# LRC-JEPA: DISENTANGLING DYNAMICS AND RESIDUAL CONTEXT FOR EFFICIENT WORLD MODELS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Compact JEPA world models enable efficient latent-space planning, but low-dimensional representation trained under reward-free self-supervision must encode both action-conditioned dynamics and predictable visual context. This competition can entangle controllable state with high-rank nuisance appearance and degrade planning as scenes become more complex. We introduce LRC-JEPA, a lightweight end-to-end world model that routes information into a compact predictive latent $\mathbf{z}$ and learned-query residual-context embeddings $\mathbf{u}$. Only $\mathbf{z}$ is propagated by the dynamics model and used for planning, while $\mathbf{u}$ captures temporally persistent information for cross-attention reconstruction; a differentiable residual connection encourages the latent to retain complementary dynamic content. Under explicit assumptions, we show that the resulting representation is sufficient, minimal, nuisance-invariant, and disentangled. Across four simulated control environments, LRC-JEPA improves average planning success over a parameter-matched JEPA baseline by 9 percentage points and matches or exceeds substantially larger pretrained models. On the real-world Bridge-v2 set, its 5.5M-parameter active encoder outperforms DINO-WM (22.1M) and V-JEPA2 (303.9M) encoders while also enabling faster planning. Physical-state probes, reconstruction interventions, and ablations confirm the effectiveness of LRC-JEPA's representation disentanglement.

## 1 INTRODUCTION

Learning an internal model of the physical world is a central objective of embodied intelligence. World models infer state from high-dimensional observations, predict the effects of candidate actions, and support planning toward unseen goals. Learning them from offline, reward-free interaction is attractive because the same trajectories can serve multiple tasks without reward annotation, online exploration, or policy retraining Ha & Schmidhuber (2018); Sobal et al. (2025). Visual world models have consequently shifted from pixel prediction Finn et al. (2016); Ebert et al. (2018) toward dynamics in compressed representation spaces Hafner et al. (2019; 2025); Hansen et al. (2024), enabling policy learning or efficient test-time optimization.

One line of latent world models predicts rich patch-level features from large pretrained encoders, as in DINO-WM and V-JEPA2 Zhou et al. (2025); Assran et al. (2025). These features retain fine-grained semantic and spatial information but make repeated planning expensive Maes et al. (2026). Compact alternatives jointly learn dynamics over one or a few scene embeddings Hafner et al. (2019); Sobal et al. (2025); Maes et al. (2026); LeWorldModel (LWM), for example, predicts a single latent state and avoids pixel reconstruction. However, this efficiency creates a *capacity–compactness trade-off*: low-dimensional encoders can discard action-relevant transition information or entangle controllable factors with predictable but weakly action-dependent context Pan et al. (2022); Cui et al. (2026). The problem is acute in reward-free image-goal planning, where future goals may depend on context that cannot be discarded a priori. Figure 1 exposes this failure mode: a single-latent JEPA degrades under PCA compression and task-irrelevant visual perturbations, indicating entanglement of nuisance with control-relevant dynamics and inefficient use of its limited capacity.



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.

Motivated by optimal representation theory Achille & Soatto (2018a;b), we propose **LRC-JEPA**, a lightweight end-to-end JEPA world model with two functionally distinct representations: a compact predictive latent $\mathbf{z}_t$ and learned-query residual-context embeddings $\mathbf{u}_t$. Our contributions are:

- We introduce a functional routing architecture that assigns predictive dynamics to compact $\mathbf{z}$ and slow-varying or persistent context to complementary $\mathbf{u}$.
- Under explicit assumptions, we show that LRC-JEPA satisfies sufficiency, minimality, nuisance invariance, and coordinate disentanglement, with $\mathbf{u}_t$ and $\mathbf{z}_t$ sufficient for context and dynamic state, respectively.
- Across simulated and real-world settings, LRC-JEPA's 5.5M active parameter encoder improves over a parameter-matched JEPA and surpasses up to 55$\times$ larger pretrained competitors on key planning metrics and planning time consumption. Probes and interventions confirm the intended allocation of kinematics to $\mathbf{z}$ and residual context to $\mathbf{u}$.

