.2 I2V-A14B	40	11
WAN 2.2 TI2V-5B	30	15

For WAN I2V-A14B, the block count and alignment index apply to each noise expert. The Cosmos v2.5 architecture figure and spatial diagnostics use block 14, whereas the depth sweep examines several hidden layers of the same trained model.

The Enlarge dataset. The Enlarge dataset combines physics-focused data with broad visual coverage. It includes physics-oriented video, rigid-body simulation, general web video, and embodied robot simulation (Wang et al., 2025b; 2023; NVIDIA, 2026a;b). We fine-tune Physical Registry using only its RGB videos and corresponding captions, without additional physical annotations or simulator states. The Enlarge dataset excludes Physics-IQ videos. Table 5 and Figure 6 report the exact number of clips and video hours used from each source.

Dataset	Clips	Mean (s)	Hours
WISA	74,370	7.1	146.7
InternVid	8,115	7.5	16.9
SDG-PhysXSim	101,422	5.0	140.9
SDG-RobotSim	28,193	9.0	70.6
Total	**212,100**	**6.4**	**375.2**

Table 5: **The Enlarge dataset.** Exact clips used by the evaluated Physical Registry models.

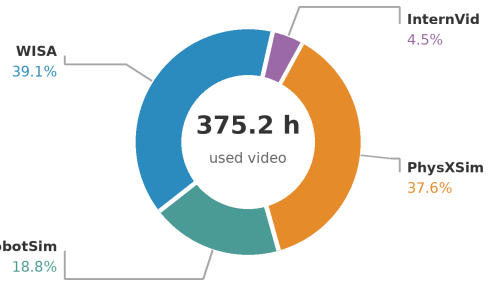


Figure 6: **Used-hour composition.** Projected hours for the exact used clips.

Fine-tuning settings. Table 6 summarizes the pretrained and additional model sizes, effective batch, and GPU resources for Physical Registry models fine-tuned on the Enlarge dataset.

Table 6: **Model sizes, fine-tuning resources, and optimization settings.** Physical Registry models are fine-tuned on the Enlarge dataset with AdamW and an effective global batch size of 16. Registry modules denote the projection and readout parameters added to the backbone, which is itself fine-tuned.

Model	Pretrained	Registry modules	Batch/update	GPUs
Cosmos 3 Nano 16B	16B	72.63M	16	8× H100
Cosmos v2.5 2B	2B	35.93M	16	8× A100 40 GB
WAN 2.2 I2V-A14B	2×14B	129.39M	16	8× H100
WAN 2.2 TI2V-5B	5B	105.02M	16	8× A100 40 GB

Model	Learning rate	Schedule	Weight decay	AdamW (β_1, β_2)
Cosmos 3 Nano 16B	1×10^{-5}	constant	0	$(0.9, 0.95)$
Cosmos v2.5 2B	3×10^{-5}	100-step warm-up, then decay	1×10^{-3}	$(0.9, 0.999)$
WAN 2.2 I2V-A14B	1×10^{-5}	constant	1×10^{-2}	$(0.9, 0.95)$
WAN 2.2 TI2V-5B	1×10^{-5}	constant	1×10^{-2}	$(0.9, 0.999)$

The separate matched-budget WAN control in Appendix B uses 8,000 iterations and effective batch 16.

Sampling settings. Table 7 lists the sampling settings of the evaluated models. Apart from the expanded-prompt setting (†, Table 2), guidance, shift, and sampler follow the official inference defaults of each release, as do the step counts. Each model generates at least five seconds after its conditioning frames, and, following the official Physics-IQ protocol, the first five seconds are evaluated. Each backbone and its Physical Registry counterpart share the same settings.

Model	Steps	Guidance	Shift	Sampler	Resolution	Gen. frames @ fps
Cosmos 3 Nano 16B	35	6.0	10.0	UniPC	1280×720	120 @ 24
Cosmos 3 Nano 16B†	50	3.0	4.75	UniPC	1280×720	120 @ 24
Cosmos v2.5 2B	35	7.0	5.0	UniPC	1280×704	92 / 80 @ 16
WAN 2.2 I2V-A14B	40	3.5 / 3.5	5.0	UniPC	1280×720	80 @ 16
WAN 2.2 TI2V-5B	50	5.0	5.0	UniPC	1280×704	120 @ 24

Table 7: **Sampling settings.** Steps are denoising steps and guidance is the classifier-free guidance scale (WAN I2V-A14B: high-/low-noise experts). Generated frames exclude conditioning frames (Cosmos v2.5: I2V / V2V). † Expanded-prompt setting of Table 2.

Sampling cost. The registry lengthens the latent sequence, so a Physical Registry sampler step costs more than a backbone step (Table 8). Both arms of a backbone are measured in one session on a single H100 80 GB over three Physics-IQ scenarios with three repeats each, reporting the median time per sampler step after discarding the first five forwards of every run. Because Nano's sampler reuses cached activations on part of its schedule, its per-step time is the total sampling time divided by the number of sampler steps. On Cosmos v2.5 2B the registry adds $28{,}160$ tokens, a third of the sequence, and a step costs $1.67\times$. On Cosmos 3 Nano 16B, the registry adds $7{,}360$ tokens to each guidance branch and a step costs $1.55\times$. Times are compared only within each backbone.

Model	Arm	Sequence (tokens)	s / sampler step	Peak (GiB)
Cosmos v2.5 2B	Backbone	84,480	6.514	6.20
	+Physical Registry	112,640	10.909	7.00
Cosmos 3 Nano 16B	Backbone	28,607 / 31,364	1.384	11.44
	+Physical Registry	35,967 / 38,724	2.148	11.94

Table 8: **Sampling cost of the registry.** Both arms of a backbone are measured in one session on a single H100 80 GB. Times are per sampler step; peak memory is the generation peak above the weights resident before sampling. Two sequence lengths are given for Cosmos 3 Nano 16B because its two classifier-free guidance branches carry different text lengths.

Benchmark metrics. Physics-IQ Verified measures physical plausibility against a recorded future, with higher scores indicating better results (Rädsch et al., 2026). It normalizes four subscores per

sample and averages equally across subscores and inputs, replacing dataset-level normalization. The input-level scores support within-input candidate comparisons. WorldModelBench (Li et al., 2025a) supplies a first-frame image and text instruction for 350 prompts spanning 7 application domains and 56 subdomains. Its human-aligned multimodal judge scores instruction following, physical adherence over five physical laws, and common sense over frame-wise and temporal quality. The total is the sum of these scores. EWMBench (Hu et al., 2025) evaluates embodied world models on real robot manipulation episodes, scoring visual scene consistency, motion correctness against the recorded trajectory, and semantic alignment with the task instruction (Appendix E). Each backbone and Physical Registry pair uses the same conditioning inputs, sampling protocol, and judge on the full set.

Tasks and candidate pools. The evaluation covers image-to-video and video-to-video generation on Physics-IQ (Motamed et al., 2026). Candidates are compared within the same scene, view, and conditioning input. Within matched pools of ten candidates per input, Plain scores average the pool, while WMReward and PRS select one candidate from it. The content ablations on Cosmos 3 Nano 16B and WAN 2.2 I2V-A14B replace the projected registry features with zeros (zero mask) or apply noise with $\sigma = 0.5$ (perturbed mask; for Nano, Gaussian noise in the configuration without prompt expansion). Their full-model references repeat the Plain values in Table 2, and the results are in Table 4(a).

Center-trained conditioning ablation. All conditioning interfaces use the same pretrained video backbone, center-view adaptation data, training budget, and held-view evaluation, and the JEPA-based variants use a fixed JEPA source. The REPA arm retains the standard video-latent sequence and uses JEPA as a teacher for ordinary denoising hidden states. Joint generation (Kouzelis et al., 2025) adapts ReDi's joint image–feature synthesis to video and JEPA targets. It concatenates noised video and future-feature tokens along the sequence dimension for joint denoising in the same video DiT, retaining observed JEPA context as clean conditioning and applying decoded-feature alignment to the future outputs. Registry only keeps the forecast-initialized registry tokens but removes the registry-targeted alignment that full Physical Registry applies.

Mechanistic diagnostics. Terminal registry readouts are compared with separately computed completed-video encodings. The balloon example, layer visualization, and additional scenarios use different measurement windows. The visualization protocol below and Appendix H.1 specify their reductions.

Cross-view JEPA representation check. This check tests whether JEPA representations from different views of the same scenario remain closer than representations from different scenarios. Figure 7 compares all eight temporal tokens. Same-scenario cross-view pairs have lower mean L2 distance and higher mean cosine similarity than different-scenario pairs, indicating that JEPA features preserve scenario correspondence across viewpoints.

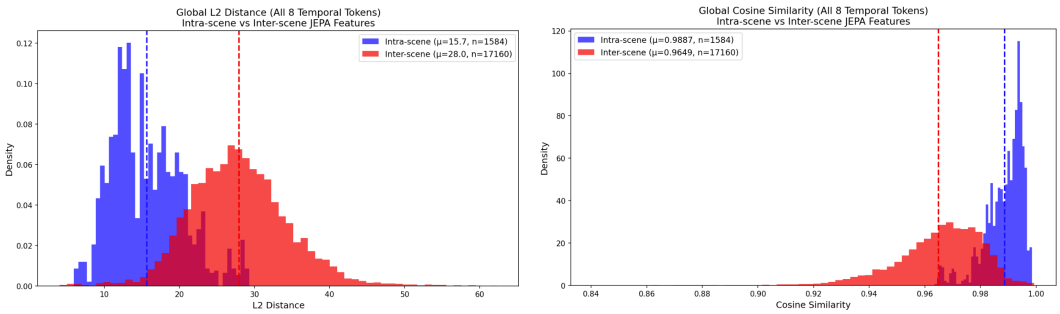


Figure 7: **Cross-view JEPA feature correspondence.** Left: L2 distance; right: cosine similarity. Blue compares views of the same scene; red compares different scenes. Dashed lines mark the group means: 15.7 versus 28.0 for L2 and 0.9887 versus 0.9649 for cosine. The all-eight-token diagnostic reports 1,584 intra-scene and 17,160 inter-scene entries; n counts plotted entries.

Registry projection architecture. The adapter G_ϕ of Equation (4) combines established video-modeling components: local spatio-temporal convolutions (Tran et al., 2015) and per-position temporal attention, a factorization used in video diffusion architectures (Ho et al., 2022). The

convolutions mix neighboring locations and adjacent forecast features. Temporal attention then integrates the slot sequence at each spatial position. Spatial resizing and channel projection match the backbone's latent geometry while retaining the forecast's temporal indexing. Figure 8 shows this adapter.

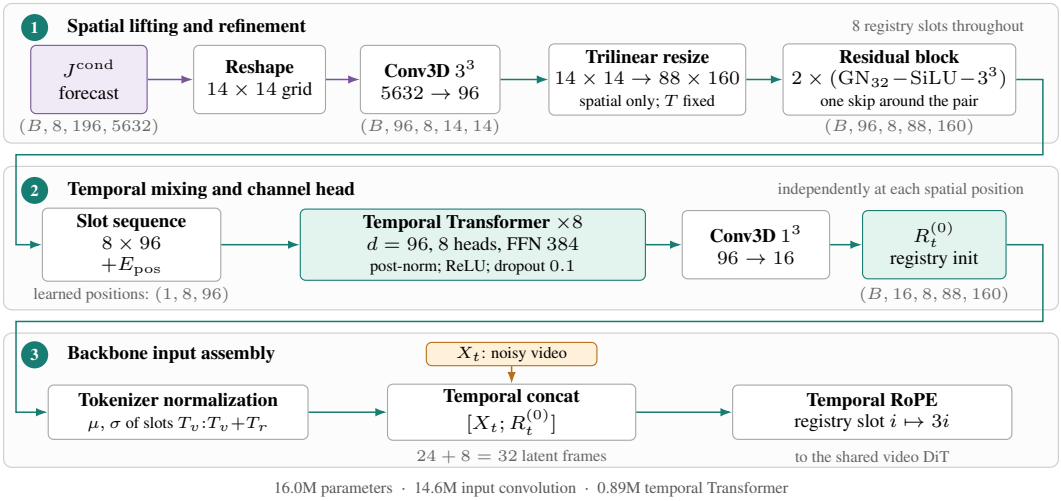


Figure 8: **The registry projection layer G_ϕ.** The three bands show spatial lifting with 3D convolutions (Tran et al., 2015), per-position temporal attention (Ho et al., 2022) and channel projection, and backbone input assembly. Shapes are for the Cosmos v2.5 2B configuration. The "Projection stages" paragraph below specifies the operations.

Projection input and output. For the Cosmos v2.5 2B configuration, the input is the forecast J^{cond}, a sequence of $T_r{=}8$ slots, each a 14×14 grid of 5632-dimensional features (the encoder embedding width of 1408 concatenated over four blocks). The output is $R_t^{(0)}$ in the video-latent format of the backbone: 16 channels over 8 registry slots at the 88×160 latent resolution.

Projection stages. Figure 8 gives the operations and shapes of each stage. Upsampling acts on the two spatial axes only, and the temporal Transformer is applied independently at each spatial position with the eight registry slots as its sequence, so spatial mixing comes from the convolutions and, after concatenation, from the backbone's self-attention. The adapter is initialized randomly and trained jointly with the backbone.

Projection placement. The module emits the backbone's native latent channel count. Its output is normalized with the tokenizer's own per-frame statistics for the registry slots, which places the registry on the scale of the video latent. The normalized output is then concatenated along the temporal axis to give $T_v{+}T_r = 24 + 8 = 32$ latent frames. Registry slot i receives the video-latent RoPE coordinate $3i$, the temporal alignment of Equation (5). With 2×2 spatial patchification, each registry slot contains 44×80 tokens, giving $8 \times 44 \times 80 = 28{,}160$ registry tokens in this configuration.

Projection size. The module has 16.0M parameters: 14.6M in the first convolution, 0.89M in the temporal Transformer, and 0.5M in the remaining stages. Relative to the 2B-parameter backbone, the adapter is under one percent of the trained parameters.

Registry readout modules. The readout operator q_ψ in Equations (8) and (12) includes spatial mean-pooling within each registry slot, before its projection. Registry slots remain separate until loss or score aggregation. For training, q_ψ uses a three-layer MLP with output widths 2048, 2048, and 5632, and SiLU activations (19.9M parameters). It maps the spatially averaged registry hidden state to the encoder feature dimension.

The slot-level target $\overline{Z}_s^{\text{tgt}}$ and reference $\overline{J}_s^{\text{cond}}$ are spatial means of the frozen encoder's target grid and the unprojected shared-forecast grid, respectively. These reductions preserve the temporal indexing

of their source features. The alignment loss trains a single readout, which the selection policy reuses frozen. No additional module is trained for selection. The conditioning adapter G_ϕ operates in the reverse direction, mapping the forecast into the video-latent geometry.

Visualization protocol. Each heat row uses a fixed row-specific scale across the five timestamps. Registry activation is the hidden-state variance across early denoising steps 0–10 at each aligned registry slot and spatial position (averaged over feature channels), shown with log scaling ($\gamma = 1$). Entity and action maps use a linear 60th–98th percentile scale, $\gamma = 1$, and screen blending with maximum opacity 0.75. Native maps are bicubically interpolated before display normalization, and values below the lower limit are transparent. For Figure 5, foreground queries use the Physics-IQ generated-foreground mask (the motion mask of pixels whose intensity changes against an adaptive background (Motamed et al., 2026), computed on the generated video) downsampled by area averaging to the 44×80 video-latent grid. The attention rows average exact native QK-softmax responses over denoising steps $\{0, 5, 11, 16, 22, 28, 34\}$. For the examples in Appendix I.3, text-attention maps sum the lexical token groups, including descriptive state verbs, and divide by valid-text attention at each spatial location. Text and video self-attention use steps 0–10, registry self-attention uses steps 11–22, and in-panel labels identify the video-query and key-slot pairs used for each map.

Prompt expansion. For the expanded-prompt Nano setting in Tables 2 and 12, Claude Opus 5 (model id `claude-opus-5`) expands the Physics-IQ original description (OP) together with the conditioning image into NVIDIA's Cosmos I2V structured prompt format. The image specifies appearance and layout, whereas OP specifies temporal dynamics. The expansions are generated once and reused in every run. The balloon-inflation example contrasts OP with the benchmark's best-practice prompt (BPP), and shows the full expanded prompt produced from OP and the image.

Original description (OP)

```
A black balloon is attached to a fixed electric air pump hose on a table
with a plain wall in the background.  Air is being pumped in the balloon.
Static shot with no camera movement.
```

Benchmark best-practice prompt (BPP)

```
Static locked-off single-shot with fixed frame throughout filmed with
constant framerate in real-time.  The scene shows a realistic scientific
demonstration.  The scene only contains the described setup and actions.
The balloon is attached to a black hose that points upwards from behind
a wooden table.  Air is being pumped into the black balloon.
```

Expanded prompt (OP + image)

Temporal dynamics. `Air is pumped continuously into the black balloon so that it swells steadily larger over the five seconds, its rubber stretching thinner and its highlights spreading while the neck stiffens and lifts the balloon more upright on the hose.`

Subject. `A matte black rubber balloon inflated to a firm round sphere, its neck pulled down over the black nozzle of a hose and clamped there with a band of bright blue tape.`

Setting. `A bare indoor corner used as a test bay: a plain pale grey-green painted wall fills the frame, and a honey-toned oak tabletop runs straight across the bottom as a level horizontal band. A broad soft shadow of the rig falls on the wall to the left. No other furniture, decoration or clutter is present.`

Camera. `Static`

0:00-0:01. `The pump runs on and the black balloon, already a firm centred sphere, grows visibly wider; its paired highlights drift apart as the surface curvature flattens.`

0:01-0:02. `The balloon continues to swell, its lower bulge rounding out and the pear-shaped taper above the blue tape shortening as rubber is drawn up into the body.`

0:02-0:03. `Pressure rises enough to stiffen the braided hose; the balloon climbs a little higher on the vertical midline and its slight lean straightens out entirely.`

0:03-0:04. `Expansion carries on at the same steady rate, the rubber now thinner and its blacks a shade lighter where it is most stretched, with a small wobble passing through the balloon on each pump surge.`

0:04-0:05. `The balloon reaches its largest size in the clip, taut and nearly upright on the stiff hose, still growing and still intact as the pump keeps running.`

B Separating Fine-Tuning on the Enlarge Dataset from the Registry Interface

The controlled comparisons in Table 1 and Table 4(b) use Cosmos v2.5 2B. To separate fine-tuning on the Enlarge dataset from the Physical Registry in a second backbone family, we track WAN TI2V-5B for 8,000 iterations with batch size 16, evaluating Physics-IQ Verified every 1,000 iterations. Vanilla fine-tuning on the Enlarge dataset uses the same data, schedule, and optimizer as Physical Registry, differing only in the absence of registry tokens and predictive alignment.

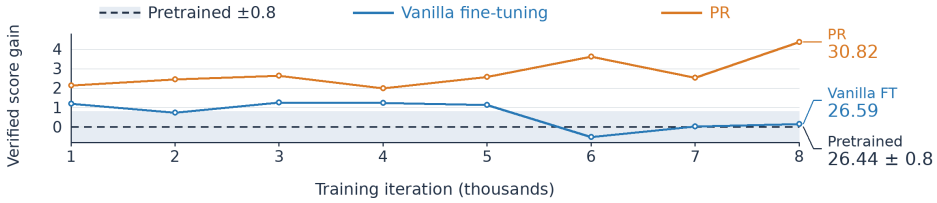


Figure 9: Physics-IQ Verified gains over pretrained WAN TI2V-5B. Shading marks one pretrained standard deviation (± 0.8) around zero.

The vanilla fine-tuning stays within 1.3 points of the pretrained score throughout and ends at $+0.15$ after 8,000 iterations. Physical Registry exceeds vanilla at all eight evaluation points. Even our smallest gain ($+1.98$) exceeds vanilla's largest ($+1.25$).Under the same data and update budget, vanilla fine-tuning produces little change from pretrained performance, whereas Physical Registry improves consistently across checkpoints. This control supports the interpretation that the WAN TI2V-5B gain in Table 2 arises from the registry interface rather than from additional training data alone.

C PHYSICS-IQ COMPONENTS AND ORIGINAL AGGREGATION

Table 9 reports the four Physics-IQ Verified components of the Physical Registry models fine-tuned on the Enlarge dataset in Tables 2 and 12, retaining both Nano prompt settings. The Original column recomputes the original Physics-IQ aggregation rule from the same evaluation metrics (Motamed et al., 2026; Rädsch et al., 2026). Across both tasks, the model ordering under the Original aggregation is unchanged from Verified.

I2V	Spatial	Spatiotemp.	Weighted	Inv. MSE	Verified	Original
Nano 16B + PR	<u>46.63</u>	31.01	<u>32.76</u>	33.63	36.01	35.00
Nano 16B† + PR	**60.70**	33.56	**42.46**	**37.53**	**43.56**	**42.95**
Cosmos v2.5 2B + PR	44.46	<u>34.44</u>	30.29	25.65	33.71	34.00
WAN I2V-A14B + PR	44.77	**42.06**	32.67	<u>34.19</u>	<u>38.44</u>	<u>38.90</u>
WAN TI2V-5B + PR	36.82	30.81	22.33	26.99	29.24	27.87
V2V	Spatial	Spatiotemp.	Weighted	Inv. MSE	Verified	Original
Nano 16B + PR	57.72	<u>38.31</u>	<u>43.43</u>	<u>41.02</u>	<u>45.12</u>	<u>45.60</u>
Nano 16B† + PR	**68.05**	**38.37**	**49.90**	**41.85**	**49.54**	**50.54**
Cosmos v2.5 2B + PR	<u>57.98</u>	37.43	43.37	30.56	42.33	44.39

Table 9: **Physics-IQ components and Original aggregation.** Component scores, the Verified aggregate, and the Original aggregation for our Physical Registry (PR) models, each normalized to 0–100. **Bold**/<u>underline</u> mark the highest/second-highest value per column within a task. † Expanded-prompt setting of Table 2.

D WorldModelBench Reference Leaderboard

Table 10 collects published WorldModelBench reference scores (Li et al., 2025a) together with the backbone comparisons in Table 3. The leaderboard entries are reported by the benchmark authors under their own generation settings. The Cosmos v2.5 backbone-versus-Physical Registry pair uses the matched protocol described in Appendix A.

Model	Instruction following (0–3)	Physical adherence (0–5)	Common sense (0–2)	Total (0–10)
Cosmos 3 Nano 16B (NVIDIA et al., 2026)	2.20 ± 0.02	4.50 ± 0.08	1.43 ± 0.06	8.13 ± 0.13
+ Physical Registry	2.20 ± 0.00 (+0.001)	**4.82** ± 0.03 (+0.32)	1.81 ± 0.02 (+0.38)	8.83 ± 0.04 (+0.69)
Cosmos v2.5 2B (NVIDIA et al., 2025b)	2.14 ± 0.03	4.43 ± 0.07	1.55 ± 0.05	8.12 ± 0.14
+ Physical Registry	2.07 ± 0.01 (-0.07)	4.54 ± 0.03 (+0.11)	1.58 ± 0.04 (+0.03)	8.19 ± 0.01 (+0.07)
WAN 2.2 I2V-A14B (Wan Team, 2025)	2.23 ± 0.03	4.45 ± 0.04	1.60 ± 0.03	8.28 ± 0.02
+ Physical Registry	2.30 ± 0.04 (+0.07)	4.53 ± 0.04 (+0.07)	1.60 ± 0.02 (+0.004)	8.43 ± 0.05 (+0.15)
WAN 2.2 TI2V-5B (Wan Team, 2025)	**2.33** ± 0.01	4.35 ± 0.10	1.57 ± 0.04	8.24 ± 0.14
+ Physical Registry	2.24 ± 0.05 (-0.09)	4.50 ± 0.05 (+0.14)	1.68 ± 0.06 (+0.11)	8.41 ± 0.11 (+0.17)
KLING (Kuaishou Technology, 2024)	<u>2.32</u>	<u>4.81</u>	**1.95**	**9.10**
Minimax (MiniMax, 2024)	2.28	4.72	<u>1.92</u>	<u>8.92</u>
Mochi 1-official (Genmo Team, 2024)	2.00	4.80	1.86	8.66
Runway (Runway, 2024)	2.17	4.60	1.86	8.64
Luma (Luma AI, 2024)	1.98	4.49	1.76	8.24
OpenSoraPlan-T2V (Lin et al., 2024)	1.72	4.64	1.68	8.04
Mochi 1 (Genmo Team, 2024)	2.06	4.39	1.45	7.91
CogVideoX-T2V (Yang et al., 2025b)	2.03	4.27	1.35	7.65
CogVideoX-I2V (Yang et al., 2025b)	1.78	4.17	1.12	7.08
Pandora (Xiang et al., 2024)	1.56	4.31	1.03	6.90
T2V-Turbo (Li et al., 2024)	1.37	4.10	1.08	6.56
OpenSora-T2V (Zheng et al., 2024)	1.61	3.88	0.69	6.17
OpenSora-I2V (Zheng et al., 2024)	1.42	3.85	0.55	5.82

Table 10: **WorldModelBench reference scores.** Backbone pairs report mean ± std, rounded to two decimals. **Bold**/<u>underline</u> mark the highest/second-highest mean in each displayed column. Green/red parentheses show positive/negative Physical Registry-minus-backbone changes before rounding. Rows from KLING onward are published references under their own generation settings (Li et al., 2025a). Main results: Table 3.

E EMBODIED WORLD MODEL BENCHMARK

EWMBench (Hu et al., 2025) evaluates embodied world models on real robot manipulation episodes along three axes: visual scene consistency, motion correctness against the recorded gripper trajectory, and semantic alignment with the task instruction. It complements Physics-IQ, which scores physical plausibility on passive scenes, by asking whether a generated rollout follows an embodiment through a task.

Protocol. We run the official evaluator on the full public ground-truth split, 21 episodes over seven tasks. Each episode is generated three times from its initial frame and instruction, giving 126 videos across the two arms. Our clips are 121 frames at 24 fps, whereas the recorded episodes run at 30 fps. Both arms share the backbone, sampling recipe, and conditioning inputs, and differ only in the registry interface.

	Scene	Motion			Semantics			Diversity
Model	Consist.	HSD	NDTW	DYN	BLEU	CLIP	Logic	Div.
Cosmos 3 Nano 16B (NVIDIA et al., 2026)	0.855	0.328	0.305	0.383	<u>0.242</u>	0.884	0.921	0.029
+Physical Registry	0.895 (+0.040)	0.351 (+0.023)	0.378 (+0.073)	0.459 (+0.076)	**0.244** (+0.002)	**0.897** (+0.013)	0.921	0.018 (-0.011)
EnerVerse-FT* (Huang et al., 2025)	<u>0.943</u>	**0.536**	**0.596**	<u>0.536</u>	0.180	0.864	**0.978**	<u>0.069</u>
LTX-FT* (HaCohen et al., 2024)	**0.944**	<u>0.476</u>	<u>0.521</u>	**0.620**	0.174	0.855	<u>0.967</u>	0.016
Kling* (Kuaishou Technology, 2024)	0.889	0.323	0.316	0.305	0.168	0.854	<u>0.967</u>	0.049
Hailuo* (MiniMax, 2024)	0.858	0.223	0.179	0.134	0.185	<u>0.886</u>	0.911	0.037
COSMOS* (NVIDIA et al., 2025a)	0.796	0.250	0.253	0.205	0.123	0.846	0.733	**0.080**
OpenSora* (Zheng et al., 2025)	0.921	0.155	0.142	0.047	0.160	0.851	0.822	0.042
LTX* (HaCohen et al., 2024)	0.916	0.158	0.143	0.100	0.069	0.832	0.733	0.017

* Published in the EWMBench evaluation table (Hu et al., 2025, Table 2), evaluated on ten tasks with ten ground-truth episodes each; the public ground-truth release used here covers 21 episodes across seven tasks. EnerVerse-FT and LTX-FT are adapted to the benchmark's robot domain; our models are not.

Table 11: **EWMBench (Hu et al., 2025).** All metrics are higher-is-better and shown on a 0–1 scale. Bold/underline mark the highest/second-highest value in each column across all rows, and green parentheses give the Physical Registry change over the backbone above it.

Results. Physical Registry improves scene consistency and all three motion metrics (Table 11). The largest gains are in trajectory agreement, where NDTW rises by 0.073 and DYN by 0.076. Among the semantic metrics, BLEU rises by 0.002, the logic pass rate is unchanged at the reported precision, and CLIPScore rises by 0.013. Diversity, the spread across the three seeds, decreases from 0.029 to 0.018 alongside the improvements in the trajectory metrics.

F THE PHYSICAL REGISTRY SELECTOR

F.1 AGGREGATION AND PROTOCOL

We apply the PRS rule in Equation (12) to candidates generated from the same conditioning input. Each cosine is evaluated before slot–step averaging. Prefix selection follows the PRS settings in Table 12. Cosmos v2.5 and Nano without † average registry-slot cosines 3–5 over evaluations $1, \ldots, K$. Nano† averages registry slots 4–7 over the latest $\min(K, 10)$ evaluations, reproducing its reported steps 41–50 window at $K = 50$. Prefix-selected quality is measured on the chosen candidate's completed video under its original 35- or 50-step schedule.

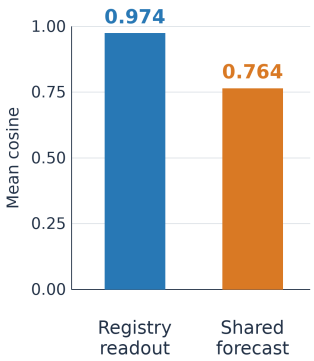
With N candidates and a full trajectory of D steps, continuing one candidate after K prefix steps gives a serial denoising budget of $NK + (D - K)$ candidate-steps, compared with ND for completing all candidates. All selection budgets are reported in these serial denoising candidate-steps. In the completed-video best-of-N (BoN) comparison, WMReward scores decoded candidates with its external world model (Yuan et al., 2026). Relative gains of PRS over Plain are computed as $100(\mathrm{PRS} - \mathrm{Plain})/\mathrm{Plain}$.

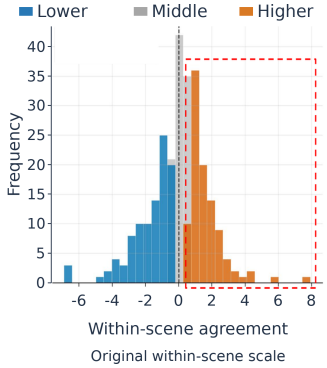
WAN selection. Both WAN backbones apply the cosine-only PRS rule in Equation (12): PRS minimizes forecast cosine, and Inverse maximizes the same cosine. All reported selection policies and

Oracles use ten candidates per input. Published Plain references remain external to these candidate pools.

I2V	Plain	WMReward	PRS	Oracle
Cosmos 3 Nano 16B	30.18 ± 0.84	**31.69**		41.48
+ Physical Registry	36.01 ± 0.95	38.85	**40.27**	51.88
Cosmos 3 Nano 16B†	37.3 ± 0.9*	**44.28**		50.59
+ Physical Registry†	43.56 ± 0.95	45.41	**45.81**	54.72
Cosmos v2.5 2B	**30.65** ± 0.37	30.61		36.07
+ Physical Registry	33.71 ± 1.11	37.74	**39.08**	49.37
WAN 2.2 I2V-A14B	32.2 ± 0.6*	**35.31**		43.71
+ Physical Registry	38.44 ± 1.06	**41.82**	40.58	51.87
WAN 2.2 TI2V-5B	27.7 ± 0.9*	**31.19**		37.76
+ Physical Registry	29.24 ± 1.30	**33.18**	32.03	39.88
V2V	Plain	WMReward	PRS	Oracle
Cosmos 3 Nano 16B	38.78 ± 0.78	**42.46**		49.09
+ Physical Registry	45.12 ± 0.94	**49.70**	46.78	59.43
Cosmos 3 Nano 16B†	46.45 ± 0.76	**50.41**		56.98
+ Physical Registry†	49.54 ± 0.63	51.69	**52.06**	59.69
Cosmos v2.5 2B	37.40 ± 0.53	**39.38**		42.60
+ Physical Registry	42.33 ± 1.90	46.59	**47.67**	55.84

Table 12: **Selection policies on Physics-IQ Verified.** WMReward and PRS each select one candidate from our ten-candidate pool; Oracle is the best in that pool, an upper bound rather than a policy. * marks published Plain values, external to the pool, and † the configuration in Table 2. Bold marks the strongest non-Oracle score in each row, gray unavailable PRS.





		Selection		
Model	Task	PRS	Inverse	Random
Cosmos v2.5	I2V	**39.08**	25.60	33.71
	V2V	**47.67**	33.76	42.33
Cosmos Nano	I2V	**40.27**	33.12	36.01
	V2V	**46.78**	42.62	45.12
WAN 14B	I2V	**40.58**	34.73	38.44
WAN 5B	I2V	**32.03**	26.91	29.24

Figure 10: **Candidate-conditioned registry readouts and selection direction.** (a) Mean cosine to the generated-video reference: the terminal readout tracks the candidate more closely than the shared forecast (0.974 versus 0.764). (b) Within-scene readout–generated-video agreement. (c) PRS against controls per backbone and task: Random is uniform selection in the same pool and Inverse the reversed ranking.

F.2 Direction and consistency

The registry readout tracks the realized candidate. For each completed candidate $v_{\text{gen}}^{(n)}$, we encode the generated video with the same frozen target encoder to obtain a diagnostic reference $Z_{\text{gen}}^{(n)}$. Figure 10(a) compares this reference with the terminal registry readout and with the shared observation-conditioned forecast. The terminal readout has a higher mean cosine similarity to the generated-video reference than the shared forecast (0.974 versus 0.764). This shows that, after interacting with the evolving video tokens, the registry more closely reflects the future realized by the candidate than the common forecast from which it was initialized.

This candidate dependence is visible within scenes. Figure 10(b) resolves the same comparison within individual scenes, showing variation in readout–generated-video agreement across candidates

that share the same observation-conditioned forecast. Thus, the terminal registry is not merely preserving the common initialization: its readout varies with the trajectory produced by each candidate.

The resulting departure has a consistent selection direction. Figure 10(c) compares PRS with the reversed ranking (Inverse) and uniform selection (Random) for each reported backbone and task. PRS outperforms both controls in every configuration, supporting the direction used in Equation (12): larger departure from the shared forecast is associated with better candidate selection.

We further test whether this direction remains stable as the candidate pool changes. For each pool size, we average exactly over all pools of that size drawn from the ten-candidate bank, so no particular pool is chosen. The full ten-candidate configuration appears in Figure 10(c). Random is the bank mean and does not depend on pool size. Every pool covers the same 198 inputs and 66 three-view physical scenarios, and each policy receives identical candidates within a pool.

Entries below are PRS-minus-control gains in Physics-IQ Verified points, with percentile 95% intervals from 10,000 paired scenario-bootstrap replicates. Each draw, shared by PRS and the control, resamples scenarios and keeps their three views and both tasks together.

Model	Pool	Task	Inverse	Random	WMReward
Cosmos v2.5 2B	6	I2V	$+12.58\,[9.19, 16.08]$	$+5.07\,[3.01, 7.23]$	$+1.33\,[-0.45, 3.33]$
	6	V2V	$+12.69\,[8.60, 17.04]$	$+4.90\,[2.85, 7.09]$	$+0.60\,[-0.64, 1.92]$
	5	I2V	$+11.46\,[8.39, 14.66]$	$+4.73\,[2.94, 6.59]$	$+1.24\,[-0.24, 2.98]$
	5	V2V	$+11.37\,[7.70, 15.34]$	$+4.50\,[2.65, 6.48]$	$+0.22\,[-0.84, 1.36]$
	4	I2V	$+10.05\,[7.34, 12.85]$	$+4.32\,[2.80, 5.87]$	$+1.05\,[-0.16, 2.49]$
	4	V2V	$+9.86\,[6.62, 13.30]$	$+4.08\,[2.48, 5.81]$	$-0.05\,[-0.91, 0.87]$
WAN 2.2 I2V-A14B	6	I2V	$+4.35\,[2.53, 6.28]$	$+1.82\,[0.92, 2.74]$	$-0.51\,[-1.31, 0.30]$
	5	I2V	$+3.99\,[2.32, 5.76]$	$+1.71\,[0.88, 2.55]$	$-0.47\,[-1.19, 0.25]$
	4	I2V	$+3.53\,[2.04, 5.10]$	$+1.56\,[0.83, 2.31]$	$-0.42\,[-1.05, 0.20]$

The direction effect holds at every pool size: maximizing departure has positive paired gains over inverse ranking and uniform selection, with intervals excluding zero in every row. Against completed-video WMReward, the differences are not statistically distinguishable from zero under these bootstrap intervals, while PRS decides from internal registry readouts before full candidate completion. Within-input rank agreement between departure and Verified quality is also positive throughout.

For Cosmos v2.5 I2V at pool size four, mean within-input Spearman correlation is 0.297, with 95% interval $[0.209, 0.383]$, and pairwise concordance, which is identical at every pool size, is 0.630, with interval $[0.591, 0.668]$.

F.3 Early selection

Figure 11 compares table-aligned prefix selection for Cosmos v2.5 and the two Nano configurations. Cosmos v2.5 first exceeds WMReward after 8/10 steps on I2V/V2V. At ten steps, Cosmos v2.5 selection already slightly exceeds completed-video WMReward on both tasks. It agrees with the full-schedule PRS decision on 72.7%/56.1% of I2V/V2V inputs, which indicates that the ranking has not yet fully stabilized at this point.

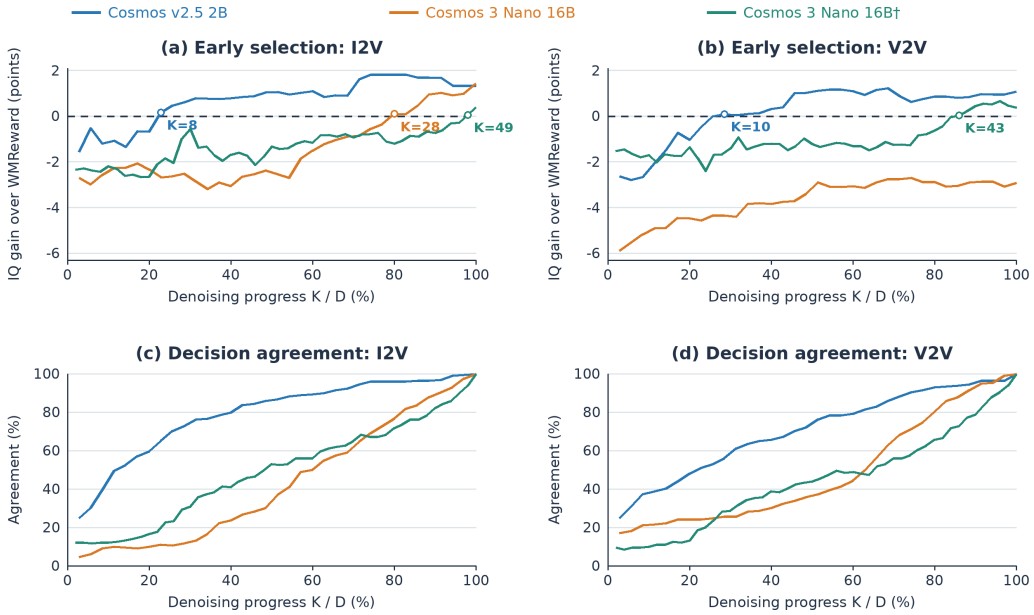


Figure 11: **Early selection and decision agreement.** (a,b) Prefix-selected Physics-IQ Verified gain over WMReward using the PRS settings in Table 12; zero marks the WMReward reference. (c,d) Agreement with the full-schedule PRS decision. Horizontal axes show the fraction of denoising steps used for selection. † denotes the expanded-prompt Nano configuration in Table 2. Protocol: Appendix F.1.

F.4 Selection cost

Continuing only the selected candidate saves candidate-steps as well as the decodes and reward evaluations that completed-video selection needs. Table 13 times both policies end to end on the same Physical Registry model. On Cosmos v2.5 the measured wall-clock ratio is $2.84\times$ against the $2.80\times$ candidate-step reduction reported in Section 4.4. Across backbones, measured wall-clock speedups range from $2.84\times$ to $4.06\times$ and broadly track the reduction in candidate-steps, while also reflecting backbone-specific sampling, decoding, and reward-evaluation costs.

Backbone	Policy	Candidate-steps	Decodes	Wall-clock (s)	Speed-up
Cosmos v2.5 2B	Complete BoN + WMReward	350	10	4,031	1.00×
	PRS ($K = 10$ of $D = 35$)	125	1	1,420	2.84×
Cosmos 3 Nano 16B†	Complete BoN + WMReward	500	10	1,985	1.00×
	PRS ($K = 10$ of $D = 50$)	140	1	614	3.23×
WAN 2.2 I2V-A14B	Complete BoN + WMReward	400	10	10,748	1.00×
	PRS ($K = 10$ of $D = 40$)	130	1	3,289	3.27×
WAN 2.2 TI2V-5B	Complete BoN + WMReward	500	10	2,834	1.00×
	PRS ($K = 10$ of $D = 50$)	140	1	697	4.06×

Table 13: **End-to-end cost of selection.** One input, $N = 10$ candidates, both policies run on the same Physical Registry model inside one job on a single H100 80 GB; model loading is excluded. Complete BoN times sampling, all ten VAE decodes, and the ten WMReward evaluations; PRS times the ten prefixes, the continuation of the selected candidate, and its single decode. † Expanded-prompt setting (Table 7).

G REGISTRY-MEDIATED MOTION TRANSFER

G.1 SETUP AND SCOPE

To test whether registry content can influence generated motion, we replace the native forecast of a receiver scene (the white-sphere scene) with source-video JEPA features on Cosmos 3 Nano 16B with Physical Registry. Both configurations use the white-sphere image and general prompt shown in Figures 12 and 13. The conditioning image, text, initial noise, and sampler are held fixed. The text specifies no falling, bouncing, or settling behavior.

Experimental setup. The native-forecast configuration combines one source-context feature with seven predicted features. The transfer configuration instead uses eight features from a frozen JEPA video encoder applied to the source video. Both representations have shape $8 \times 196 \times 5632$ and enter the same trained projection and video-generation interface. The source features are full video representations, containing scene content as well as dynamics. For this comparison, both configurations use the same factor-two amplification of the projected registry residual in the conditional branch, $P(0) + 2[P(J) - P(0)]$. The unconditional branch uses $P(J)$ without amplification. Here P is the trained projection and J the chosen registry input. Both receiver outputs use the same inference configuration: 50 denoising steps, classifier-free guidance scale 3.0, flow shift 4.75, and 1280×720 outputs with 121 frames at 24 fps.

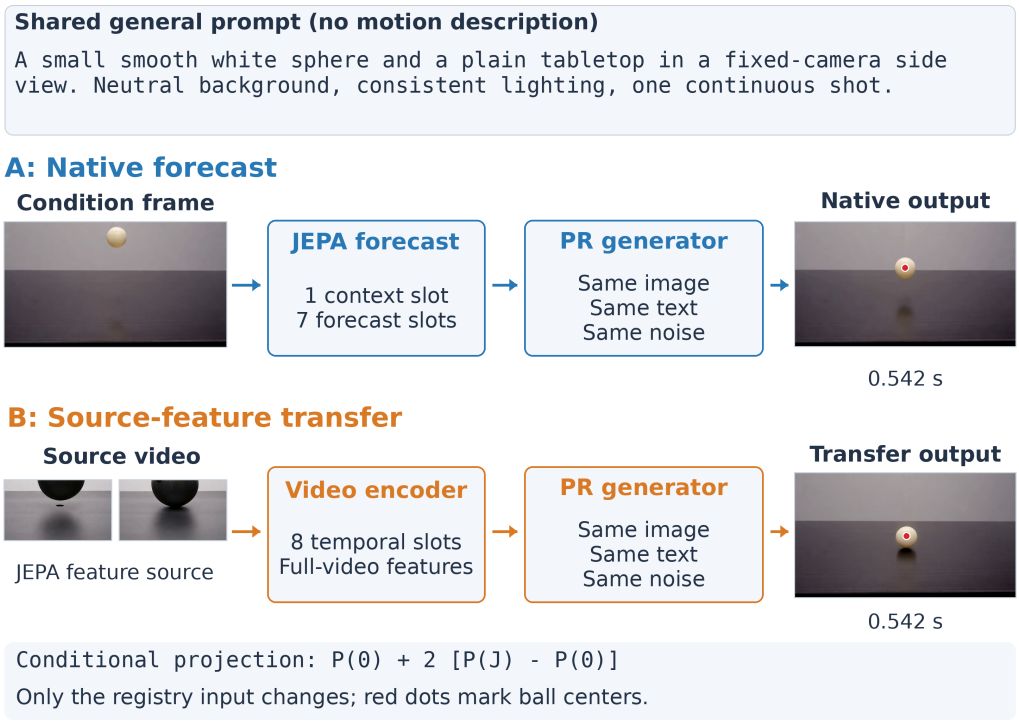


Figure 12: **Registry-source configurations for motion transfer.** The native configuration uses the conditioning image's JEPA context and forecast; the transfer configuration uses video-encoded features from the source video. Both use the same conditioning frame, general text, initial noise, and projection recipe. The exact prompt is shown in a monospace block and contains no motion description.

G.2 SOURCE EXAMPLES

Each example is a matched pair: A and B share the conditioning image, prompt, initial noise, and sampler, and differ only in the registry source. The two example pairs use the same generation settings. The two sources carry different parts of a fall, so each example is analyzed where its source differs from the native forecast: after landing for the bowling-ball source and during the approach for the balloon source.

Near-contact source. The native-forecast output (A) shows three diminishing upward returns before settling, whereas the transfer output (B) settles with much smaller upward motion, resembling the source's near-contact behavior.

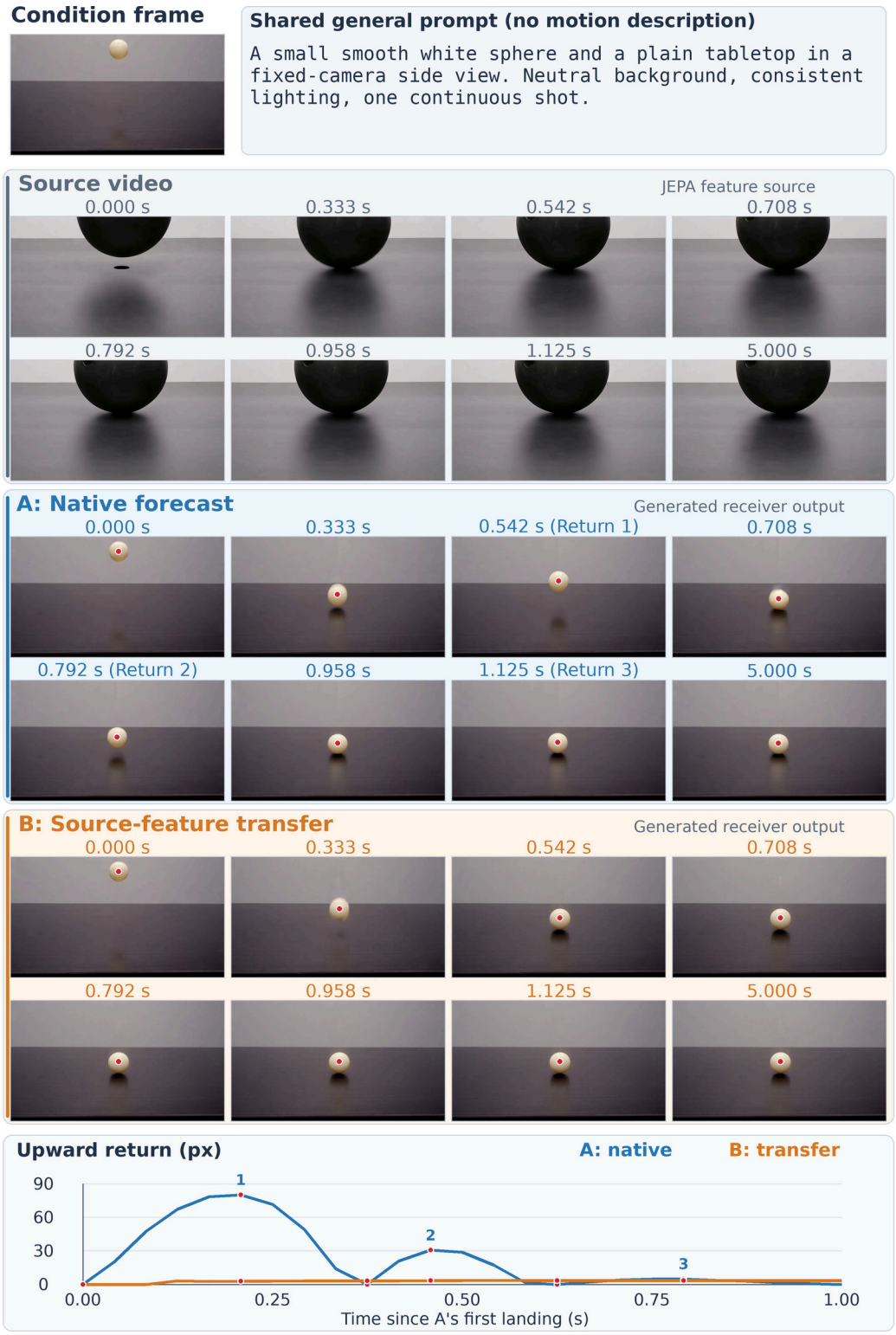


Figure 13: **Motion transfer with matched inputs.** The source supplies JEPA features; A and B are generated receiver outputs. Red dots track ball centers and rows share timestamps. The plot shows the upward return $r(t) = \max_{s \le t} y(s) - y(t)$, with image y increasing downward, over the second after A's first landing (0.333 s). B first lands at 0.417 s. Numbers mark A's three diminishing returns.

Center-motion comparison. Across the full clips, maximum upward return decreases from 80.1 to 3.6 pixels (Table 14).

Configuration	Valid frames	Vertical span	Upward return	Path length
A: Native forecast	121/121	289.7	80.1	526.9
B: Transfer	121/121	281.1	3.6	290.9

Table 14: **Image-plane ball-center movement in the matched pair.** Both outputs use Cosmos 3 Nano 16B with Physical Registry, over all 121 frames. Image y increases downward, and distances are image-plane pixels. Definitions in Figure 13.

Slow-descent source. A balloon descends continuously to the table, so its transfer is read before contact, on the descent itself. The receiver model, prompt, sampler, and factor-two conditional amplification follow the settings of Figure 13, and its native-forecast output A is a separate generation.

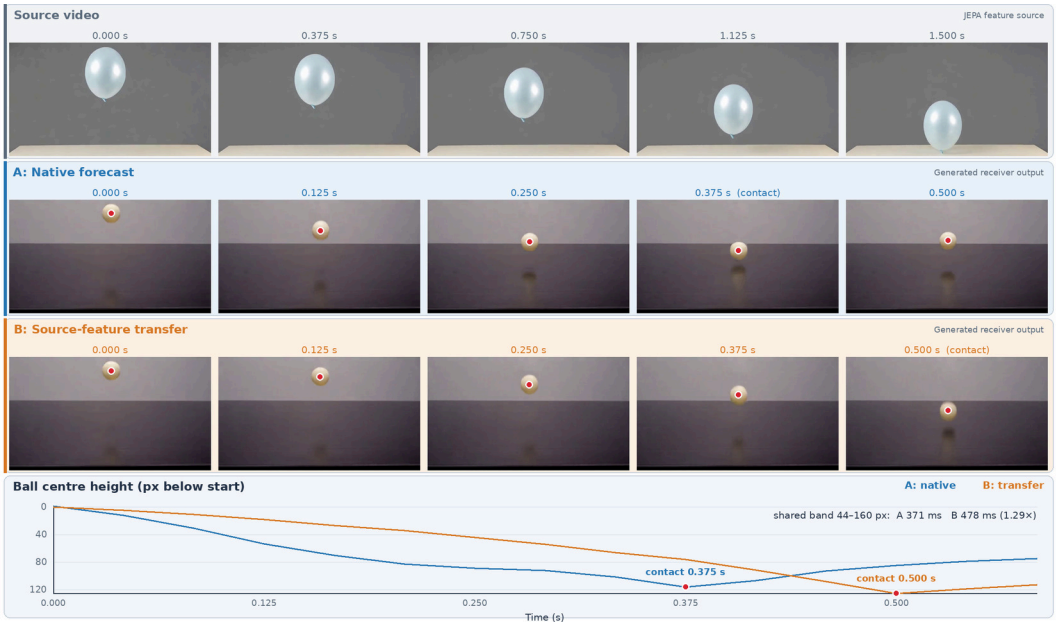


Figure 14: **Descent transfer from a slow-descent source.** A and B share the prompt, initial noise, sampler, and amplification, and differ only in the registry source. Settings follow Figure 13, but A is a separate native-forecast generation. Red dots track ball centers and rows share timestamps. The plot follows each center to first table contact, with image y increasing downward and contact marked on each curve.

Pre-contact descent. With the source's features in the registry, the receiver ball descends more slowly and reaches the table later (Figure 14). The native forecast takes 371 ms to descend and the transfer 478 ms, a ratio of $1.29\times$. Mean descent speed falls from 312 to 243 pixels per second. First table contact moves from frame 9 to frame 12, from 0.375 s to 0.500 s. The upward return after that contact also drops, from 40.9 to 12.1 pixels, consistent with the settling behavior in Figure 13. Silhouette areas at the first frame differ by 1.0%, making an object-scale explanation of the slower descent unlikely.

Across these two source interventions, changing the registry source while holding the receiver inputs fixed shifts the generated motion of the same receiver.

H Extended Mechanistic Diagnostics

H.1 Registry Feature Visualization

Registry activation maps characterize the spatial organization of each registry slot. The main-text visualization in Section 4.6 aligns five future times from one balloon-inflation trajectory and visualizes entity and action conditioning, foreground-query attention to video tokens and to the aligned registry slots, and slot-wise registry activation.

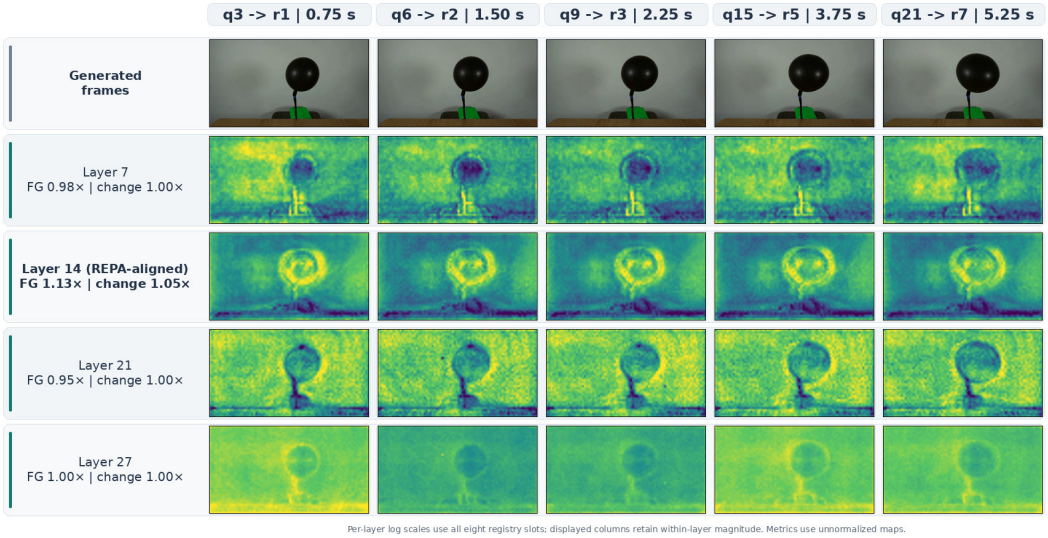


Figure 15: **Registry feature organization across DiT layers.** Columns align five generated frames with their registry slots. Heat rows show hidden channel variance at denoising step 16 for layers 7, 14, 21, and 27. Each layer uses an independent viridis log scale over all eight slots, so colors preserve within-layer slot magnitude.

For the depth sweep, we fix denoising step 16 and measure variance across hidden feature channels at every registry token. Layers 7, 21, and 27 retain broad scene-level activation, while the alignment-supervised layer 14 exhibits a response around the balloon and expanding contour. At this layer, foreground enrichment is $1.13\times$ and change-region enrichment is $1.05\times$. This layer shows the largest measured enrichment among these four layers. Each layer's log scale uses the 0.5th–99.5th percentiles over all eight registry slots. Foreground and change-region enrichment use the original unnormalized maps, so the spatial measurements are separate from the layer-specific display normalization.

I QUALITATIVE COMPARISONS

I.1 ADDITIONAL QUALITATIVE COMPARISONS

Figure 16 covers the five physics categories of Figure 3 on Cosmos v2.5, and Figures 17 and 18 provide additional comparisons on the same backbone. In each block, the top row shows the backbone and the bottom row shows Physical Registry built on the same backbone. Both rows start from the same conditioning input, and frames are sampled at eight equally spaced times of the generated sequence.

In these examples, Physical Registry completes several intended events that the backbone leaves unrealized, while avoiding some duplicated or spurious objects. On Cosmos v2.5 (Figure 16), Physical Registry propagates the domino topple after contact and extinguishes a lit match once it reaches the water. In set A (Figure 17), the backbone duplicates the released ball and the magnet-sequence block, while Physical Registry keeps each single. In set B (Figure 18), Physical Registry flattens the styrofoam cup under the descending kettlebell, while the backbone leaves it standing.



Figure 16: **Qualitative comparisons on Cosmos v2.5.** One scenario per physics category, with the layout of Figure 3; here the top row is the Cosmos v2.5 backbone.

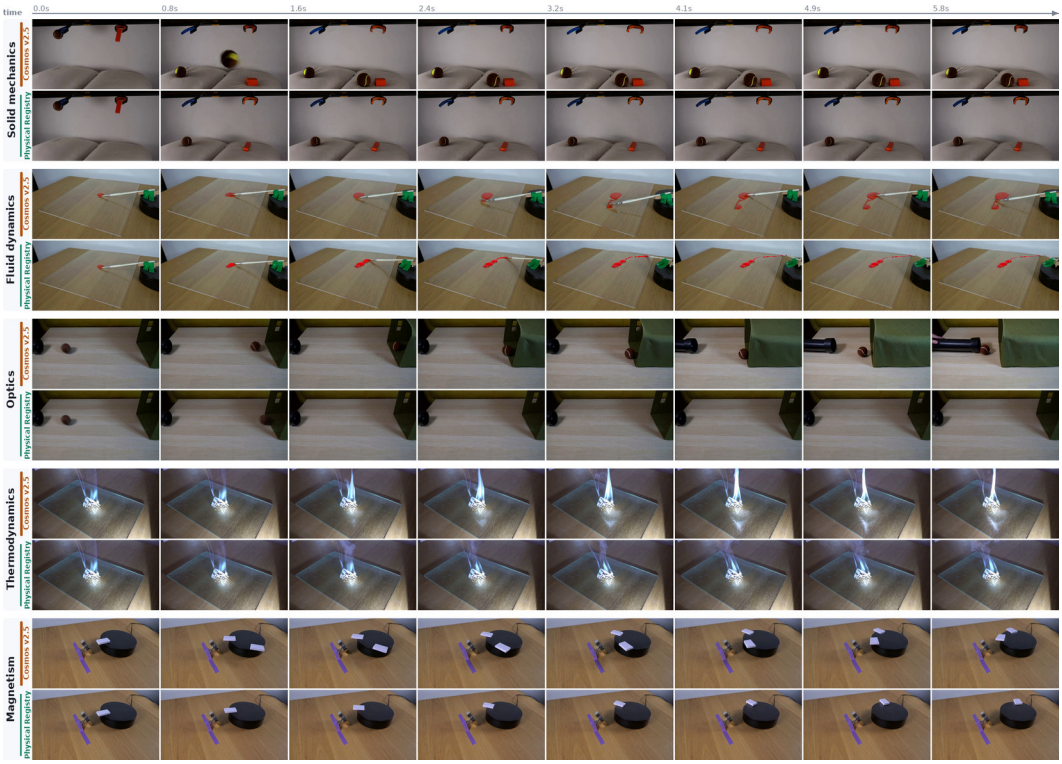


Figure 17: **Additional qualitative comparisons (set A).** Layout as in Figure 3.

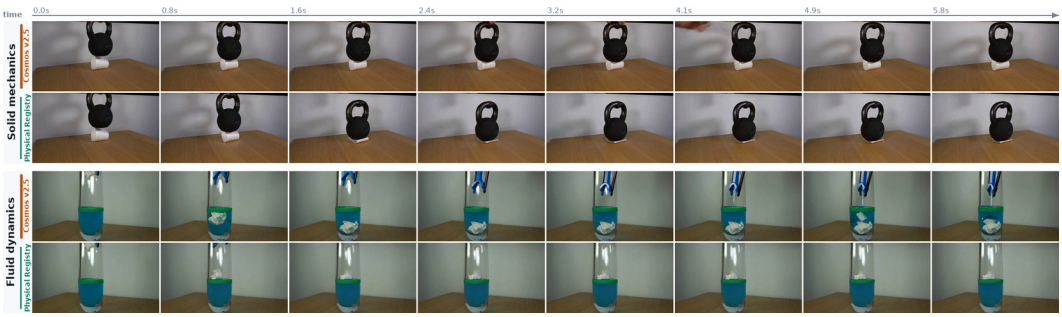


Figure 18: **Additional qualitative comparisons (set B).** Layout as in Figure 3.

I.2 WorldModelBench Qualitative Examples

Figures 19 and 20 show eight WorldModelBench generations from Cosmos 3 Nano 16B with Physical Registry, grouped into robotics and driving/industrial/natural scenes. Each text prompt is followed by eight evenly spaced generated frames in a large 4×2 storyboard, read left to right and then top to bottom. These examples illustrate the range of robotics, driving, industrial, and natural-scene prompts covered by the evaluation.

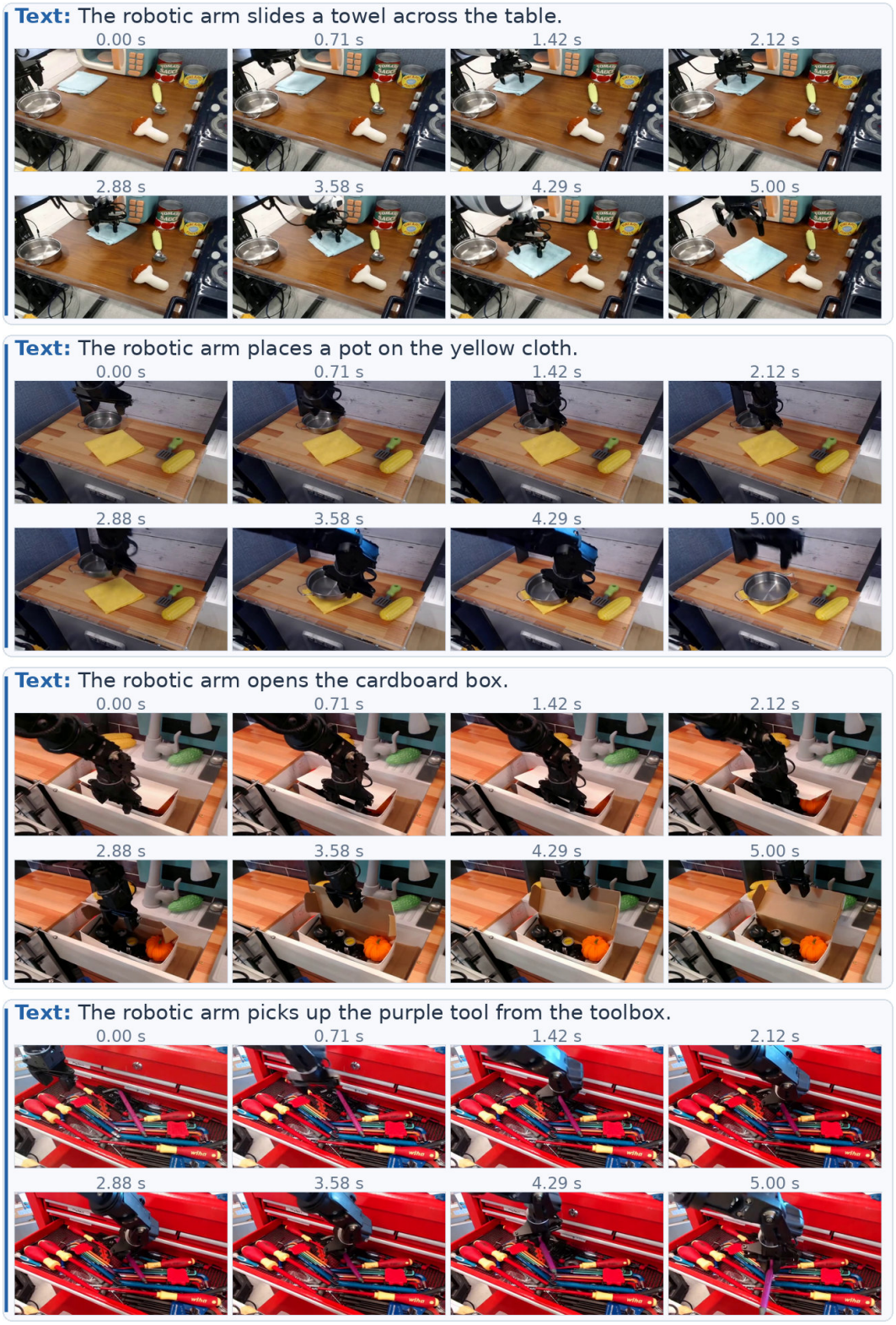


Figure 19: **WorldModelBench robotics: Ours (Cosmos 3 Nano 16B with Physical Registry).** Each prompt precedes eight video frames in a 4×2 storyboard, read left to right and then top to bottom. Times above the frames are seconds from video start, shared across both WorldModelBench figures.



Figure 20: **WorldModelBench driving, industry, and nature: Ours (Cosmos 3 Nano 16B with Physical Registry).** Prompts, frame order, and times in seconds follow Figure 19.

I.3 Additional Attention and Registry Visualizations

We extend Figure 5 with six Physics-IQ examples spanning solid mechanics, fluid dynamics, optics, and magnetism (Motamed et al., 2026). All panels show the same six rows: generated frames, entity and action cross-attention, self-attention to video and registry slots, and registry activation. Captions identify entity words and event verbs. Appendix A gives the visualization protocol. Across these examples, registry attention and activation tend to concentrate more locally around interacting or changing regions, whereas text cross-attention is often broader.

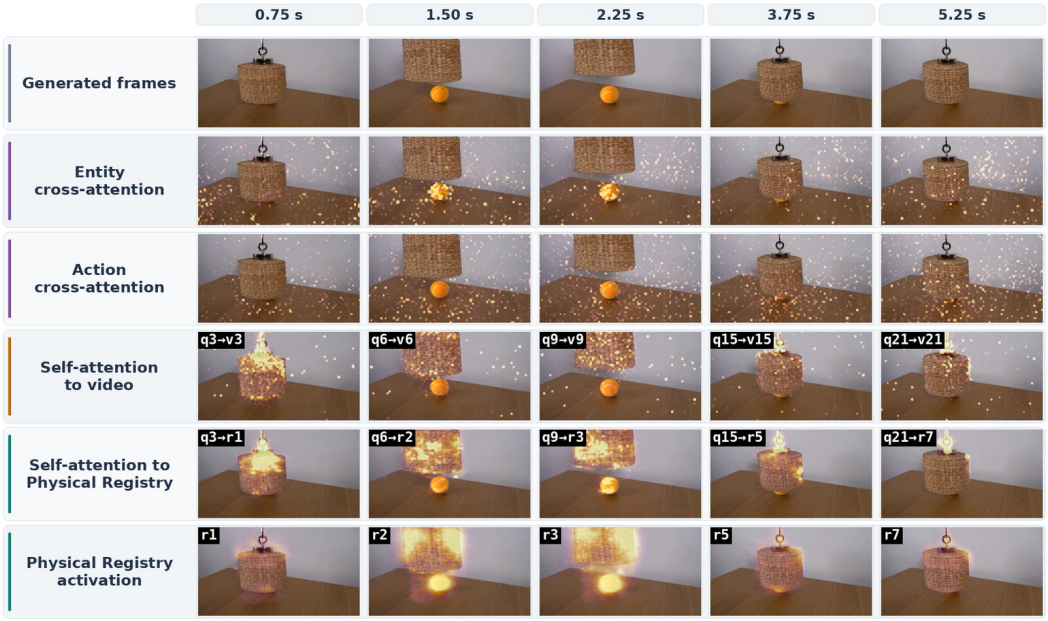


Figure 21: **Solid mechanics: object occlusion.** Entity text: *ball*. Action text: *lowered*.

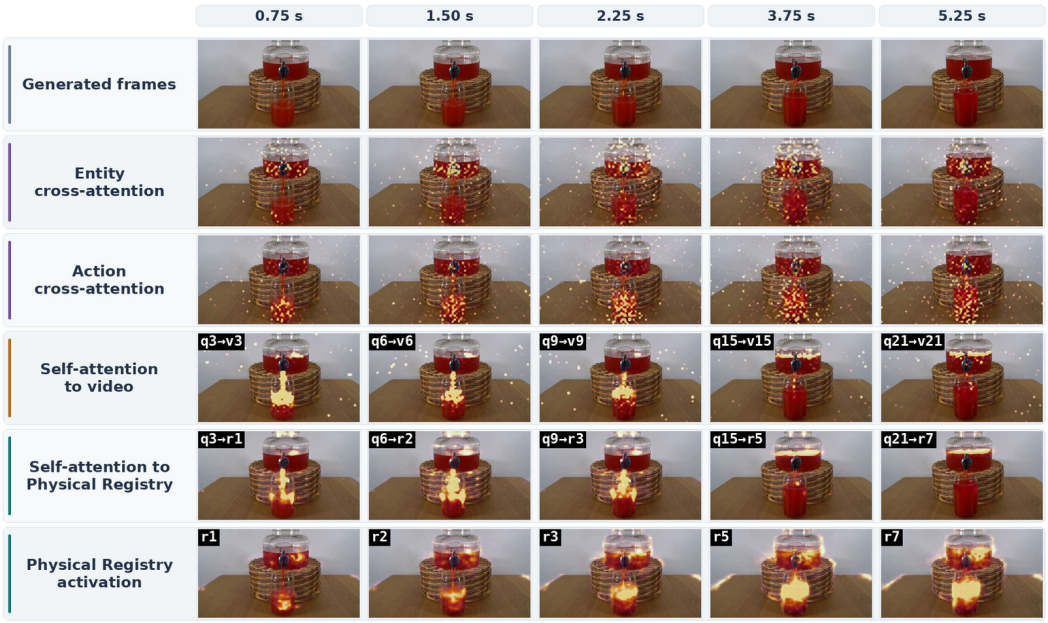


Figure 22: **Fluid dynamics: filling a glass.** Entity text: *glass, red*. Action text: *filled*.

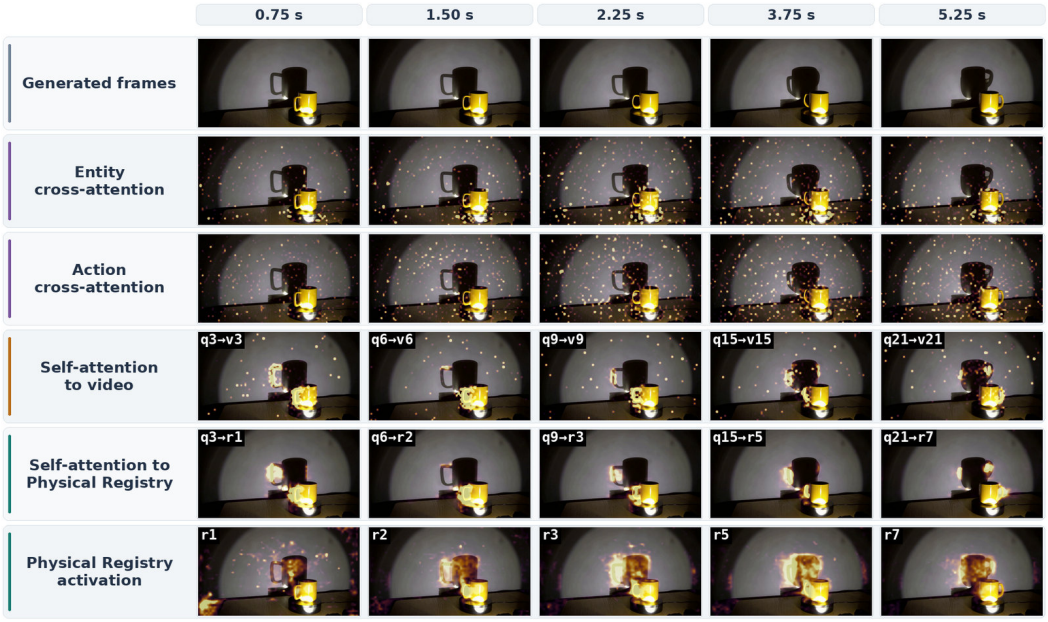


Figure 23: **Optics: a mug and its cast shadow.** Entity text: *mug*. Action text: *placed, rotating, rotates, casting*.

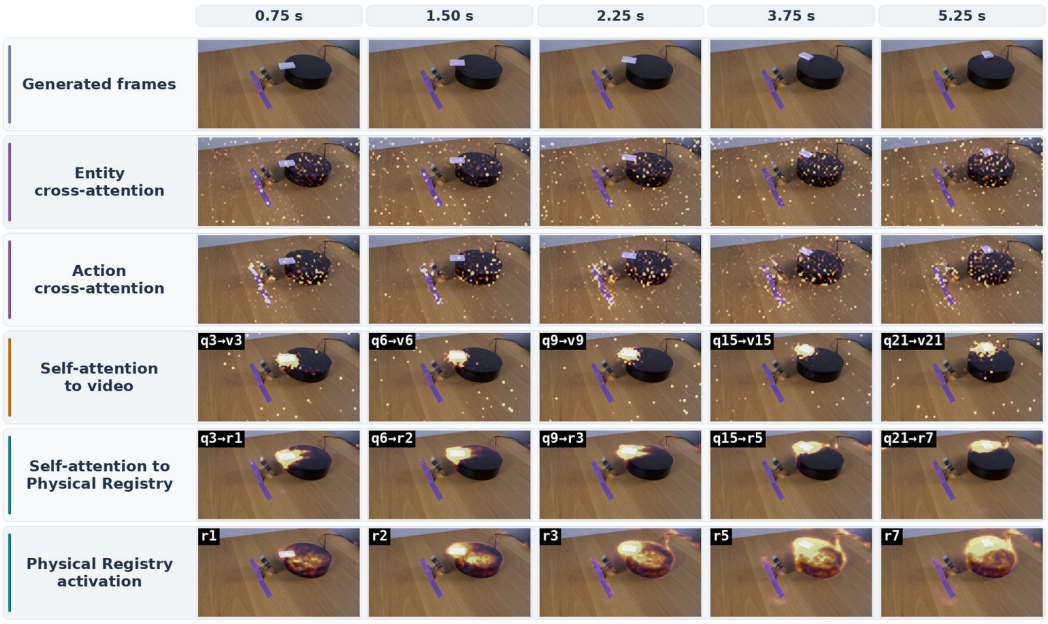


Figure 24: **Magnetism: a magnet–domino interaction.** Entity text: *domino, magnet*. Action text: *placed, rotating*.

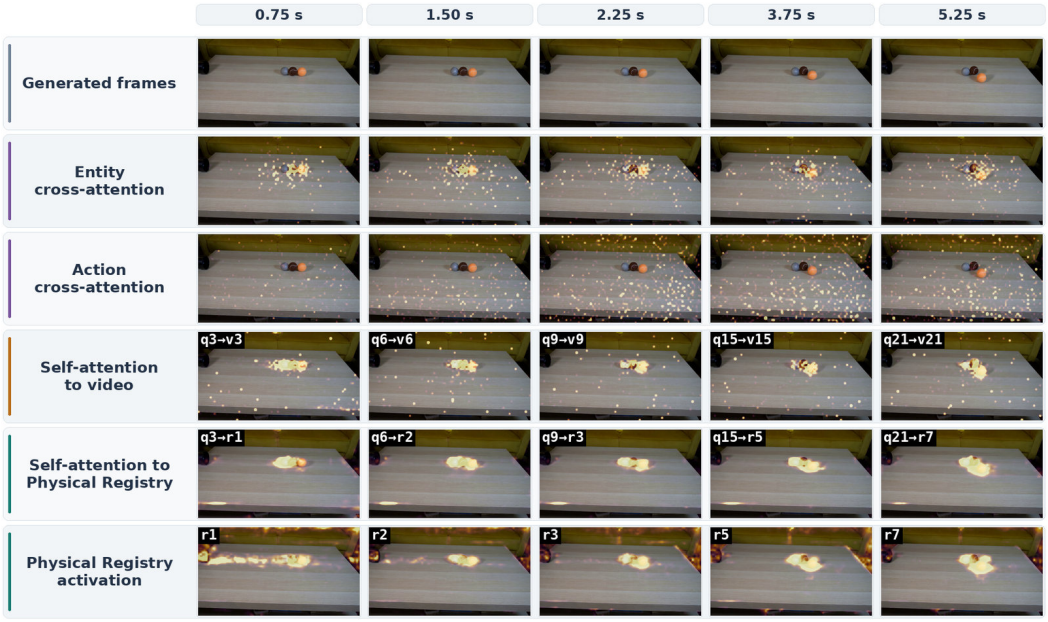


Figure 25: **Solid mechanics: rolling-ball interaction.** Entity text: *ball, balls*. Action text: *rolls*.

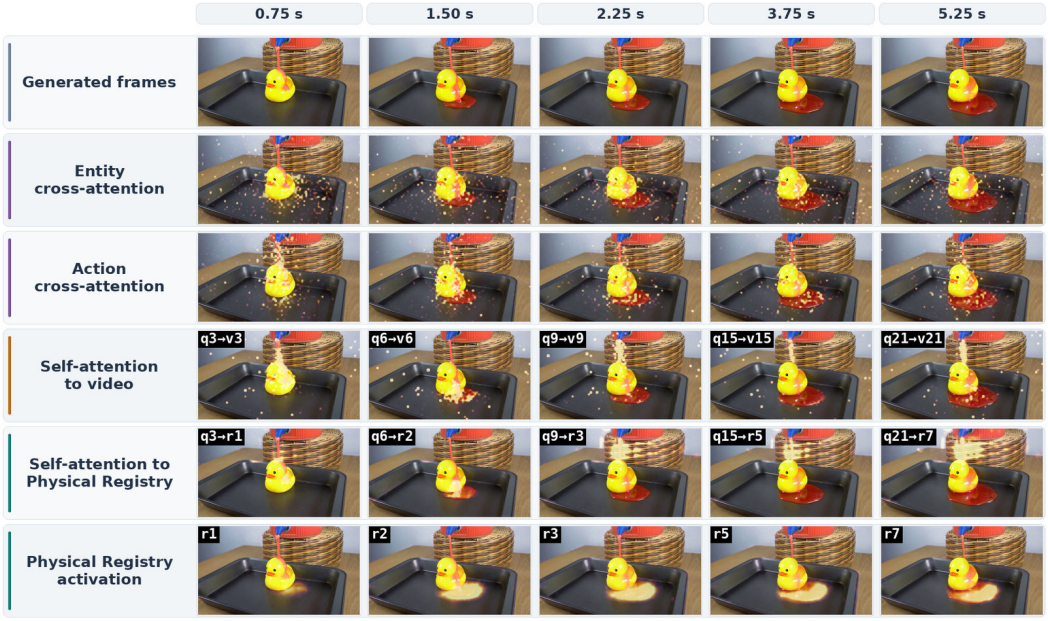


Figure 26: **Fluid dynamics: pouring liquid over a duck.** Entity text: *liquid, duck*. Action text: *pours*.