MOTION-GUIDED CONTRASTIVE REGULARIZATION FOR LATENT WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

JEPA-style world models learn action-conditioned latent dynamics without reconstructing observations, but distributional regularization alone does not ensure that their representations preserve information needed for control. We identify *static collapse*, a failure mode in which the encoder preserves episode-specific static content while suppressing manipulation-relevant state changes. Because static content is temporally predictable yet diverse across episodes, this failure can persist despite regularizers such as SIGReg. A controlled study on OGBench-Cube shows that introducing randomized backgrounds substantially degrades planning performance while leaving the underlying manipulation task unchanged. To address this problem, we propose *motion-guided contrastive regularization*. Our approach compares predicted latent transitions with transitions to candidate observations from the same episode, reducing the usefulness of static background identity for distinguishing action outcomes. Optical-flow similarities provide soft contrastive targets, allowing candidates with similar motion to receive similar supervision. Combined with latent prediction and SIGReg, our objective encourages representations to preserve within-episode state distinctions and similarities between related transitions. Our method improves planning success over LeWorldModel (LeWM) from 52% to 78% on OGBench-Cube with randomized backgrounds and from 14% to 54% on LIBERO-Object, while maintaining comparable performance on four other visual-control benchmarks.

1 INTRODUCTION

World models have attracted considerable attention due to their wide applications in robotics and autonomous driving. By learning environment dynamics to predict the consequences of actions, they enable agents to plan and learn behaviors through imagined experience (Dawid & LeCun, 2024; Hafner et al., 2025; Micheli et al., 2022). However, learning representations that capture critical state information without being overwhelmed by high-dimensional visual details is challenging.

Joint Embedding Predictive Architectures (JEPAs) offer a promising solution by predicting target representations from context, completely bypassing pixel-level reconstruction (Assran et al., 2023). It allows an encoder to map observations into latent states while a predictor estimates future embeddings based on observation history and actions (Maes et al., 2026). However, abandoning pixel reconstruction introduces a critical vulnerability to representational collapse, where the unconstrained objective trivially maps all observations to a constant embedding (Maes et al., 2026).

Considerable effort has been devoted to learning non-collapsing representations in predictive world models. Early approaches to prevent representation collapse primarily focused on constraining encoder updates, either through moving-average target encoders (Assran et al., 2023; Bardes et al., 2024) or by freezing large pretrained visual encoders (Zhou et al., 2024; Assran et al., 2025). These methods stabilize prediction targets and achieve strong empirical results. However, they do not explicitly force representations to distinguish between different action outcomes, and frozen-encoder approaches additionally incur significant computational overhead from their large pretrained backbones. Alternatively, other approaches prevent collapse by supervising action information (Ivashkov et al., 2026; Zhang et al., 2026; Sobal et al., 2026). By employing inverse dynamics to recover actions from latent transitions or differences, these methods successfully establish meaningful state representations. Nevertheless, their supervisory signal is inherently focused on recovering the ex-

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.

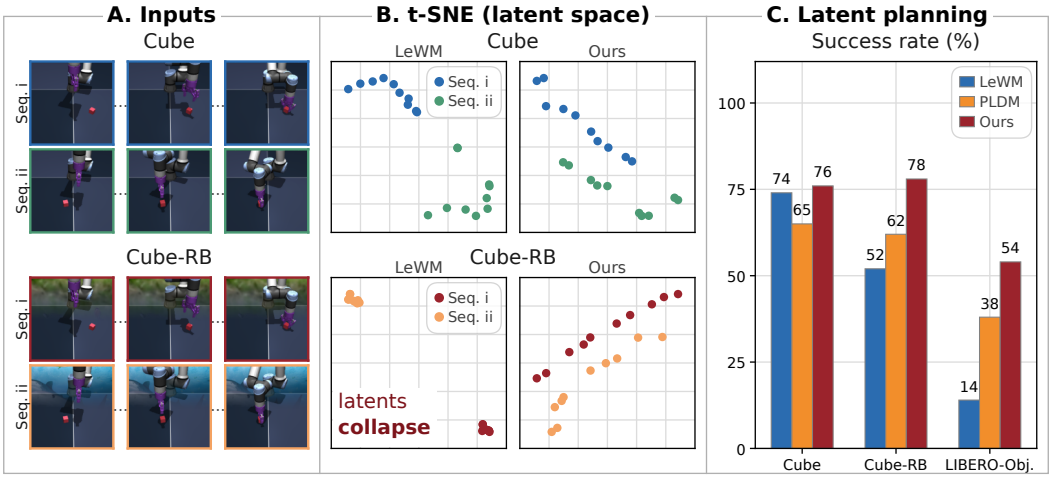


Figure 1: **Static collapse and our solution.** **(A)** We train JEPA world models end-to-end on *OGBench-Cube* (Cube) and a variant with randomized backgrounds (Cube-RB). In Cube-RB, one background image is sampled per episode and held fixed throughout it. Only distant pixels (depth $\geq$ 1.0 m) are replaced, so the robot, cube, and dynamics are unchanged and only task-irrelevant background variation is added. **(B)** t-SNE visualizations of the encoder embeddings, where each point is one input frame and each color one episode. On Cube-RB, LeWorldModel (LeWM; Maes et al., 2026) maps distinct states within each episode to tightly clustered latents, whereas our motion-guided contrastive regularization preserves within-episode state variation. **(C)** Planning success of LeWM, Planning with Latent Dynamics Models (PLDM; Sobal et al., 2026), and our method. LeWM and our method perform similarly on Cube, but LeWM's success drops on Cube-RB and is low on LIBERO-Object. Our method achieves the highest success of the three in all settings.

ecuted action itself, rather than explicitly distinguishing a predicted future outcome from competing alternatives. More recent efforts have turned to regularizing the embedding distribution itself (Bardes et al., 2021; Balestriero & LeCun, 2025; Maes et al., 2026). For instance, by employing Sketched Isotropic Gaussian Regularization (SIGReg) to encourage isotropic Gaussian latent spaces, LeWorldModel (LeWM; Maes et al., 2026) trains a lightweight ViT that achieves results comparable to DINO-WM (Zhou et al., 2024) while planning more than $50\times$ faster (Table 1).

However, SIGReg constrains embedding statistics without specifying which information to preserve, allowing static backgrounds to dominate while control-relevant state changes are suppressed. We term this failure **static collapse**. Figure 1 illustrates it on OGBench-Cube-RB, constructed by replacing distant background pixels in OGBench-Cube (Park et al., 2025) while preserving the robot and cube. LeWM's t-SNE embeddings cluster tightly within each episode despite changes in manipulation state, and its planning success drops substantially relative to the original dataset. These results show that statistical regularization alone does not ensure preservation of the state distinctions needed for control.

In this paper, we introduce motion-guided contrastive regularization for JEPA-style world models. Our approach is motivated by two key observations. First, distributional regularization, e.g., SIGReg, alone does not ensure that representations preserve state differences within the same environment. We therefore introduce a contrastive objective that compares the predicted action outcome with the observed next state and alternative states from the same episode. Because these observations share the same static scene, background identity alone cannot resolve the comparison, encouraging the model to capture changes in the robot and manipulated objects. Second, alternative states are not equally different from the observed outcome. Treating them all as negatives can push apart states that reflect similar manipulation outcomes, while measuring their similarity through pixel differences can be sensitive to irrelevant appearance variations. We therefore use optical flow to guide the contrastive objective, assigning greater similarity to alternatives whose motion resembles the observed transition. Together, these two ideas encourage the model to distinguish meaningful state changes while preserving similarities between related outcomes. Combined with latent prediction

and SIGReg, our approach targets static collapse without relying on pixel reconstruction losses or auxiliary action-decoding objectives.

We evaluate our framework on the OGBench (Park et al., 2025) and LIBERO (Liu et al., 2023) benchmarks under varying levels of visual complexity. Through rigorous representation analysis and downstream planning evaluations, our experiments reveal that while static collapse fundamentally impairs standard SIGReg-regularized models, our proposed contrastive objective effectively circumvents this failure. Ultimately, our approach yields substantial improvements in planning performance amid increased background diversity and scene clutter. As a result, our method substantially boosts planning robustness under challenging visual conditions, increasing success rates from 52% to 78% on OGBench-Cube with randomized backgrounds and from 14% to 54% on LIBERO-Object compared to LeWM.

This paper contributes to JEPA world models as follows. First, we propose a motion-guided contrastive objective that explicitly preserves action-conditioned state changes in the learned latent space. Second, we introduce an optical-flow-based similarity measure that focuses the contrastive objective on meaningful scene dynamics rather than low-level visual differences. Third, we show across multiple benchmark datasets that our approach mitigates static-background-induced representation collapse and substantially improves planning performance under visually complex conditions.

2 Related Work

Latent world models. Latent world models learn compact environment dynamics to support planning and policy learning. World Models (Ha & Schmidhuber, 2018), PlaNet (Hafner et al., 2019b), Dreamer and DreamerV3 (Hafner et al., 2019a; 2023), and IRIS (Micheli et al., 2022) learn to model dynamics and control in latent space, while supervising the process with observation reconstruction. MuZero (Schrittwieser et al., 2020), EfficientZero and EfficientZero V2 (Ye et al., 2021; Wang et al., 2024a), TD-MPC (Hansen et al., 2022), and TD-MPC2 (Hansen et al., 2024) learn reconstruction-free, task-oriented models whose representations are shaped by reward and value supervision. DreamerPro (Deng et al., 2022), MuDreamer (Burchi & Timofte, 2024), and Dreamer-CDP (Hauri & Zenke, 2026) also combine reconstruction-free representation learning with reward-driven control.

JEPA methods, including I-JEPA (Assran et al., 2023) and V-JEPA (Bardes et al., 2024), learn visual representations by predicting target embeddings without reconstructing pixels. For control, DINO-WM (Zhou et al., 2024), V-JEPA 2-AC (Assran et al., 2025), and JEPA-WM (Terver et al., 2025) learn action-conditioned dynamics over pretrained visual representations, enabling image-goal planning without task-reward supervision. PLDM (Sobal et al., 2026) and LeWM (Maes et al., 2026) jointly learn the encoder and dynamics predictor from offline trajectories without observation reconstruction or task rewards. We focus on this end-to-end setting, where representations and dynamics can be learned together from offline interactions for image-goal planning. However, under joint optimization, latent prediction alone admits collapsed solutions, making the preservation of informative representations a central challenge.

Representation collapse and anti-collapse regularization. Anti-collapse regularization promotes representation diversity during joint encoder–predictor training. VICReg (Bardes et al., 2021) constrains feature variance and covariance, while SIGReg (Balestriero & LeCun, 2025), adopted by LeWM (Maes et al., 2026), encourages an isotropic Gaussian embedding distribution. However, statistical diversity alone does not ensure that representations preserve control-relevant information. Prior work has shown that JEPA models can favor static background features while failing to capture moving-object states (Sobal et al., 2022). Other approaches address this issue through task- or action-based supervision. Bisimulation-based encoders (Toso et al., 2026) encourage representations that preserve behaviorally relevant distinctions while remaining invariant to distractors. PLDM (Sobal et al., 2026) uses inverse-dynamics supervision to retain action information. Concurrent with our work, Sensorimotor World Models (Ivashkov et al., 2026) and Delta-JEPA (Zhang et al., 2026) also explore inverse-dynamics supervision for learning action-sensitive representations. Our method instead introduces a within-episode contrastive objective guided by motion similarity. It encourages predicted latent transitions to distinguish alternative outcomes while preserving similarities between related transitions, without requiring an auxiliary action decoder.

Contrastive and motion-guided representation learning. Contrastive objectives enable latent prediction without requiring observation reconstruction (van den Oord et al., 2018; Kipf et al., 2019). Building on this idea, recent approaches apply contrastive supervision to action-conditioned prediction (Zheng et al., 2023; Burchi & Timofte, 2025) and the learning of physical dynamics (Martini et al., 2026). The choice of negative samples also influences the learned representation: Gulati & Nemenman (2026) show that within-trajectory negatives suppress shortcuts based on trajectory-specific static noise and improve the recovery of underlying dynamical variables. In video representation learning, synthetic motion augmentations and complementary visual-motion objectives have been used to encourage motion-sensitive features (Thoker et al., 2023; Han et al., 2020; Xiao et al., 2022). Most closely related to our approach, Mahendran et al. (2018) supervise representations by aligning pixel-embedding similarity distributions with targets derived from optical-flow similarity. Our method combines within-episode contrastive sampling with motion-derived soft targets to explicitly structure action-conditioned latent transitions.

3 Preliminaries and static collapse

LeWorldModel. We build on LeWM (Maes et al., 2026), a JEPA-style world model trained end-to-end from visual observations. Given an observation $\mathbf{x}_t$ and action $\mathbf{a}_t$, its ViT encoder (Dosovitskiy et al., 2020) and transformer predictor compute

$$\mathbf{z}_t = \mathrm{enc}_\theta(\mathbf{x}_t), \qquad \hat{\mathbf{z}}_{t+1} = \mathrm{pred}_\phi(\mathbf{z}_t, \mathbf{a}_t). \tag{1}$$

For clarity, we write the predictor in single-step form; in practice, it conditions on the embeddings of the three most recent observations and their associated actions (Appendix A).

The prediction is supervised by the encoded next observation $\mathbf{z}_{t+1} = \mathrm{enc}_\theta(\mathbf{x}_{t+1})$:

$$\mathcal{L}_{\mathrm{pred}} = \mathbb{E}\left[\|\hat{\mathbf{z}}_{t+1} - \mathbf{z}_{t+1}\|_2^2\right], \tag{2}$$

where the expectation is over training transitions. To discourage constant collapse, LeWM additionally uses SIGReg (Balestriero & LeCun, 2025), which encourages minibatch embeddings $\mathbf{Z}$ to follow an isotropic Gaussian distribution. The encoder and predictor are jointly optimized with

$$\mathcal{L}_{\mathrm{LeWM}} = \mathcal{L}_{\mathrm{pred}} + \lambda\,\mathrm{SIGReg}(\mathbf{Z}), \tag{3}$$

where λ controls the regularization strength. Although this objective encourages prediction accuracy and batch-level diversity, it does not ensure that the representations preserve action-relevant state changes.

Static collapse. Although SIGReg is an effective regularizer and LeWM achieves promising results, it does not specify which information should be preserved in the latent representation. We identify a failure mode arising from this limitation: in environments with visually complex backgrounds, the encoder can satisfy the SIGReg objective by primarily encoding static scene appearance. Because background features already exhibit substantial diversity across observations, they provide an easy way to satisfy the Gaussianity constraint enforced by SIGReg, reducing the pressure to encode manipulation-relevant dynamics. Specifically, when background features alone are sufficient to achieve a low SIGReg loss, the learned representation tends to become dominated by those features:

$$\mathrm{SIGReg}(\mathbf{Z}_B) \approx 0 \implies \mathbf{z}_t \approx \mathbf{Z}_B + \epsilon_t, \qquad \|\epsilon_t\| \ll \|\mathbf{Z}_B\|, \tag{4}$$

where $\mathbf{Z}_B$ denotes background features and ϵ_t captures the remaining dynamic information. Consequently, changes induced by robot-object interactions contribute only weakly to the representation, leading to poor discrimination between manipulation states and degraded downstream control performance.