We also make the codes and pretrained checkpoints of LRC-JEPA available at Codes and Models, respectively.

## 2 Related Work

**Visual world models and model-based control.** World models predict action consequences for planning and policy learning Ha & Schmidhuber (2018). Early visual methods predict future pixels for robotic control Finn et al. (2016); Ebert et al. (2018), while recent foundation models scale video generation to broad physical-AI data NVIDIA (2025). Such models preserve visual detail but must also spend substantial capacity on appearance variation that may be irrelevant to a particular control objective. Latent approaches instead support planning or policy learning through compressed states Hafner et al. (2019; 2025); Hansen et al. (2022; 2024). PlaNet and Dreamer combine latent dynamics with observation and reward prediction, whereas TD-MPC methods emphasize task-oriented latent dynamics, reward prediction, and value estimation. These objectives provide signals for retaining task-relevant information. Reward-free visual planning poses a different challenge: the representation must preserve action-dependent state and distinguish image goals without reward supervision. LRC-JEPA addresses this setting by learning a compact predictive state alongside a separate context representation used for reconstruction during training.

**Joint-embedding predictive architecture (JEPA).** JEPAs learn by predicting target representations rather than reconstructing every input detail. I-JEPA introduces masked representation prediction for images Assran et al. (2023), and V-JEPA extends this approach to video Bardes et al. (2024). Existing methods prevent collapse with asymmetric or stop-gradient target encoders. LeJEPA instead uses teacher-free distributional regularization Balestriero & LeCun (2025), which LWM applies to compact action-conditioned dynamics learned from pixels Maes et al. (2026). Alternatively, DINO-WM and V-JEPA2 retain much richer pretrained features Oquab et al. (2024); Zhou et al. (2025); Assran et al. (2025). These methods successfully establish feature prediction for control. However, preventing collapse does not inherently determine which physical variables a predictive representation preserves.

**Functional separation in representation learning.** Separating appearance and dynamics has a long history in video generation and prediction Denton & Birodkar (2017); Tulyakov et al. (2018); Hsieh et al. (2018); MotionRNN further distinguishes motion trends from transient variation Wu et al. (2021). MC-JEPA jointly learns motion and content features through optical-flow and self-supervised representation objectives Bardes et al. (2023). For control, bisimulation uses reward and transition equivalence to suppress irrelevant variation Zhang et al. (2021). Without rewards, predictive objectives can favor slowly varying distractors over changing task state Sobal et al. (2022). Recent work addresses related allocation problems through progression–content subspaces Thil et al. (2026) or Gaussian regularization of temporally centered residuals Liu et al. (2026). LRC-JEPA uses a different division of responsibility: action-conditioned prediction shapes $\mathbf{z}$, while reconstruction and temporal invariance encourage $\mathbf{u}$ to retain complementary, slowly varying context. This is a functional bias rather than a prescribed semantic partition; slowly varying quantities can still matter for control. The context stream and decoder support representation learning, while planning operates only on $\mathbf{z}$, connecting the separation directly to inference efficiency.

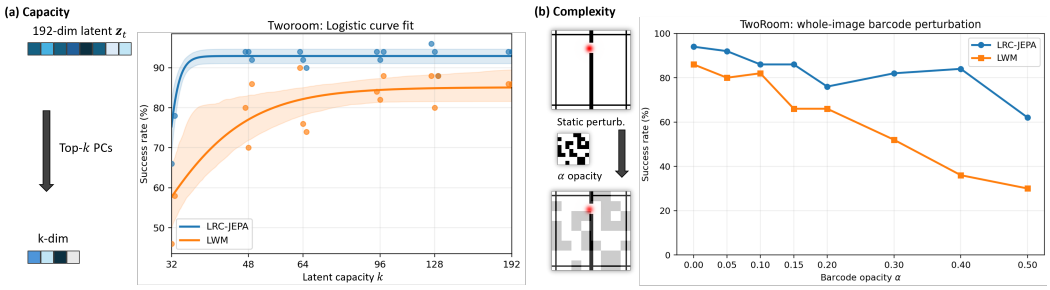



Figure 1: **Zero-shot capacity–complexity stress tests for compact JEPA planning.** *Left:* Starting from the original 192-dimensional predictive latent, PCA is fitted on clean representations and latents are projected onto top-$k$ components during planning. *Right:* In TwoRoom, a trajectory-persistent grayscale pattern is blended into both current and goal observations with opacity $\alpha$, where $\alpha = 0$ denotes the clean environment. Details are provided in Appendix A.2.

# 3 METHOD

We first use optimal representation theory to identify a limitation of single-latent JEPAs, then present LRC-JEPA and establish its minimality and sufficiency properties.

## 3.1 THEORETICAL ANALYSIS FOUNDATIONS

We adopt the Achille–Soatto criteria of sufficiency, minimality, invariance, and disentanglement Achille & Soatto (2018b); definitions appear in Appendix B.1.

**Theorem 1** (Achille–Soatto criterion; Achille & Soatto, 2018a, Prop. 3.1 and §5)**.** *Let $Y$ be a task, $N$ a nuisance ($I(Y;N) = 0$), and $Z$ a sufficient representation of the input $X$ such that the mutual information $I(Y;Z) = I(Y;X)$, with $N \to X \to Z$. (a)* Sufficiency, minimality $\Rightarrow$ Invariance: *every nuisance satisfies $I(Z;N) \leq I(Z;X) - I(X;Y)$ (Remark 1); since $I(X;Y)$ is constant on given data, a sufficient $Z$ is maximally invariant w.r.t $N$ if $I(Z;X)$ minimizes. (b)* Minimality $\Rightarrow$ coordinate disentanglement: *For a stochastic-network model $Z = g_z(X)$, the sum of mutual information and total correlation (TC) is upper bounded by the information of training data, so a weight-information regularization enforcing minimality drives the total correlation $\mathrm{TC}(Z)$ down with it. Thus, invariance and coordinate disentanglement follow one mechanism: a penalty that decreases $I(Z;X)$ while sufficiency is maintained.*

We next characterize the single-latent LWM under these criteria.

**Proposition 1.** ***Sufficiency.*** *Under Assumptions 1, 2, 3, the population SIGReg objective vanishes if and only if $q(\boldsymbol{z}_t) = \mathcal{N}(0, I_d)$, in which case $\mathrm{TC}(Z_t) = 0$ and the coordinates are mutually independent and disentangled. Moreover, the latent representation learned by LWM is sufficient for predicting the underlying state.* ***No guaranteed minimality.*** *For every $\lambda \geq 0$,* SIGReg *is unchanged by encoder modifications that preserve $q(\boldsymbol{z}_t)$, including those changing the mutual information $I(Z_t;X_t)$. In fact, $I(Z_t;X_t)$ is pushed up when $I(Z_t;N)$ maximizes; and $\mathcal{L}_{\mathrm{pred}}(Z_t)$ only rewards information about the transition but not penalizes $I(Z;N)$. Hence, during training no term of $\mathcal{L}_{\mathrm{LWM}}$ decreases $I(Z_t;X_t)$, and the minimizers of $\mathcal{L}_{\mathrm{LWM}}$ does not generally guarantee minimal prediction-sufficient statistics.*

The proof is in Appendix B. SIGReg's unique population minimizer $\mathcal{N}(0, I_d)$ prevents collapse, fixes scale, and decorrelates features, but does not prevent control-irrelevant nuisance information from occupying the latent. This limitation is consistent with Figure 1.

## 3.2 MODEL ARCHITECTURE

An observation $\mathbf{i}_t \in \mathbb{R}^{3\times H\times W}$ contains action-dependent state and slowly varying context such as background, lighting, camera configuration, and appearance. LRC-JEPA assigns them comple-

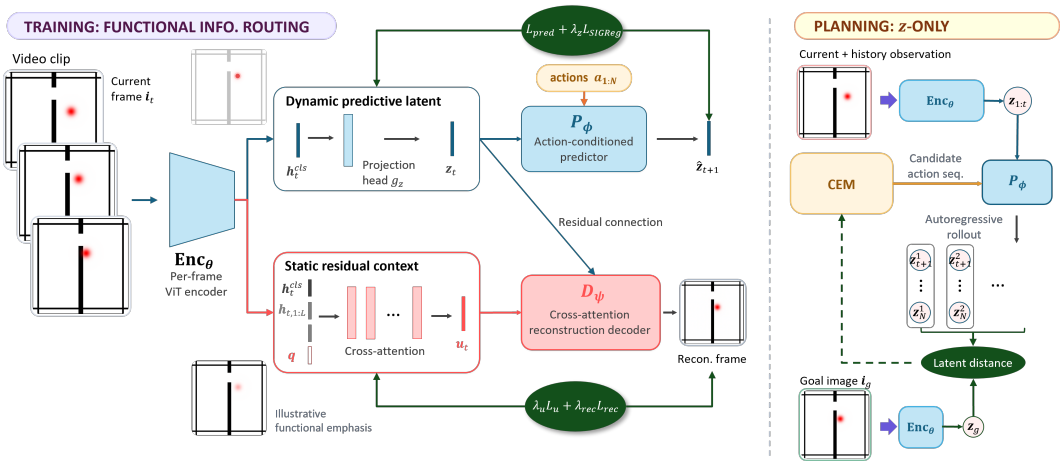



Figure 2: **LRC-JEPA architecture.** During training, functional information routing separates a compact, dynamics-predictive latent $z_t$ from an auxiliary representation $u_t$ that supports reconstruction through cross-attention. This separation is the key novelty of our method: it concentrates control-relevant dynamics in $z$, enabling efficient $z$-only planning with CEM and latent-distance objectives at test time.

mentary roles: $\mathbf{z}_t$ is a compact *predictive latent state* used by the prediction model and planning objective, whereas $\mathbf{u}_t$ is a bottlenecked *residual-context embedding* used only for reconstruction.

**Shared visual encoder.** Each frame ($\mathbf{i}_t$) is processed independently by a standard Vision Transformer Dosovitskiy et al. (2021), denoted by $\mathrm{Enc}_\theta$. We represent its final-layer class and patch tokens as

$$\left[\mathbf{h}_t^{\mathrm{cls}}, \mathbf{h}_{t,1}, \ldots, \mathbf{h}_{t,L}\right] = \mathrm{Enc}_\theta(\mathbf{i}_t) \in \mathbb{R}^{(L+1)\times d_h}, \tag{1}$$

where $\mathbf{h}_t^{\mathrm{cls}}, \mathbf{h}_{t,\ell}$ denote the class token and the $\ell$-th patch token, respectively. The predictive latent is obtained from the class token using a two-layer projection head,

$$\mathbf{z}_t = g_z\left(\mathbf{h}_t^{\mathrm{cls}}\right) \in \mathbb{R}^{d_z}, \qquad g_z = \mathrm{Linear} \circ \mathrm{GELU} \circ \mathrm{BN} \circ \mathrm{Linear}, \tag{2}$$

**Learned-query residual context.** To retain spatial and appearance information without passing all patch tokens to the dynamics model, we attentively pool the complete token set using $Q$ learnable queries $\mathbf{q} \in \mathbb{R}^{Q\times d_u}$. With $\mathbf{H}_t = [\mathbf{h}_t^{\mathrm{cls}}; \mathbf{h}_{t,1:L}]$, $W_u$ as a linear projection from ViT feature to context token domain, the context encoder is

$$\widetilde{\mathbf{u}}_t = \mathrm{MHA}\left(\mathrm{LN}(\mathbf{q}), \mathrm{LN}(W_u\mathbf{H}_t), \mathrm{LN}(W_u\mathbf{H}_t)\right), \tag{3}$$

$$\mathbf{u}_t = \widetilde{\mathbf{u}}_t + \mathrm{MLP}(\widetilde{\mathbf{u}}_t) \in \mathbb{R}^{Q\times d_u}. \tag{4}$$

Thus, $\mathbf{u}_t$ summarizes the spatial feature map through a constrained reconstruction-only bottleneck, discouraging frame-wise shortcuts.

**Action-conditioned latent predictor.** An action encoder $g_a$ maps the normalized action associated with each visual transition to an embedding $\mathbf{e}_t$,

$$\mathbf{e}_t = g_a(\mathbf{a}_t) \in \mathbb{R}^{d_z}. \tag{5}$$

When one visual transition spans $F$ low-level control steps, the $F$ actions are concatenated before being passed to $g_a$. For multi-dataset training, dataset-specific action adapters map heterogeneous action spaces into the common embedding space Maes et al. (2026).

Given a history of $T$ predictive states, a causal transformer $P_\phi$ predicts the subsequent states by teacher forcing:

$$\widetilde{\mathbf{z}}_{1:T} = P_\phi\left(\mathbf{z}_{1:T} + \mathbf{r}_{1:T}; \mathbf{e}_{1:T}\right), \tag{6}$$

$$\widehat{\mathbf{z}}_{t+1} = g_p(\widetilde{\mathbf{z}}_t), \qquad t = 1, \ldots, T. \tag{7}$$

Here $\mathbf{r}_{1:T}$ are learned temporal embeddings and $g_p$ has the same structure as $g_z$. Causal self-attention restricts position $t$ to $\mathbf{z}_{1:t}$, while AdaLN-Zero is adapted to every transformer block to condition on actions Peebles & Xie (2023). Crucially, $\mathbf{u}_t$ never enters $P_\phi$: only $\mathbf{z}_t$ is rolled forward and used for planning.

**Cross-attention reconstruction decoder.** Reconstruction is used to shape the allocation of information between the two latent streams. For each frame, the decoder context is

$$\mathbf{c}_t = [\mathbf{z}_t; \mathbf{u}_{t,1}; \ldots; \mathbf{u}_{t,Q}]\,. \tag{8}$$

We use one learned query $\mathbf{s}_n^{(0)}$ for each output image patch. The $l$-th decoder layer applies

$$\mathbf{s}_t^{(\ell)} \leftarrow \mathbf{s}_t^{(\ell-1)} + \mathrm{MHA}\left(\mathrm{LN}(\mathbf{s}_t^{(\ell-1)}), \mathrm{LN}(W_c\mathbf{c}_t), \mathrm{LN}(W_c\mathbf{c}_t)\right), \tag{9}$$

$$\mathbf{s}_t^{(\ell)} \leftarrow \mathbf{s}_t^{(\ell)} + \mathrm{MLP}\left(\mathrm{LN}(\mathbf{s}_t^{(\ell)})\right). \tag{10}$$

A linear output head maps each final query to a flattened RGB patch, which is unpatchified to $\widehat{\mathbf{i}}_t$.

### 3.3 Training Objective

LRC-JEPA is trained with four complementary objectives: one-step latent prediction, anti-collapse regularization, temporal context regularization, and patch reconstruction. The action-conditioned predictor maps $\mathbf{z}_{1:T}$ to targets $\mathbf{z}_{2:T+1}$ using

$$\mathcal{L}_{\mathrm{pred}} = \frac{1}{BTd_z}\sum_{b=1}^{B}\sum_{t=1}^{T}\left\|\widehat{\mathbf{z}}_{t+1}^{(b)} - \mathbf{z}_{t+1}^{(b)}\right\|_2^2. \tag{11}$$

Following LeJEPA and LWM Balestriero & LeCun (2025); Maes et al. (2026), we regularize the predictive latent with SIGReg loss $\mathcal{L}_{\mathrm{SIG}}$, without a stop-gradient target or momentum encoder. We further apply a temporal VICReg-style objective Bardes et al. (2022) $\mathcal{L}_u$ to encourage the residual context $\mathbf{u}_t$ to remain stable within a trajectory while retaining variation across trajectories. Finally, a patch reconstruction loss

$$\mathcal{L}_{\mathrm{rec}} = \mathrm{MSE}\left(D_\psi(\mathbf{z}_t, \mathbf{u}_t), \mathbf{i}_t\right) \tag{12}$$

encourages both latent streams to preserve scene information. The complete objective is

$$\mathcal{L}_{\text{LRC-JEPA}} = \mathcal{L}_{\mathrm{pred}} + \lambda_z\mathcal{L}_{\mathrm{SIG}} + \lambda_u\mathcal{L}_u + \lambda_{\mathrm{rec}}\mathcal{L}_{\mathrm{rec}}. \tag{13}$$

We use $(\lambda_z, \lambda_u, \lambda_{\mathrm{rec}}) = (0.05, 0.05, 0.2)$. Full definitions and implementation details of each regularizer are provided in Appendix A.4.

### 3.4 Theoretical Analysis for The Proposed Method

LRC-JEPA splits the representation into a predictive latent $\boldsymbol{z}_t \in \mathbb{R}^{d_z}$ (SIGReg-regularized, rolled forward by the action-conditioned predictor, alone used for planning) and a query-bottlenecked *residual context* $\boldsymbol{u}_t$ (regularized to be stable within a trajectory), joined only in the reconstruction decoder $D_\psi(\boldsymbol{z}_t, \boldsymbol{u}_t)$. Theorem 2 shows that this split structurally supplies the ingredient of optimal representation theory in 1.

**Theorem 2** (Optimality of LRC-JEPA)**.** *Under Assumptions 1, 4, and 5 (two-factor Gaussian world $X_t = g(s_t, v)$: hidden state $s_t \in \mathbb{R}_s^d$, trajectory-constant appearance nuisance $v \perp s$), every global minimizer of $\mathcal{L}_{\mathrm{LRC}}$ satisfies: (i)* sufficiency: *$\boldsymbol{u}_t$ is sufficient for $v$, and $\boldsymbol{z}_t$ is a sufficient statistic of $s_t$ conditioned on $u_t$; (ii)* minimality: *$I(Z_t; X_t)$ is minimal among prediction-sufficient representations; (iii)* invariance: *$I(Z_t; V) = 0$; and (iv)* disentanglement: *$q(\boldsymbol{z}_t) = \mathcal{N}(0, I_n)$ and* $\mathrm{TC}(Z_t) = 0$.

Theorem 2 formalizes how two components induce this separation. First, under squared reconstruction loss, information already represented by $U_t$ cannot further reduce the reconstruction residual, so residual-connected reconstruction rewards complementary dynamic information in $Z_t$. Second, detaching $Z_t$ removes this reconstruction constraint and can yield a predictable yet insufficient state. Appendix C gives the full analysis and proof.

# 4 RESULTS

We evaluate latent planning, physical-state accessibility, reconstruction interventions, and architectural ablations in simulation and on Bridge-v2 real-world dataset Walke et al. (2023).

## 4.1 LATENT PLANNING PERFORMANCE

**Evaluation Protocol.** We evaluate four simulated environments and Bridge-v2, spanning 2D and 3D manipulation and motion planning. