PHYLATENT: LEARNING DYNAMICS-RELEVANT REPRESENTATIONS FOR JEPA WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose PhyLatent, a dynamics-relevant training objective for Joint-Embedding Predictive Architecture (JEPA) world models. Our key observation is that preventing global latent collapse does not necessarily ensure that the learned representation preserves physically meaningful state and action relationships. We identify three failure modes: sensitivity to appearance changes that leave the physical state unchanged, insufficient separation of distinct physical states, and insufficient separation of different action-conditioned futures. We refer to these as Physical Invariance Collapse, Physical Distinguishability Collapse, and Counterfactual Dynamics Collapse, respectively. PhyLatent targets these failures through three coordinated training pathways, implemented with static visual invariance, physical state grounding, future representation alignment, counterfactual branch separation, and latent denoising. On OGBench-Cube, PhyLatent reduces the three collapse rates by **43.9%**, **27.9%**, and **47.1%**, respectively, while improving model predictive control (MPC) success by **12.0 percentage points** (**17.2% relative**). Across four visual-control tasks, average planning success increases by **6.62 percentage points** (**8.3% relative**). These results show that global non-collapse alone is insufficient for learning a reliable JEPA world-model state space, and that explicitly preserving dynamics-relevant structure can improve closed-loop planning.

1 INTRODUCTION

JEPAs learn compact representations for future prediction without reconstructing pixel-level details (LeCun, 2022). Recent work extends this idea to action-conditioned world models, where predicted latents support control and planning (Assran et al., 2025; Maes et al., 2026b). LeWorldModel (LeWM), for example, combines latent prediction with isotropic Gaussian regularization to maintain a globally non-degenerate latent distribution, and performs MPC directly in the learned representation space (Balestriero & LeCun, 2025; Maes et al., 2026b).

However, we observe a limitation of global anti-collapse regularization: a globally non-degenerate latent space is not necessarily reliable for planning. The latent space may still change with irrelevant visual appearance, fail to distinguish different physical states, or fail to reflect different action outcomes. We examine these failures in Sec. 3. Specifically, we distinguish:

- **Dynamics-relevant structure:** physical state information needed for prediction and planning, such as robot and object positions and poses, together with how actions change them.
- **Appearance variation:** changes in brightness, color, or texture that leave this physical information unchanged.

This raises a direct question: **does global non-collapse preserve dynamics-relevant structure?**

In LeWM, we identify three forms of dynamics-relevant collapse. *Physical Invariance Collapse* occurs when changing only visual appearance breaks a physical relation that was previously represented correctly. *Physical Distinguishability Collapse* occurs when differences in current physical states are not sufficiently reflected in latent space. *Counterfactual Dynamics Collapse* occurs when different actions from the same current state yield divergent physical outcomes but indistinguishable predictions from the model. Figure 1 summarizes the three failures.

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.

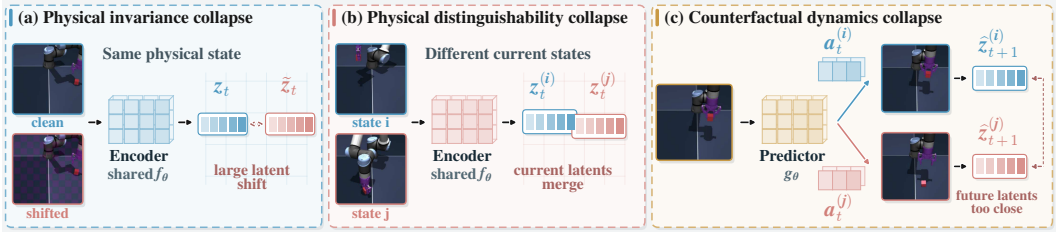


Figure 1: **Motivation for PhyLatent.** Global non-collapse can coexist with **(a)** appearance-induced latent shifts despite unchanged physical state, **(b)** insufficient separation of distinct physical states, and **(c)** insufficient predicted separation of divergent action-conditioned futures from the same state. Here, z_t and $\tilde{z}_t$ denote clean and appearance-shifted latents, $(i), (j)$ index alternative states or action branches, and $\hat{z}$ denotes a predicted future latent.

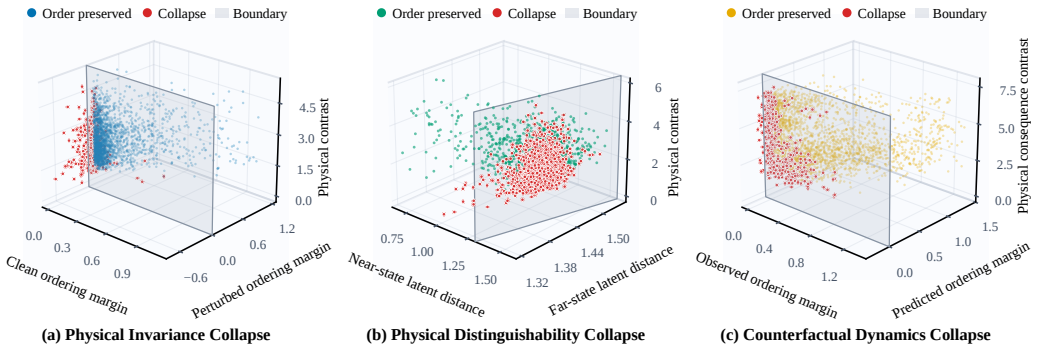
To address these failures, we propose **PhyLatent**, a dynamics-relevant representation learning method for JEPA world models. PhyLatent organizes its training objectives into three coordinated pathways. It suppresses appearance-induced changes, preserves task-relevant physical information, aligns predicted and observed futures, preserves differences among action-conditioned futures, and regularizes local future structure through latent denoising. During training, the Physical Grounding pathway uses physical state information available from simulator trajectories, such as robot and object positions and poses. These physical-state targets and all additional auxiliary components are not required at inference time.

On OGBench-Cube (Park et al., 2025), PhyLatent reduces Physical Invariance, Physical Distinguishability, and Counterfactual Dynamics Collapse from **19.58%**, **45.51%**, and **14.47%** to **10.99%**, **32.82%**, and **7.65%**, respectively, while improving planning success from **69.67%** to **81.67%**. Across four tasks, average success increases from **79.92%** to **86.54%**. These results show that global non-collapse is only a starting point: effective world-model representations benefit from preserving dynamics-relevant structure.

2 Related Work

JEPA world models. JEPAs learn abstract representations by predicting target representations rather than reconstructing pixels (LeCun, 2022). I-JEPA applies joint-embedding prediction to masked image regions (Assran et al., 2023), building on masked-image modeling paradigms (He et al., 2022; Bao et al., 2022), while V-JEPA extends feature prediction to video (Bardes et al., 2024). More recent work brings joint-embedding prediction closer to control: V-JEPA 2 and LeWM introduce action-conditioned prediction for embodied understanding and planning (Assran et al., 2025; Maes et al., 2026b), while PLDM and DINO-WM use latent dynamics or pretrained visual features for planning (Sobal et al., 2025; Zhou et al., 2025). SD-JEPA instead explicitly structures the latent space by separating progression from content (Thil et al., 2026). Despite these differences, their usefulness for control depends on whether the learned representation preserves the physical state variables and action consequences required for planning.

Representation collapse. Joint-embedding models require mechanisms that prevent representations from degenerating into trivial or low-rank embeddings (Grill et al., 2020; Chen & He, 2021; Zbontar et al., 2021; Bardes et al., 2022; Balestriero & LeCun, 2025). Existing approaches either use predictive training mechanisms, including contrastive objectives, asymmetric prediction, momentum targets, self-distillation, and masked prediction, or directly regularize the representation distribution through redundancy reduction and variance–covariance objectives (van den Oord et al., 2018; Chen et al., 2020; He et al., 2020; Caron et al., 2020; Grill et al., 2020; Chen & He, 2021; Caron et al., 2021; Zbontar et al., 2021; Bardes et al., 2022; Assran et al., 2023; Bardes et al., 2024). Along the latter direction, LeJEPA introduces Sketched Isotropic Gaussian Regularization (SIGReg) to maintain a globally non-degenerate embedding distribution, which LeWM applies to action-conditioned latent prediction (Balestriero & LeCun, 2025; Maes et al., 2026b). Sub-JEPA relaxes the full-space Gaussian constraint to multiple random subspaces (Zhao et al., 2026), while QQWorld further constrains deviations in the



(a) Physical Invariance Collapse **(b) Physical Distinguishability Collapse** **(c) Counterfactual Dynamics Collapse**

Figure 2: **Sample-level dynamics-relevant collapse on Cube.** Each point is an eligible LeWM comparison; red marks ordering violations, other colors mark preserved orderings, and translucent planes show their boundaries. Panels (a)–(c) show Physical Invariance, Physical Distinguishability, and Counterfactual Dynamics Collapse, respectively.

tails of the latent distribution (Yu et al., 2026). These methods maintain global non-degeneracy, but do not directly constrain the local structure associated with physical states and action-conditioned futures.

Dynamics-aware representations and latent control. A complementary line of work focuses on representations that retain structure useful for control, including temporal continuity, controllability, transition prediction, bisimulation, and inverse dynamics (Jonschkowski & Brock, 2015; Sermanet et al., 2018; Gelada et al., 2019; Zhang et al., 2021; Wang et al., 2022; Pathak et al., 2017). For example, Sensorimotor World Models (SMWM) use inverse-dynamics regularization to learn action-aligned latent representations (Ivashkov et al., 2026). Learned latent dynamics are also used directly for decision making: PlaNet performs planning in latent space, Dreamer learns behavior through latent imagination, and TD-MPC2 performs local trajectory optimization using a learned latent world model (Hafner et al., 2019; 2020; Hansen et al., 2024). These works show that latent dynamics can support effective control, but they do not specifically ask whether global non-collapse in JEPA world models preserves the local physical and counterfactual relations required for planning. PhyLatent targets this representation-level gap.

3 Why Does LeWorldModel Exhibit Dynamics-Relevant Collapse?

Figure 2 shows eligible LeWM comparisons on Cube. Ordering violations occupy non-negligible regions of all three diagnostic spaces, indicating that collapse extends beyond isolated rollout examples.

We inspect three representative decisions from failed Cube rollouts of the same LeWM checkpoint by restoring the simulator at the relevant planning step. For the Distinguishability and Counterfactual cases, we execute alternative action sequences from the original cross-entropy method (CEM) candidate pool from the same state. Lower latent-space goal cost indicates an outcome judged closer to the goal. Supplement: “Collapse Diagnostic Definitions”; “Decision-Level Failure Cases”.

Physical Invariance Collapse. Appearance-only perturbations change LeWM’s action preferences and lead to a failed rollout despite unchanged robot, cube, and goal states. Restoring the original appearance only for the first planning call, from the same simulator state and with the remaining procedure unchanged, yields success. Thus, visual appearance alone can change the planning decision.

Physical Distinguishability Collapse. The original plan leaves the cube **10.00 cm** from the goal; an alternative candidate reaches **1.49 cm** and succeeds. Yet directly encoding the realized futures assigns the failed outcome a lower goal cost (**207.37 vs. 345.24**). The physically better state is therefore ranked worse by the encoder, locating the inconsistency in the realized-state representation rather than only in future prediction.

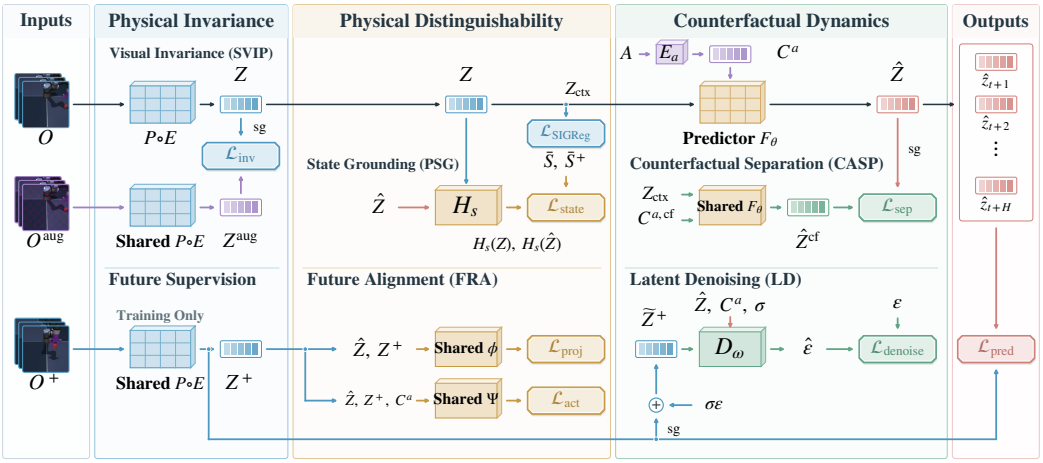


Figure 3: **Overview of PhyLatent.** Three training-time pathways augment the shared LeWM backbone: Physical Invariance (SVIP) constrains appearance consistency; Physical Grounding (PSG, FRA) preserves physical-state information and aligns predicted with observed futures; Counterfactual Dynamics (CASP, LD) structures action-conditioned futures. The original prediction objective is retained.

Counterfactual Dynamics Collapse. For the original and alternative plans, directly encoded realized-future costs (**227.16 vs. 81.32**) correctly favor the alternative, but predicted costs (**2.87 vs. 55.00**) reverse this ordering and favor the failed plan. Executing the alternative from the same state succeeds. Here, the observed inconsistency arises during action-conditioned prediction, despite correct encoded future ordering.

These interventions locate observed planning inconsistencies without isolating a unique internal cause. Together with Figure 2, they show that global non-collapse does not guarantee preservation of local physical and action-conditioned relations needed for planning.

4 PhyLatent

Figure 3 summarizes the overall PhyLatent architecture. PhyLatent retains the LeWM encoder, action encoder, predictor, and inference-time planning pipeline, while introducing three coordinated training pathways corresponding to the failure modes identified in Sec. 3. Physical Invariance suppresses appearance-induced variation, Physical Grounding preserves task-relevant state distinctions and aligns future representations, and Counterfactual Dynamics structures action-conditioned future predictions. Although motivated by distinct failure modes, the pathways jointly shape the same latent representation; all additional heads and supervision are removed at inference time.

See "Implementation and Evaluation Details" in the supplement.

4.1 Physical Invariance

For each observation sequence O, we construct an appearance-augmented view O^{aug} while leaving the underlying physical state unchanged. The original and augmented observations pass through the same JEPA encoder and projector, producing Z and Z^{aug}, respectively. We introduce the **Static Visual Invariance Penalty (SVIP)**:

$$\mathcal{L}_{\text{inv}} = d_{\text{N}}(Z^{\text{aug}}, \text{sg}(Z)) \,. \tag{1}$$

For feature vectors $X_i, Y_i \in \mathbb{R}^D$, we define

$$d_{\text{N}}(X, Y) = \frac{1}{ND} \sum_{i=1}^{N} \|\nu(X_i) - \nu(Y_i)\|_2^2 \,, \qquad \nu(x) = \frac{x}{\max(\|x\|_2, 10^{-12})} . \tag{2}$$

Here, N counts all non-feature positions, and $\mathrm{sg}(\cdot)$ denotes stop-gradient, which preserves its forward value while blocking gradients through the target branch (Chen & He, 2021). The detached clean representation serves as the reference, so SVIP penalizes appearance-induced latent changes without altering the underlying physical state.

4.2 Physical Grounding for Distinguishability

Physical State Grounding (PSG). When selected physical-state variables are available in the training trajectories, PSG uses them as auxiliary supervision to ground the latent representation in task-relevant physical information. Let $\bar{S}$ denote the per-dimension standardized physical targets and $\bar{S}^+$ the corresponding targets temporally aligned with the predicted future $\hat{Z}$. A shared state-prediction head H_s maps both encoded and predicted representations to these targets using mean squared error (MSE):

$$\mathcal{L}_{\text{state}} = \mathrm{MSE}\big(H_s(Z), \bar{S}\big) + \mathrm{MSE}\Big(H_s(\hat{Z}), \bar{S}^+\Big). \tag{3}$$

Sharing H_s across the two branches encourages physical distinctions to remain recoverable from both encoded and predicted latents.

Future Representation Alignment (FRA). FRA extends this grounding to future prediction by aligning $\hat{Z}$ with the corresponding encoded future Z^+. A shared projection head ϕ compares the two representations globally, whereas an action-query module Ψ summarizes each future relative to the executed action context. Let C^a denote the context-aligned action embeddings; their temporal summary forms the query in Ψ, and the future latent sequence provides the keys and values. We use

$$\mathcal{L}_{\text{align}} = \mathcal{L}_{\text{proj}} + \mathcal{L}_{\text{act}}, \tag{4}$$

where $\mathcal{L}_{\text{proj}} = d_{\text{N}}(\phi(\hat{Z}), \mathrm{sg}(\phi(Z^+)))$ and $\mathcal{L}_{\text{act}} = d_{\text{N}}(\Psi(\hat{Z}, C^a), \mathrm{sg}(\Psi(Z^+, C^a)))$. The two terms align the overall future representation and its action-conditioned summary, respectively.

4.3 Counterfactual Dynamics

Counterfactual Action Separation Penalty (CASP). From the same latent context, factual and counterfactual action sequences produce $\hat{Z}$ and $\hat{Z}^{\text{cf}}$ through the shared action encoder and predictor. Let $\delta^a_{b,\tau}$ and $\delta^z_{b,\tau}$ denote the normalized action difference and predicted-future separation for mini-batch sample b and prediction position τ, respectively. We define

$$\mathcal{L}_{\text{sep}} = \frac{1}{|\Omega|} \sum_{(b,\tau)\in\Omega} \big[\min\big(\gamma\delta^a_{b,\tau}, m_{\max}\big) - \delta^z_{b,\tau}\big]_+ . \tag{5}$$

Here, $[x]_+ = \max(x, 0)$, Ω retains sufficiently different action pairs, and γ and $m_{\max}$ are fixed task-specific hyperparameters. CASP uses action difference as a proxy for counterfactual branch diversity and does not explicitly identify outcome-equivalent action pairs.

Latent Denoising (LD). LD further regularizes the local neighborhood of the observed future. We perturb the detached target future as $\tilde{Z}^+ = \mathrm{sg}(Z^+) + \sigma\epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$ and σ is sampled from a task-specific noise range. A denoising head D_ω receives the noisy target future $\tilde{Z}^+$, predicted future $\hat{Z}$, action embedding C^a, and noise scale σ, and predicts the injected noise:

$$\mathcal{L}_{\text{denoise}} = \mathrm{MSE}\Big(D_\omega(\tilde{Z}^+, \hat{Z}, C^a, \sigma), \epsilon\Big). \tag{6}$$

The detached target provides a fixed reference, and the resulting denoising signal regularizes local future structure around the observed trajectory.

4.4 Training Objective

PhyLatent combines the three pathways with the original LeWM training objective. Let $\mathcal{L}_{\text{base}}$ denote the original LeWM objective, including latent prediction and SIGReg, and $\mathcal{A} = \{\text{inv}, \text{state}, \text{align}, \text{sep}, \text{denoise}\}$ the auxiliary objectives. The full objective is

$$\mathcal{L}_{\text{PhyLatent}} = \mathcal{L}_{\text{base}} + \sum_{i\in\mathcal{A}} \lambda_i \mathcal{L}_i. \tag{7}$$

Table 1: Closed-loop planning success (%) across four tasks. PhyLatent, LeWM, SMWM, and Sub-JEPA use matched paired evaluation (mean ± across-seed standard deviation). PLDM and DINO-WM use LeWM's reported baselines (Maes et al., 2026b), outside our paired-seed evaluation.

Method	Cube	TwoRooms	Reacher	PushT	Avg.
PhyLatent	81.67 ± 3.01	96.83 ± 2.14	82.17 ± 3.60	85.50 ± 3.62	**86.54**
LeWM	69.67 ± 5.13	83.17 ± 1.47	82.33 ± 2.58	84.50 ± 3.39	79.92
SMWM	**83.67 ± 2.16**	**100.00 ± 0.00**	73.17 ± 2.99	85.83 ± 3.13	85.67
Sub-JEPA	76.83 ± 3.37	94.17 ± 2.71	**84.83 ± 2.14**	**89.50 ± 2.59**	86.33
PLDM	65.00	97.00	78.00	78.00	79.50
DINO-WM (w/o proprio.)	86.00	100.00	79.00	74.00	84.75

Table 2: Collapse rates (%; lower is better); CF-H5 uses the planner horizon $H{=}5$. Definitions: "Collapse Diagnostic Definitions" in the supplement.

Metric	Cube		TwoRooms		Reacher		PushT	
	LeWM	PhyLatent	LeWM	PhyLatent	LeWM	PhyLatent	LeWM	PhyLatent
Inv. ↓	19.58	**10.99**	17.18	**4.69**	18.58	**7.66**	9.86	**3.38**
Dist. ↓	45.51	**32.82**	47.68	**26.67**	47.60	**33.46**	38.38	**30.71**
CF-H5 ↓	14.47	**7.65**	14.10	**9.23**	15.48	15.10	6.63	**2.09**

All auxiliary objectives are optimized jointly with fixed coefficients within each training run; there is no sequential priority mechanism. The coefficients are task-specific and were manually adjusted using preliminary experiments.

5 Experiments

We evaluate PhyLatent on OGBench-Cube (Park et al., 2025), TwoRooms (Sobal et al., 2025; Maes et al., 2026a;b), Reacher from the DeepMind Control Suite (Tunyasuvunakool et al., 2020), and PushT (Chi et al., 2023). Cube's robot–object geometry supports controlled analysis of all three failures (Sec. 3).

Primary evaluation protocol. Unless stated otherwise, models share sampled start–goal pairs and simulator initializations. CEM-MPC uses 300 candidates, 30 elites, prediction and receding horizons of 5 blocks, an action block size of 5, and a budget of 50 environment interactions. Cube, TwoRooms, and Reacher use 10 CEM updates; PushT uses 30. Primary success is reported as mean ± sample standard deviation across six paired evaluation seeds (supplement: "Implementation and Evaluation Details").

5.1 Cross-Task Planning Performance

Table 1 compares PhyLatent with LeWM (Maes et al., 2026b), SMWM (Ivashkov et al., 2026), and Sub-JEPA (Zhao et al., 2026) under the same paired evaluation protocol. PhyLatent raises the four-task average over LeWM from 79.92% to **86.54%** (+6.62 percentage points), with gains of 12.00 and 13.66 points on Cube and TwoRooms, a modest PushT improvement, and effectively tied Reacher performance. The backbone and planner are unchanged.

Matched baselines lead on different tasks: SMWM on Cube and TwoRooms, and Sub-JEPA on Reacher and PushT. PhyLatent's average gain is not uniform per-task dominance.

5.2 Dynamics-Relevant Latent Structure

Table 2 evaluates the three collapse modes from Sec. 3. Invariance and Distinguishability collapse decrease substantially on all four tasks. Counterfactual collapse also decreases in point estimate, but the Reacher improvement at the planner-aligned horizon $H{=}5$ is small.

These diagnostics are not the training losses: SVIP, PSG, and CASP optimize feature consistency, state prediction, and action-based future separation, whereas the diagnostics test relational ordering under appearance changes, physical-state differences, and realized action consequences. Sampling and comparison rules are fixed and shared across models.

Reacher separates improved current-state geometry from improved control: static diagnostics improve, but planning remains effectively tied. Its counterfactual gains at shorter horizons do not establish a gain at the planning horizon, where the paired confidence interval includes zero. Held-out rank agreement also improves on fresh random-policy states without the PSG head or binary collapse definitions. Absolute correlations remain modest, supporting improved distance ordering rather than an isometric physical-state representation or a causal explanation of planning gains (supplement: "Paired Statistics and Horizon Sensitivity").

5.3 Robustness to Visual Appearance Shifts

Figure 4 shows task-dependent robustness. On Cube, PhyLatent retains advantages of 18.00 and 20.00 percentage points under background and texture changes. TwoRooms and Reacher instead degrade sharply under these shifts; TwoRooms grounds only agent position, leaving other scene information unconstrained. The results support improved invariance of planning-relevant relations, but not invariance to arbitrary appearance shifts.

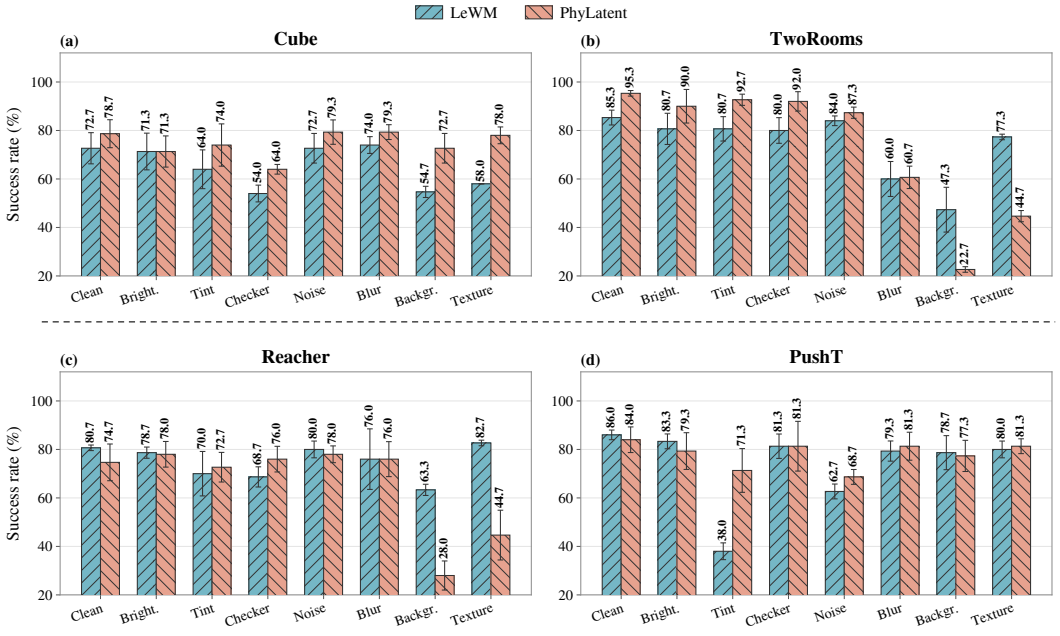


Figure 4: **Closed-loop robustness under visual appearance shifts.** Success (%) of LeWM and PhyLatent under clean observations and controlled appearance perturbations across four tasks.

5.4 Supervision and Hyperparameter Controls

In Table 3(a), the matched PSG-only control recovers much of the Cube and TwoRooms improvement but averages 81.75%, versus 86.54% for Full. The PSG-only gains on Cube and TwoRooms coexist with declines relative to LeWM on Reacher and PushT. Full largely avoids those declines, indicating a different cross-task profile rather than merely a larger average gain. This control does not isolate which remaining objective accounts for that difference. Jointly scaling all five auxiliary losses from $0.5\times$ to $2\times$ keeps Cube success within 80.17–81.67% while collapse rates improve (panel b). The result is stable over this tested weight range, but better diagnostics do not translate monotonically into better control.

Table 3: Controls for privileged-state supervision and auxiliary-loss weighting. (a) Mean closed-loop success (%) for LeWM, its matched PSG-only control, and PhyLatent across tasks. (b) Joint rescaling of all five PhyLatent auxiliary coefficients on Cube; success is mean ± across-seed standard deviation.

(a) Matched privileged-state control					
Method	**Cube**	**TwoRooms**	**Reacher**	**PushT**	**Avg.**
LeWM	69.67	83.17	**82.33**	84.50	79.92
LeWM + PSG	80.67	96.33	67.17	82.83	81.75
PhyLatent	**81.67**	**96.83**	82.17	**85.50**	**86.54**

(b) Cube auxiliary-weight sensitivity				
Scale	**Success ↑**	**Inv. ↓**	**Dist. ↓**	**CF-H5 ↓**
0.5×	80.17±3.43	12.54	34.70	9.48
1×	**81.67±3.01**	10.99	32.82	7.65
2×	80.83±3.60	**8.54**	**30.73**	**7.55**

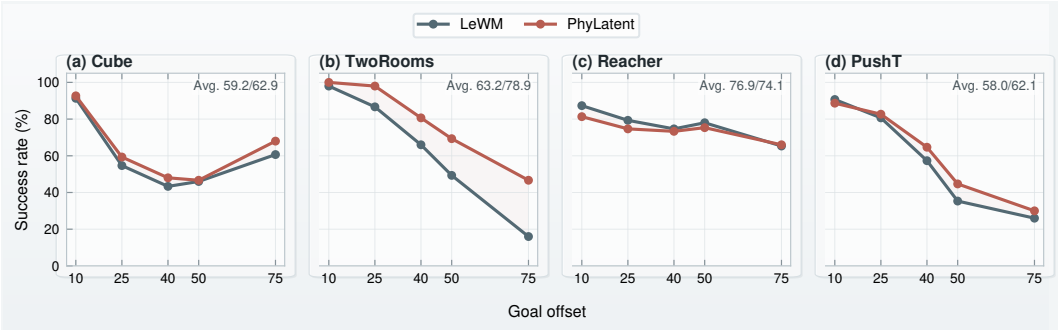
5.5 Pathway Interactions and Ablations on Cube

Table 4 shows non-additive pathway effects. Removing Physical Grounding or PSG reduces Cube success from 81.67% to 76.50% or 77.00%, respectively, while distinguishability and counterfactual ordering also deteriorate. Removing FRA instead lowers Dist. collapse from 32.82% to 28.85% while reducing success to 79.33%. Improving one diagnostic therefore need not improve control.

Table 4: Cube ablations. Non-Full columns remove the named component or group. Success is mean ± across-seed standard deviation; lower diagnostics are better. CF-H5 uses the ablation-family eligibility intersection (N=10,162), not the LeWM–Full intersection in Table 2, explaining the different Full values (supplement: "Cross-Task Pathway and Component Ablations").

	Full	**Physical Grounding**			**Invariance**	**Counterfactual Dynamics**		
Metric	**PhyLatent**	**Group**	**PSG**	**FRA**	**SVIP**	**Group**	**CASP**	**LD**
Success ↑	**81.67** ±3.01	76.50 ±3.15	77.00 ±3.63	79.33 ±1.51	80.67 ±4.27	81.17 ±2.23	79.00 ±0.89	79.50 ±1.97
Inv. ↓	10.99	11.85	13.44	8.46	9.67	9.74	9.46	6.93
Dist. ↓	32.82	40.11	39.33	28.85	29.71	27.26	28.42	28.81
CF-H5 ↓	5.934	13.629	13.836	7.567	8.611	5.737	6.908	6.288

5.6 Planning under Increasing Goal Separation



(e) Cube physical-displacement stratification					
Method	**Q1** **<0.039 cm**	**Q2** **0.039–9.101 cm**	**Q3** **9.101–19.966 cm**	**Q4** **19.966–49.888 cm**	**Q1→Q4** **Rel. Drop ↓**
LeWM	**100.00**	81.33	48.00	26.67	73.33
PhyLatent	**100.00**	**88.00**	**52.00**	**60.00**	**40.00**

Figure 5: **Planning under increasing goal separation.** (a)–(d) Closed-loop success (%) at temporal offsets 10, 25, 40, 50, and 75 on Cube, TwoRooms, Reacher, and PushT, respectively. (e) Cube results by shared true-displacement quartiles Q1–Q4, with 75 paired instances per quartile. Larger temporal offsets need not imply monotonically increasing difficulty.



Figure 6: **Qualitative rollouts across the three collapse modes.** Rows compare LeWM and PhyLatent under matched Cube configurations for Physical Invariance, Physical Distinguishability, and Counterfactual Dynamics Collapse.

Figure 5 shows higher average success on Cube, TwoRooms, and PushT, with Reacher the main exception. On TwoRooms, the gap over LeWM grows from 2.00 points at temporal offset 10 to 30.67 points at offset 75.

Cube provides a physical counterpart: Q1–Q4 are shared quartiles of true start–goal cube displacement. Both models solve near-zero-motion cases, but success in the largest quartile rises from 26.67% with LeWM to 60.00% with PhyLatent. With temporal offset fixed and bin boundaries shared, the displacement test compares the models on the same physical separations. Its largest-quartile advantage shows that the benefit extends beyond near-stationary goals, although displacement alone does not isolate every source of planning difficulty.

5.7 Qualitative Rollout Analysis

Figure 6 compares matched Cube rollouts: PhyLatent preserves a more useful representation and completes interactions where LeWM fails.

6 Conclusion

PhyLatent exposes a limitation of global anti-collapse objectives in JEPA world models: a globally non-degenerate representation can still lose the local physical and action-conditioned relations required for planning. By targeting Physical Invariance, Physical Distinguishability, and Counterfactual Dynamics Collapse, PhyLatent reduces the three collapse rates on Cube from 19.58%, 45.51%, and 14.47% to 10.99%, 32.82%, and 7.65%, while improving four-task average planning success from 79.92% to 86.54%. The decision-level cases further connect these representation failures to planning behavior: in the Distinguishability case, an alternative candidate reaches 1.49 cm from the goal instead of 10.00 cm when the latent ordering fails. Together, these results show that preserving dynamics-relevant structure can improve both latent representations and closed-loop planning.

Limitation and Future Work. PhyLatent uses selected physical-state variables during training, and its auxiliary objectives are not fully orthogonal in the shared representation space. Future work should reduce this supervision requirement and better understand and coordinate the interactions among the three training pathways.

AI USE STATEMENT

In this work, we used generative AI tools for retrieval and discovery of relevant literature, to draft and polish portions of the manuscript, to format tables, and to discuss the interpretation and presentation of experimental results. All AI-assisted material was reviewed and revised by the authors, who take full responsibility for the final content of the paper.

REPRODUCIBILITY STATEMENT

The supplement provides "Collapse Diagnostic Definitions", "Decision-Level Failure Cases", and "Implementation and Evaluation Details", covering diagnostic sampling, case-study replays, training, hyperparameters, and evaluation. Further results appear in "Additional Diagnostic Analyses", "Additional Robustness and Planning Results", and "Full Ablation Results". Together with the main paper, these sections provide the methodological and evaluation details needed to reproduce PhyLatent.

REFERENCES

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.

Mido Assran, Adrien Bardes, David Fan, et al. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Randall Balestriero and Yann LeCun. LeJEPA: Provable and scalable self-supervised learning without the heuristics. *arXiv preprint arXiv:2511.08544*, 2025.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEiT: BERT pre-training of image transformers. In *International Conference on Learning Representations*, 2022.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*, 2024.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9912–9924, 2020.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660, 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pp. 1597–1607. PMLR, 2020.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin C. M. Burchfiel, and Shuran Song. Diffusion Policy: Visuomotor policy learning via action diffusion. In *Proceedings of Robotics: Science and Systems*, Daegu, Republic of Korea, July 2023. doi: 10.15607/RSS.2023.XIX.026.

Pieter-Tjerk de Boer, Dirk P. Kroese, Shie Mannor, and Reuven Y. Rubinstein. A tutorial on the cross-entropy method. *Annals of Operations Research*, 134:19–67, 2005. doi: 10.1007/s10479-005-5724-z.

Carles Gelada, Saurabh Kumar, Jacob Buckman, Ofir Nachum, and Marc G. Bellemare. DeepMDP: Learning continuous latent space models for representation learning. In *International Conference on Machine Learning*, 2019.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, 2020.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 2555–2565, 2019.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020.

Nicklas Hansen, Hao Su, and Xiaolong Wang. TD-MPC2: Scalable, robust world models for continuous control. In *International Conference on Learning Representations*, 2024.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738, 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

Petr Ivashkov, Randall Balestriero, and Bernhard Schölkopf. Sensorimotor world models: Perception for action via inverse dynamics. *arXiv preprint arXiv:2606.20104*, 2026.

Rico Jonschkowski and Oliver Brock. Learning state representations with robotic priors. *Autonomous Robots*, 39(3):407–428, 2015. doi: 10.1007/s10514-015-9459-7.

Yann LeCun. A path towards autonomous machine intelligence. Position paper, 2022. URL `https://openreview.net/forum?id=BZ5a1r-kVsf`.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=Bkg6RiCqY7`.

Lucas Maes, Quentin Le Lidec, Luiz Facury, Nassim Massaudi, Ayush Chaurasia, Francesco Capuano, Richard Gao, Taj Gillin, Dan Haramati, Damien Scieur, Yann LeCun, and Randall Balestriero. stable-worldmodel: A platform for reproducible world modeling research and evaluation. *arXiv preprint arXiv:2605.21800*, 2026a.

Lucas Maes, Quentin Le Lidec, Damien Scieur, Yann LeCun, and Randall Balestriero. LeWorldModel: Stable end-to-end joint-embedding predictive architecture from pixels. *arXiv preprint arXiv:2603.19312*, 2026b.

D. Q. Mayne, J. B. Rawlings, C. V. Rao, and P. O. M. Scokaert. Constrained model predictive control: Stability and optimality. *Automatica*, 36(6):789–814, 2000. doi: 10.1016/S0005-1098(99)00214-9.

Seohong Park, Kevin Frans, Benjamin Eysenbach, and Sergey Levine. OGBench: Benchmarking offline goal-conditioned RL. In *International Conference on Learning Representations*, 2025.

Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *International Conference on Machine Learning*, 2017.

Pierre Sermanet, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, and Sergey Levine. Time-contrastive networks: Self-supervised learning from video. In *IEEE International Conference on Robotics and Automation*, 2018.

Uladzislau Sobal, Wancong Zhang, Kyunghyun Cho, Randall Balestriero, Tim G. J. Rudner, and Yann LeCun. Learning from reward-free offline data: A case for planning with latent dynamics models. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Lucas Thil, Jesse Read, Rim Kaddah, and Guillaume Doquet. Subspace-decomposed JEPAs: Disentangling progression and content in latent world models. *arXiv preprint arXiv:2605.31111*, 2026.

Saran Tunyasuvunakool, Alistair Muldal, Yotam Doron, Siqi Liu, Steven Bohez, Josh Merel, Tom Erez, Timothy Lillicrap, Nicolas Heess, and Yuval Tassa. dm_control: Software and tasks for continuous control. *Software Impacts*, 6:100022, 2020. doi: 10.1016/j.simpa.2020.100022.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Tongzhou Wang, Simon S. Du, Antonio Torralba, Phillip Isola, Amy Zhang, and Yuandong Tian. Denoised MDPs: Learning world models better than the world itself. In *International Conference on Machine Learning*, 2022.

Zhoushun Yu, Xiaoyu Hu, and Xiangyu Xu. QQWorld: Quantile-quantile matching for world model regularization. *arXiv preprint arXiv:2607.28415*, 2026.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow Twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, 2021.

Amy Zhang, Rowan McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. Learning invariant representations for reinforcement learning without reconstruction. In *International Conference on Learning Representations*, 2021.

Kai Zhao, Dongliang Nie, Yuchen Lin, Zhehan Luo, Yixiao Gu, Deng-Ping Fan, and Dan Zeng. Sub-JEPA: Subspace gaussian regularization for stable end-to-end world models. *arXiv preprint arXiv:2605.09241*, 2026.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. DINO-WM: World models on pre-trained visual features enable zero-shot planning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 79115–79135. PMLR, 2025.