We use the single-cube manipulation environment from OGBench (Park et al., 2025), denoted *Cube*, and construct a paired variant, *Cube-RB*. For each episode in *Cube-RB*, we sample a background image from ImageNet (Deng et al., 2009) and keep it fixed throughout the episode. Only distant pixels ($depth \geq 1.0$ m) are replaced, leaving the robot arm, cube, actions, and underlying dynamics unchanged. As a result, *Cube* and *Cube-RB* differ only in the presence of episode-specific background variation. We train LeWM on both datasets using identical hyperparameters, thereby isolating the effect of complex static visual content. Figure 1(B) visualizes the learned latent representations using

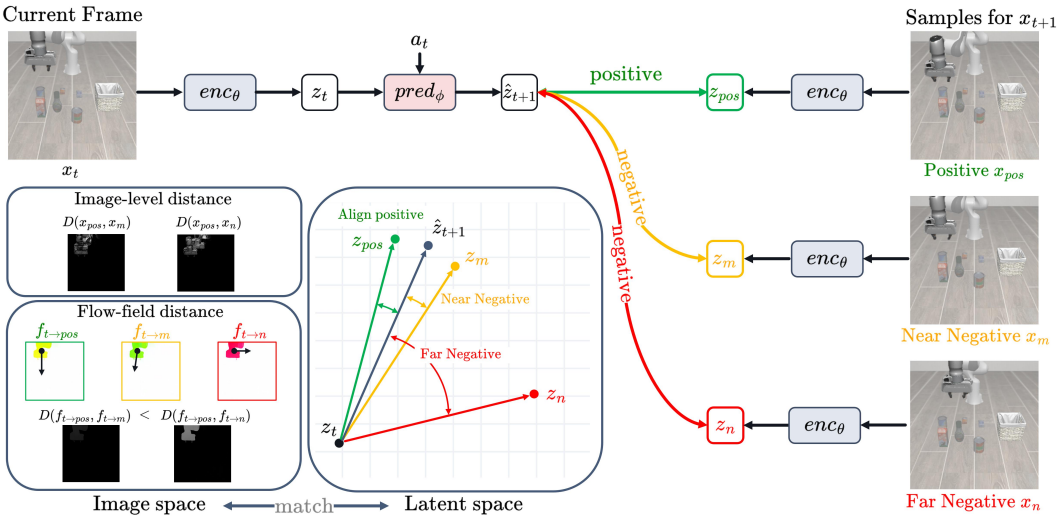


Figure 2: **Motion-guided contrastive learning.** The world model encodes the current observation $\mathbf{x}_t$ and predicts $\hat{\mathbf{z}}_{t+1}$. The true next observation $\mathbf{x}_{\text{pos}} = \mathbf{x}_{t+1}$ provides the positive sample, while other frames from the same episode provide negative candidates $\mathbf{x}_m$ and $\mathbf{x}_n$. All observations share the same encoder. **Image-space:** Optical flow is computed from $\mathbf{x}_t$ to each candidate. Comparing these fields with the positive flow assigns higher target similarity to candidates with similar motion ($\mathbf{x}_m$) than to those with different motion ($\mathbf{x}_n$). **Latent-space:** Angles illustrate cosine similarities between the predicted transition $\hat{\mathbf{z}}_{t+1} - \mathbf{z}_t$ and candidate transitions $\mathbf{z}_j - \mathbf{z}_t$. Our motion-guided contrastive loss matches the resulting latent similarity distribution to the motion-derived target distribution through cross-entropy. The prediction loss aligns $\hat{\mathbf{z}}_{t+1}$ with $\mathbf{z}_{\text{pos}}$. The grippers in the sample frames are highlighted for illustration purposes only.

t-SNE (van der Maaten & Hinton, 2008). On *Cube*, the embeddings distinguish different manipulation states within an episode. In contrast, on *Cube-RB*, embeddings cluster primarily according to background identity, causing visually distinct manipulation states to collapse into compact episode-specific groups despite substantial changes in the robot and cube configurations. This representation collapse is accompanied by a significant drop in planning performance. As shown in Figure 1(C), LeWM's success rate decreases from 74% on *Cube* to 52% on *Cube-RB*.

This failure mode motivates our motion-guided contrastive objective. By contrasting candidate states within the same scene, the objective prevents the model from relying on static background cues and instead encourages representations that capture task-relevant dynamics.

4 Motion-Guided Contrastive Regularization

We introduce our motion-guided contrastive objective in this section. We train an action-conditioned world model to distinguish the observed next state from alternative states within the same episode, using optical-flow similarity to assign soft contrastive targets. Comparing states within an episode reduces the usefulness of static background identity, while motion-based targets account for alternatives that represent similar transitions.

Within-episode candidates. Given an observed transition $(\mathbf{x}_t, \mathbf{a}_t, \mathbf{x}_{t+1})$, we use the next observation $\mathbf{x}_{t+1}$ as the positive sample and sample N additional observations from the same episode. Let $\mathcal{N}_t$ contain their time indices, excluding t and $t+1$, and let $\mathcal{C}_t = \{t+1\} \cup \mathcal{N}_t$ index all candidates. These observations share the same scene but may differ in the configuration of the robot and manipulated objects. Static background identity alone therefore cannot distinguish the positive from the other candidates.

Latent transition similarities. A shared encoder maps each observation to a latent state, $\mathbf{z}_i = \text{enc}_\theta(\mathbf{x}_i)$, and an action-conditioned predictor estimates the next latent state, $\hat{\mathbf{z}}_{t+1} = \text{pred}_\phi(\mathbf{z}_t, \mathbf{a}_t)$. We express the prediction and each candidate as changes from the current latent state:

$$\widehat{\Delta \mathbf{z}}_t = \hat{\mathbf{z}}_{t+1} - \mathbf{z}_t, \qquad \Delta \mathbf{z}_{t,j} = \mathbf{z}_j - \mathbf{z}_t, \quad j \in \mathcal{C}_t. \tag{5}$$

Inspired by Mahendran et al. (2018), who match pixel embedding-similarity distributions to optical-flow-derived targets, we adopt a distribution-matching formulation for latent transitions. Using the predicted latent change $\widehat{\Delta z}_t$ as the contrastive anchor, we apply a softmax with temperature τ_z to its cosine similarities with the candidate changes $\Delta z_{t,j}$:

$$p_{t,j} = \frac{\exp\Big(\mathrm{sim}(\widehat{\Delta \mathbf{z}}_t, \Delta \mathbf{z}_{t,j})/\tau_z\Big)}{\sum_{k \in \mathcal{C}_t} \exp\Big(\mathrm{sim}(\widehat{\Delta \mathbf{z}}_t, \Delta \mathbf{z}_{t,k})/\tau_z\Big)}, \tag{6}$$

where $\mathrm{sim}(\cdot, \cdot)$ denotes cosine similarity and $\tau_z > 0$ is a temperature. This distribution describes how closely each candidate's latent change matches the predicted change.

Motion-guided supervision. Other observations in the episode may depict transitions similar to the observed next transition. As shown in Figure 2, image-level distance $D(\mathbf{x}_{pos}, \mathbf{x}_m)$ can be sensitive to irrelevant appearance variations. We therefore assign soft targets based on their motion similarity, rather than treating every alternative as equally dissimilar.

For each candidate, we compute an optical-flow field from the current observation:

$$\mathbf{f}_{t \to j} = \mathrm{Flow}(\mathbf{x}_t, \mathbf{x}_j), \qquad j \in \mathcal{C}_t. \tag{7}$$

We compute optical flow with a pretrained SEA-RAFT model (Wang et al., 2024b); settings are given in Appendix A.3.

To prevent large static regions from diluting motion differences, we weight each location u by the larger flow magnitude in the positive and candidate transitions:

$$w_{t,j}(u) = \max(\|\mathbf{f}_{t \to t+1}(u)\|_2, \|\mathbf{f}_{t \to j}(u)\|_2)\,. \tag{8}$$

We then measure motion dissimilarity $d_{t,j}$ for all pixels Ω using a weighted endpoint distance:

$$d_{t,j} = \frac{\sum_{u \in \Omega} w_{t,j}(u) \|\mathbf{f}_{t \to t+1}(u) - \mathbf{f}_{t \to j}(u)\|_2}{\sum_{u \in \Omega} w_{t,j}(u) + \epsilon}, \tag{9}$$

where $\epsilon > 0$ ensures numerical stability. Locations with little motion in both transitions receive little weight, while motion in either transition increases their contribution. The vector distance captures differences in both displacement magnitude and direction.

These distances define the target distribution:

$$q_{t,j} = \frac{\exp(-d_{t,j}/\tau_f)}{\sum_{k \in \mathcal{C}_t} \exp(-d_{t,k}/\tau_f)}, \tag{10}$$

where $\tau_f > 0$ controls target sharpness. The positive has zero distance and therefore receives maximal target probability, potentially tied with candidates that have identical flow. Other candidates receive probability according to their motion similarity to the positive.

The motion-guided contrastive loss matches the latent similarity distribution to these targets:

$$\mathcal{L}_{\mathrm{MGC}} = -\mathbb{E}_{t, \mathcal{N}_t} \left[\sum_{j \in \mathcal{C}_t} q_{t,j} \log p_{t,j} \right], \tag{11}$$

where the expectation is over observed transitions and sampled candidate sets. A one-hot target on $t+1$ would recover a single-positive InfoNCE loss with the same latent transition scores. Our soft targets instead encourage the predicted latent change to remain compatible with alternatives whose observed motion resembles the positive transition.

Training objective. We retain LeWM's latent prediction loss and SIGReg regularization and add the motion-guided contrastive loss:

$$\mathcal{L} = \mathcal{L}_{\mathrm{pred}} + \lambda\, \mathrm{SIGReg}(\mathbf{Z}) + \beta\, \mathcal{L}_{\mathrm{MGC}}, \tag{12}$$

where $\mathbf{Z}$ denotes the embeddings used for regularization in a training batch, and $\lambda, \beta \geq 0$ control the two regularization terms. The prediction loss learns action-conditioned dynamics, SIGReg regularizes the embedding distribution, and our contrastive loss encourages latent transitions to distinguish candidates according to their observed motion.

Table 1: Planning success rate (%) on six settings. † uses a frozen pretrained visual encoder. Peak GPU memory and planning time are measured on Cube dataset with an A5000 GPU and averaged over its 50 evaluation instances; planning time is the wall-clock time to plan one full trajectory (five replanning rounds). CEM uses 128 candidates and 10 iterations per round, with a planning horizon of up to 5 control steps.

Method	Push-T	2-Rooms	Reacher	Cube	Cube-RB	LIBERO-Object	Peak GPU mem. (MB)	Planning time (ms)
DINO-WM† (Zhou et al., 2024)	74	**100**	78	78	68	**68**	20,714	11,223
PLDM (Sobal et al., 2026)	78	97	64	65	62	38	338	28
SMWM (Ivashkov et al., 2026)	83	99	66	**84**	**78**	40	374	101
LeWM (Maes et al., 2026)	**90**	86	**80**	74	52	14	372	112
Ours	**90**	88	78	76	**78**	54	372	112

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Benchmark datasets. We evaluate on six visual-control settings. The controlled pair introduced in Section 3, *Cube* (Park et al., 2025) and *Cube-RB* (Cube with randomized backgrounds), isolates the effect of episode-specific background variation. We additionally evaluate on LIBERO-Object (Liu et al., 2023) to assess whether the benefits extend to multi-object scenes without synthetic background replacement. For LIBERO-Object, we expand the training data with noise-perturbed expert trajectories. Appendix A.1 details this data preparation and evaluation protocol. Finally, we evaluate on Push-T (Zhou et al., 2024), 2-Rooms (Sobal et al., 2026), and Reacher (Tassa et al., 2018), following prior world-model evaluations (Zhou et al., 2024; Maes et al., 2026), to assess whether our objective maintains performance in visually simpler environments.

Baseline models. We compare against four learned world-model baselines. **LeWM** (Maes et al., 2026) is our backbone, trained with its original prediction and SIGReg objectives without the proposed contrastive loss. **Sensorimotor World Models** (SMWM) (Ivashkov et al., 2026) encourages action-sensitive representations through an auxiliary action-decoding objective. **PLDM** (Sobal et al., 2026) combines latent prediction with VICReg-inspired regularization and an auxiliary inverse-dynamics objective. **DINO-WM** (Zhou et al., 2024) learns dynamics over frozen DINOv2 features (Oquab et al., 2023), providing a reference based on large-scale visual pretraining; unlike the other methods, it is not trained end-to-end on task data, and we mark it with † in Table 1.

Training details. All methods are trained on the same task data and observation augmentations for 100 epochs, using a batch size of 128 and a learning rate of 5×10^{-5}. LeWM and our method share identical architectures and optimization settings, differing only in the addition of our motion-guided contrastive objective. Following LeWM (Maes et al., 2026), we use $M = 1024$ random projections for SIGReg and set its regularization weight to $\lambda = 0.09$. We compute optical flow using SEA-RAFT (Wang et al., 2024b). For each training transition, we sample $N = 6$ additional candidate observations from the same episode. We set the flow temperature to $\tau_f = 20$, the latent temperature to $\tau_z = 0.1$, and the contrastive loss weight to $\beta = 0.3$. We analyze sensitivity to the temperatures τ_f and τ_z, the number of additional candidates N, the SIGReg weight λ, and the contrastive loss weight β in Appendix B.

Evaluation via latent planning. We evaluate the learned world models through image-goal planning using a shared, sampling-based model-predictive control (MPC) procedure. Following LeWM (Maes et al., 2026), we encode the current and goal observations and predict latent rollouts for candidate action sequences. The Cross-Entropy Method (CEM) (Rubinstein, 1999) optimizes these sequences by minimizing the squared Euclidean distance between the predicted terminal embedding and the goal embedding. Holding the controller fixed across methods evaluates how well their learned representations and dynamics support goal-directed control. Our comparison focuses on offline, reward-free world models used within this common planning framework. To compare computational cost, we also report the peak GPU memory (*Peak GPU mem.*) and the wall-clock planning time (*Planning time*) required to plan one trajectory. Appendix A.2 provides the planning configuration and evaluation protocol.

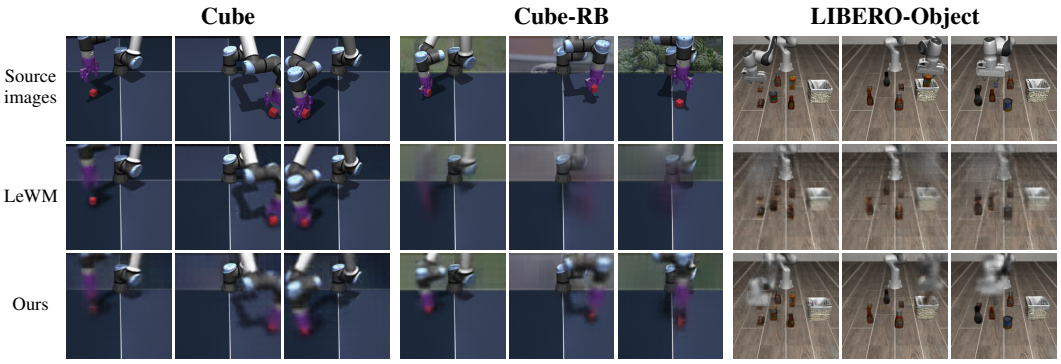


Figure 3: Latent-space decoding on Cube, Cube-RB, and LIBERO-Object. Compared to LeWM, reconstructions from our frozen latents more clearly preserve the robot and manipulated objects on the challenging Cube-RB and LIBERO-Object datasets. Check our video demo for more comparisons.

5.2 Main Results

Table 1 shows that motion-guided contrastive regularization substantially improves planning success over LeWM on Cube-RB and LIBERO-Object. On Cube-RB, success increases from 52% to 78%, a gain of 26 percentage points under background randomization. On LIBERO-Object, success under our image-goal planning protocol increases from 14% to 54%, extending the benefits to multi-object scenes, where the task-irrelevant static objects can cause similar static collapse behavior as in Cube-RB. On Cube, Push-T, 2-Rooms, and Reacher, our method's success rates remain on par. In particular, our method achieves 76% on Cube and 78% on Cube-RB, maintaining similar observed performance despite the introduction of episode-specific backgrounds, indicating the effectiveness of our method in preserving task-relevant information. Among the evaluated end-to-end world models, our method achieves the highest observed LIBERO-Object success rate. DINO-WM achieves higher success on this setting using its frozen pretrained visual encoder. On Cube-RB, our method matches SMWM at 78%. These comparisons demonstrate substantial improvements over the LeWM backbone.

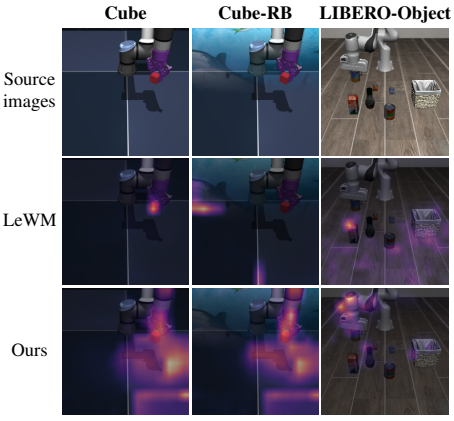


Figure 4: Encoder attention over image patches (brighter = more attention). On Cube-RB and LIBERO-Object, LeWM attends to background and static objects; ours focuses on the gripper and manipulated object.

Decodability of manipulation state. Following Maes et al. (2026), we train a lightweight transformer decoder to reconstruct observations from a single latent embedding. The world-model parameters remain frozen during decoder training, providing a qualitative probe of the visual information recoverable from each representation. Figure 3 compares reconstructions on Cube, Cube-RB, and LIBERO-Object (ground truth, LeWM, ours). Reconstructions from LeWM latents retain the coarse scene layout but blur the manipulated object and end effector. Reconstructions from our latents more clearly depict object locations and gripper configurations. These results suggest that motion-guided contrastive supervision makes manipulation-relevant visual information more accessible to the decoder.

Attention maps. We visualize the encoder's spatial focus when forming the global latent embedding by aggregating patch attention weights across all layers (Figure 4). Rows show the input image and the attention maps for LeWM and our method. On Cube-RB and LIBERO-Object, LeWM assigns substantial attention to background regions and static objects, whereas our method focuses more strongly on the manipulated object and gripper. These patterns complement the reconstruction results and suggest greater emphasis on manipulation-relevant regions.

Physical property probing. Qualitative reconstructions suggest that our latents retain manipulation-relevant visual information. To quantify physical-state recoverability, we train two-layer MLP probes on frozen encoder embeddings following Maes et al. (2026). We predict end-effector and cube positions on Cube and Cube-RB, and end-effector position and a binary object-holding indicator on LIBERO-Object. We report Euclidean position error in centimeters and the area under the receiver operating characteristic curve (AUROC) for object-holding classification. As shown in Table 2, our method improves physical-state recovery across all evaluated quantities.

Table 2: MLP probing of physical state from frozen latents. Position errors are Euclidean distances in centimeters. Holding object (AUROC) is a binary gripper-state classification on LIBERO-Object.

Dataset	Metric	LeWM	Ours
Cube	Robot pos. error (cm) ↓	2.74	**0.61**
	Cube pos. error (cm) ↓	1.27	**0.62**
Cube-RB	Robot pos. error (cm) ↓	13.01	**3.56**
	Cube pos. error (cm) ↓	10.49	**5.10**
LIBERO-Object	Robot pos. error (cm) ↓	10.67	**6.00**
	Holding object (AUROC) ↑	0.78	**0.88**

5.3 Ablation Studies

We ablate contrastive supervision and SIGReg on Cube-RB and LIBERO-Object, retaining the prediction loss in all variants. As shown in Table 3, adding InfoNCE to LeWM increases planning success from 52% to 56% on Cube-RB and from 14% to 34% on LIBERO-Object. Our motion-guided objective further increases success to 78% and 54%, exceeding InfoNCE by 22 and 20 percentage points, respectively. These gains support the benefit of incorporating observed motion similarities into contrastive supervision. Removing SIGReg from the full objective reduces success from 78% to 58% on Cube-RB and from 54% to 14% on LIBERO-Object. These results indicate that motion-guided supervision and batch-level distributional regularization make complementary contributions.

Table 3: Ablation of training objectives on Cube-RB and LIBERO-Object. Entries are planning success (%). All variants retain the prediction loss. InfoNCE uses the same within-episode candidates ($N = 6$) and latent transition scores ($\tau_z = 0.1$) as our method.

SIGReg	Contrastive objective	Cube-RB	LIBERO-Object
✓	None	52	14
✓	InfoNCE	56	34
–	Motion-guided	58	14
✓	Motion-guided	**78**	**54**

Additional ablations, including sensitivity to the temperatures τ_f and τ_z, the number of additional candidates N, the SIGReg weight λ, and the contrastive loss weight β, are provided in Appendix B.

6 Conclusion

We identified static collapse in JEPA-style world models, where representations preserve episode-specific static content while suppressing manipulation-relevant changes. We introduced motion-guided contrastive regularization, which uses optical flow to guide comparisons among observations from the same episode. Combined with latent prediction and SIGReg, it improves planning success from 52% to 78% on Cube-RB and from 14% to 54% on LIBERO-Object, with comparable performance on four other settings. Representation analyses and ablations support improved preservation of manipulation state and the complementary role of SIGReg.

Limitation and future work. Like LeWM, our method depends on sufficient action coverage and remains sensitive to shifts in action representations and visual appearance, including changes to LIBERO's rendering engine. It also assumes static backgrounds; background motion could distort the optical-flow similarities used for supervision. Future work should improve transfer across actions and appearances, isolate task-relevant motion in dynamic scenes, and extend planning to longer horizons.

AI USE STATEMENT

We used generative AI tools to assist with method implementation, refine hypotheses, provide feedback on methodology and experimental design, and interpret results. We did not use them to generate synthetic datasets; develop theoretical or conceptual frameworks; formulate mathematical claims or assist with proofs; conduct qualitative or thematic data analysis; translate text; or clean or reformat datasets.

We also used generative AI tools to search for and summarize literature, suggest experimental parameters, create or edit code and figures, draft and edit manuscript text, and suggest the paper's structure, title, keywords, and references. The authors reviewed and verified all AI-assisted work and take responsibility for the final text, claims, code, and figures.

REPRODUCIBILITY STATEMENT

We will release code, configuration files, and instructions needed to reproduce our experiments. Additional experimental details and ablations are provided in Appendices A and B.

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. *arXiv preprint arXiv:2301.08243*, 2023.

Mahmoud Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, 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, 2025. URL https://arxiv.org/abs/2511.08544.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

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.

Maxime Burchi and Radu Timofte. Mudreamer: Learning predictive world models without reconstruction. *ArXiv*, abs/2405.15083, 2024. URL https://api.semanticscholar.org/CorpusID:270045288.

Maxime Burchi and Radu Timofte. Learning transformer-based world models with contrastive predictive coding. *ArXiv*, abs/2503.04416, 2025. URL https://api.semanticscholar.org/CorpusID:276813068.

Anna Dawid and Yann LeCun. Introduction to latent variable energy-based models: a path toward autonomous machine intelligence. *Journal of Statistical Mechanics: Theory and Experiment*, 2024(10):104011, 2024.

Fei Deng, Ingook Jang, and Sungjin Ahn. Dreamerpro: Reconstruction-free model-based reinforcement learning with prototypical representations. In *International Conference on Machine Learning*, pp. 4956–4975. PMLR, 2022.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Paarth Gulati and Ilya Nemenman. Contrast encodes inductive bias: separating slow noise from dynamics in predictive representation learning, 2026. URL https://arxiv.org/abs/2606.07770.

David R Ha and Jürgen Schmidhuber. World models. *ArXiv*, abs/1803.10122, 2018.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019a.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International conference on machine learning*, pp. 2555–2565. PMLR, 2019b.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

Danijar Hafner, Wilson Yan, and Timothy P. Lillicrap. Training agents inside of scalable world models. *ArXiv*, abs/2509.24527, 2025.

Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. *ArXiv*, abs/2010.09709, 2020. URL https://api.semanticscholar.org/CorpusID:224703413.

Nick Hansen, Hao Su, and Xiaolong Wang. Td-mpc2: Scalable, robust world models for continuous control. In *International Conference on Learning Representations*, volume 2024, pp. 47376–47405, 2024.

Nicklas Hansen, Xiaolong Wang, and Hao Su. Temporal difference learning for model predictive control. In *International Conference on Machine Learning*, 2022. URL https://api.semanticscholar.org/CorpusID:247318709.

Michael Hauri and Friedemann Zenke. Dreamer-cdp: Improving reconstruction-free world models via continuous deterministic representation prediction. *ArXiv*, abs/2603.07083, 2026. URL https://api.semanticscholar.org/CorpusID:286372117.

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

Thomas Kipf, Elise Van der Pol, and Max Welling. Contrastive learning of structured world models. *arXiv preprint arXiv:1911.12247*, 2019.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. LIBERO: Benchmarking knowledge transfer for lifelong robot learning. *arXiv preprint arXiv:2306.03310*, 2023.

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

Aravindh Mahendran, James Thewlis, and Andrea Vedaldi. Cross pixel optical flow similarity for self-supervised learning. In *Asian Conference on Computer Vision*, 2018. URL https://api.semanticscholar.org/CorpusID:49864012.

K. Michael Martini, Eslam Abdelaleem, Paarth Gulati, and Ilya Nemenman. Information bottleneck for learning the phase space of dynamics from high-dimensional experimental data. *ArXiv*, abs/2604.24662, 2026. URL https://api.semanticscholar.org/CorpusID:287782626.

Vincent Micheli, Eloi Alonso, and François Fleuret. Transformers are sample-efficient world models. *arXiv preprint arXiv:2209.00588*, 2022.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

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

Reuven Y. Rubinstein. The cross-entropy method for combinatorial and continuous optimization. *Methodology And Computing In Applied Probability*, 1:127–190, 1999.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, et al. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609, 2020.

Uladzislau Sobal, Wancong Zhang, Kyunghyun Cho, Randall Balestriero, Tim GJ Rudner, and Yann LeCun. Learning from reward-free offline data: A case for planning with latent dynamics models. *Advances in Neural Information Processing Systems*, 38:43905–43941, 2026.

Vlad Sobal, S. V. Jyothir, Siddhartha Jalagam, Nicolas Carion, Kyunghyun Cho, and Yann LeCun. Joint embedding predictive architectures focus on slow features. *arXiv preprint arXiv:2211.10831*, 2022.

Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, Timothy P. Lillicrap, and Martin A. Riedmiller. Deepmind control suite. *ArXiv*, abs/1801.00690, 2018. URL https://api.semanticscholar.org/CorpusID:6315299.

Basile Terver, Tsung-Yen Yang, Jean Ponce, Adrien Bardes, and Yann LeCun. What drives success in physical planning with joint-embedding predictive world models?, 2025. URL https://arxiv.org/abs/2512.24497.

Fida Mohammad Thoker, Hazel Doughty, and Cees G. M. Snoek. Tubelet-contrastive self-supervision for video-efficient generalization. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13766–13777, 2023. URL https://api.semanticscholar.org/CorpusID:257631855.

Leonardo F. Toso, Davit Shadunts, Yunyang Lu, Nihal Sharma, Donglin Zhan, Nam H. Nguyen, and James Anderson. Learning invariant visual representations for planning with joint-embedding predictive world models. *arXiv preprint arXiv:2602.18639*, 2026.

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

Laurens van der Maaten and Geoffrey E. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008. URL https://api.semanticscholar.org/CorpusID:5855042.

Shengjie Wang, Shaohuai Liu, Weirui Ye, Jiacheng You, and Yang Gao. Efficientzero v2: Mastering discrete and continuous control with limited data. In *International Conference on Machine Learning*, 2024a. URL https://api.semanticscholar.org/CorpusID:268230777.

Yihan Wang, Lahav Lipson, and Jia Deng. Sea-raft: Simple, efficient, accurate raft for optical flow. In *European Conference on Computer Vision*, pp. 36–54. Springer, 2024b.

Fanyi Xiao, Joseph Tighe, and Davide Modolo. Maclr: Motion-aware contrastive learning of representations for videos. In *European conference on computer vision*, pp. 353–370. Springer, 2022.

Weirui Ye, Shaohuai Liu, Thanard Kurutach, Pieter Abbeel, and Yang Gao. Mastering atari games with limited data. *Advances in neural information processing systems*, 34:25476–25488, 2021.

Zhenghao Zhang, Yuanxiang Wang, Zhenyu Guan, Yujia Yang, Bingkang Shi, Tianyu Zong, Hongzhu Yi, Guoqing Chao, Xingchen Chen, Tiankun Yang, et al. Delta-jepa: Learning action-sensitive world models via latent difference decoding. *arXiv preprint arXiv:2606.31232*, 2026.

Ruijie Zheng, Xiyao Wang, Yanchao Sun, Shuang Ma, Jieyu Zhao, Huazhe Xu, Hal Daumé, and Furong Huang. Taco: Temporal latent action-driven contrastive loss for visual reinforcement learning. *ArXiv*, abs/2306.13229, 2023. URL https://api.semanticscholar.org/CorpusID:259244013.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. Dino-wm: World models on pre-trained visual features enable zero-shot planning, 2024. URL https://arxiv.org/abs/2411.04983.

APPENDIX

A ADDITIONAL EXPERIMENTAL DETAILS

A.1 LIBERO-OBJECT DATA PREPARATION

Dataset partitioning. We use all tasks in LIBERO-Object, which provides 500 expert trajectories. Before generating additional data, we reserve three expert trajectories per task for evaluation and use the remaining trajectories for training. We generate perturbed continuations only from training trajectories, ensuring that held-out trajectories and their continuations are excluded from the training data.

Perturbed trajectory generation. To increase training diversity around the demonstrated behaviors, we generate additional trajectories by perturbing expert actions. For each training trajectory, we uniformly sample three valid starting frames without replacement. At each starting frame, we restore the recorded simulator state and execute three perturbed continuations of up to 100 steps each, yielding nine candidate rollouts per training demonstration.

We perturb the continuous pose components of the expert action sequence with zero-mean Gaussian noise and clip the resulting actions to $[-1, 1]$. The three variants use noise standard deviations $\sigma \in \{0.15, 0.33, 0.5\}$, paired with per-step gripper-flip probabilities $p \in \{0.006, 0.03, 0.09\}$, respectively. For each continuation, the discrete gripper command is initialized from the expert command at the starting frame and subsequently undergoes stochastic sign flips with the corresponding probability. We record the executed actions and resulting simulator observations.

Filtering and dataset composition. We discard a rollout if it contains fewer than five recorded steps, if the target object's vertical coordinate falls below $z = -0.60$ at any recorded step in simulator coordinates, or if the Euclidean displacement between its first and last recorded positions is less than $0.03\,\mathrm{m}$. We also tag distractor motion when the first-to-last change in selected static object-state coordinates has an ℓ_2 norm greater than 0.05. Rollouts exhibiting distractor motion are retained and assigned the tag `distractor_moved=1`.

After filtering, we retain 4,108 perturbed trajectory chunks, which are combined with the training expert trajectories. All methods evaluated on LIBERO-Object use this same expanded training dataset, with one model per method trained jointly across all LIBERO-Object tasks.

A.2 LATENT PLANNING AND EVALUATION PROTOCOL

Planning objective. Given a current observation x_t and a goal observation x_g, we compute their embeddings $z_t = \mathrm{enc}(x_t)$ and $z_g = \mathrm{enc}(x_g)$. For each candidate action sequence $\mathbf{a} = (a_t, \ldots, a_{t+H-1})$, the learned dynamics model autoregressively predicts a terminal embedding $\hat{z}_{t+H}(\mathbf{a})$. We optimize the action sequence using the Cross-Entropy Method (CEM) to minimize the terminal goal-matching cost

$$\mathcal{C}(\mathbf{a}) = \|\hat{z}_{t+H}(\mathbf{a}) - z_g\|_2^2 . \tag{13}$$

All methods use the same planning procedure and optimization budget, allowing us to compare how their learned representations and dynamics support goal-directed control.

CEM and replanning. At each control step, CEM performs 10 optimization iterations, sampling 128 candidate action sequences per iteration. The 16 sequences with the lowest predicted costs form the elite set used to update the sampling distribution. One control step spans five environment timesteps, with observations sampled at this interval. Each evaluation instance has a total execution budget of five control steps, corresponding to 25 environment timesteps. After optimization, we execute only the first action of the selected sequence, obtain a new observation after one control step, and replan toward the same goal. The planning horizon matches the remaining execution budget, decreasing from five to one control step over successive replanning rounds: $H \in \{5, 4, 3, 2, 1\}$.

Evaluation instances and checkpoints. For each benchmark setting, we evaluate 50 initial–goal pairs using evaluation seed 42. Each pair is constructed by sampling an initial timestep from a held-out trajectory and selecting the goal observation five control steps later in the same trajectory. For

LIBERO-Object, the 50 instances span all ten tasks; this is the total evaluation count rather than the count per task. All methods are evaluated on the same initial–goal pairs, and we use the final saved training checkpoint for each method. Success rates are computed over the 50 evaluation instances in each benchmark setting.

Success criteria. Success is determined by matching the underlying environment state to the sampled goal state using the following criteria:

- **Push-T.** The combined Euclidean position error of the pusher and T-block must be less than 20 pixels in the native 512×512 coordinate system, and the wrapped T-block orientation error must be less than $\pi/9$ radians ($20°$).
- **Reacher.** The absolute difference between each joint angle and its goal value must be less than 0.05 radians, following the `qpos_match` configuration.
- **2-Rooms.** The Euclidean distance between the agent's center and its goal position must be less than 16 pixels in the 224×224 environment coordinate system.
- **Cube and Cube-RB.** The Euclidean distance between the cube's 3D position and its goal position must be at most $0.04\,\mathrm{m}$. Cube orientation is excluded from the success criterion. Cube-RB uses the same criterion as Cube because background replacement changes only the observations, while preserving the underlying states and dynamics.
- **LIBERO-Object.** The maximum Euclidean position error across the task's movable objects of interest must be at most $0.05\,\mathrm{m}$ relative to their positions in the sampled goal state. Object orientations and robot joint errors are excluded. This criterion evaluates attainment of the sampled goal state rather than completion of the original LIBERO task objective.

Initially satisfied goals. For Cube and Cube-RB, initial states that already satisfy the goal criterion are counted as successes, following the original LeWM (Maes et al., 2026) evaluation implementation. For LIBERO-Object, we reject initially satisfied initial–goal pairs and resample until 50 valid evaluation instances are obtained.

A.3 Model and Training Implementation

Architecture and optimization. Visual encoder. We use a ViT-Tiny/14 encoder trained from scratch, with 12 transformer blocks, 192-dimensional tokens, 3 attention heads, and 768-dimensional feed-forward layers with GELU. Each 224×224 RGB image is divided into 256 non-overlapping 14×14 patches, supplemented with a learned CLS token and learned positional embeddings. The final CLS feature passes through a projection head consisting of a linear layer from 192 to 2048 dimensions, BatchNorm, GELU, and a linear layer back to 192 dimensions.

Temporal predictor. Observations are encoded independently and passed to an action-conditioned causal transformer with 6 blocks, 16 attention heads, 64 dimensions per head, and 2048-dimensional feed-forward layers. The predictor's input and output dimensions are both 192. Each training clip contains three context observations and one future observation, sampled at intervals of five environment timesteps. Using the subsampled time index of the main text, the predictor uses (x_{t-2}, x_{t-1}, x_t) and the associated actions to predict the embedding of x_{t+1}.

Optimization. LeWM and our method share the same architecture and optimization settings, differing only in the addition of the motion-guided contrastive objective. We train for 100 epochs with a batch size of 128 using AdamW, a learning rate of 5×10^{-5}, weight decay of 10^{-3}, momentum coefficients $(0.9, 0.999)$, and $\epsilon = 10^{-8}$. The learning rate follows a linear warmup followed by cosine decay.

Candidate sampling and motion supervision. Additional candidate observations are sampled from the same episode as the current observation, excluding the current frame x_t and the positive frame x_{t+1}. Candidates may come from any other timestep in the episode.

We compute optical flow using the pretrained SEA-RAFT Spring-M checkpoint `MemorySlices/Tartan-C-T-TSKH-spring540x960-M`. The flow model is frozen and runs in evaluation mode with four refinement iterations and `scale=0`. Flow is computed at 224×224 resolution, with displacement vectors measured in pixels at that resolution. The offline pipeline converts stored RGB frames to channel-first floating-point tensors in $[0, 255]$, which

SEA-RAFT internally normalizes to $[-1, 1]$. No additional resizing or augmentation is applied during flow precomputation for the native 224×224 frames.

For each anchor frame $\mathbf{x}_t$, we randomly sample 40 candidate frames from the same episode, excluding $\mathbf{x}_t$ and $\mathbf{x}_{t+1}$, and precompute their motion-weighted distances $d_{t,j}$ (Eq. 9) from the anchor-to-positive and anchor-to-candidate flow fields. Only these scalar distances are cached, in float32 format; the dense flow fields are not retained. During training, we randomly draw N of the 40 cached candidates for each transition ($N = 6$ by default) and construct the soft contrastive targets from their cached distances. SEA-RAFT inference is therefore performed entirely offline, and no flow model is run during world-model training.

Training objective. We retain LeWM's latent prediction loss and SIGReg regularization, and add the motion-guided contrastive loss defined in the main text. SIGReg is applied to the pre-projection encoder embeddings of the current and next observations, using $M = 1024$ random projections and regularization weight $\lambda = 0.09$. Unless otherwise specified, the contrastive objective uses $\tau_f = 20$, $\tau_z = 0.1$, and $\beta = 0.3$.

A.4 Representation-Analysis implementation

Frozen-representation evaluation. We assess the information retained by the learned representations through image reconstruction and physical-state probing. For each method, we freeze the world-model parameters and train auxiliary decoders or probes on the resulting embeddings. Only the auxiliary model is optimized during these analyses.

Image reconstruction. We train a lightweight transformer decoder to reconstruct an observation from its corresponding latent embedding. These reconstructions provide a qualitative assessment of the visual information recoverable from each representation, particularly the configurations of the robot and manipulated objects.

On Cube and Cube-RB, both methods use the same episode-level partition, with 9,000 training episodes and 1,000 validation episodes, so all clips from a given episode remain within a single subset.

For LIBERO-Object, both methods use 47 source demonstrations per task for decoder training and three for validation. Perturbed continuations inherit the partition of their source demonstration, ensuring that each demonstration and its perturbed descendants remain in the same subset.

Physical-state implementation. We train two-layer MLP probes on frozen encoder embeddings to quantify the recoverability of manipulation-relevant state information. On Cube and Cube-RB, the probes predict the 3D positions of the end effector and cube. On LIBERO-Object, they predict the end-effector position and a binary object-holding indicator. We measure position-prediction errors using Euclidean distance in centimeters and evaluate object-holding classification using the area under the receiver operating characteristic curve (AUROC).

The numerical probing experiments use episode-level partitions into disjoint training, validation, and test subsets. For each of Cube and Cube-RB, these subsets contain 7,000, 1,500, and 1,500 episodes, respectively, corresponding to 56,000, 12,000, and 12,000 frames. For LIBERO-Object, they contain 350, 70, and 80 episodes, corresponding to 10,850, 2,160, and 2,487 frames. All frames from a given episode remain within a single subset.

B Additional Results

We examine the sensitivity of motion-guided contrastive regularization to the flow temperature τ_f, latent temperature τ_z, number of additional within-episode candidates N, SIGReg weight λ, and contrastive loss weight β. Each sweep varies one hyperparameter while keeping the remaining settings fixed, including the architecture, training schedule, and planning procedure. The reference settings are $\tau_f = 20$, $\tau_z = 0.1$, $N = 6$, $\lambda = 0.09$, and $\beta = 0.3$. Table 4 summarizes planning success on LIBERO-Object under the evaluation protocol in Appendix A.2.

Table 4: Hyperparameter sensitivity on LIBERO-Object. "Succ." denotes planning success (%). Bold marks the highest result within each sweep; underlining marks the reference parameter value.

(a) Flow temp.		**(b) Latent temp.**		**(c) Candidates**		**(d) SIGReg**		**(e) Contrastive**	
τ_f	Succ.	τ_z	Succ.	N	Succ.	λ	Succ.	β	Succ.
2.5	40	0.02	52	2	52	0	14	0.05	50
5	36	0.05	**54**	4	48	0.03	40	0.1	**54**
10	46	<u>0.1</u>	**54**	<u>6</u>	54	<u>0.09</u>	**54**	<u>0.3</u>	**54**
<u>20</u>	**54**	0.2	48	8	50	0.18	**54**	0.6	52
40	40	0.5	50	10	54	0.3	50	1.0	38
80	34	0.75	44	12	**58**				

B.1 Flow temperature τ_f

The flow temperature τ_f controls the concentration of the motion-derived target distribution. Smaller values assign more probability to candidates with the lowest flow distances from the positive, whereas larger values distribute probability more evenly across candidates.

As shown in Table 4(a), planning success reaches 54% at $\tau_f = 20$, the highest result among the tested settings. Success is lower at both smaller and larger temperatures, ranging from 36% to 46% for $\tau_f \in \{2.5, 5, 10\}$ and from 34% to 40% for $\tau_f \in \{40, 80\}$. These results favor an intermediate temperature within the tested range. We use $\tau_f = 20$ in the main experiments.

B.2 Latent temperature τ_z

The latent temperature τ_z controls how cosine similarities between predicted and candidate latent transitions are mapped to a probability distribution. Smaller values produce sharper distributions, while larger values produce smoother distributions. Thus, τ_z controls the predicted distribution, whereas τ_f controls the motion-derived supervision targets.

Table 4(b) shows that planning success ranges from 52% to 54% for $\tau_z \in \{0.02, 0.05, 0.1\}$, compared with 44%–50% for the larger tested values. The highest observed success is shared by $\tau_z = 0.05$ and $\tau_z = 0.1$. We use $\tau_z = 0.1$ in the main experiments.

B.3 Number of additional candidates N

We vary the number of additional observations sampled from the same episode, evaluating $N \in \{2, 4, 6, 8, 10, 12\}$ while fixing $\tau_f = 20$ and $\tau_z = 0.1$. Each comparison contains $N + 1$ candidates, including the true next observation.

As shown in Table 4(c), increasing N does not yield a monotonic improvement in planning success. The highest observed success is 58% at $N = 12$, compared with 54% at our default $N = 6$. We retain $N = 6$ to limit the number of additional observations encoded per training transition.

B.4 SIGReg weight λ

We examine the strength of distributional regularization by varying the SIGReg weight λ while fixing the contrastive loss weight at $\beta = 0.3$. All configurations retain the prediction and motion-guided contrastive losses.

As shown in Table 4(d), removing SIGReg yields 14% planning success, consistent with the ablation in Table 3. Adding SIGReg increases success to 40% at $\lambda = 0.03$ and 54% at both $\lambda = 0.09$ and $\lambda = 0.18$, while $\lambda = 0.3$ yields 50%. These results support the complementary role of distributional regularization alongside motion-guided supervision. We use $\lambda = 0.09$ as the default.

B.5 CONTRASTIVE LOSS WEIGHT β

We vary the weight β of the motion-guided contrastive loss while fixing the SIGReg weight at $\lambda = 0.09$. This experiment evaluates how the strength of motion-guided supervision affects planning performance.

Table 4(e) shows that $\beta = 0.1$ and $\beta = 0.3$ both achieve the highest observed success of 54%. The settings $\beta = 0.05$ and $\beta = 0.6$ yield 50% and 52%, respectively, whereas increasing the weight to $\beta = 1.0$ reduces success to 38%. Thus, assigning greater weight to the contrastive objective does not necessarily improve planning performance. We retain $\beta = 0.3$ as the default.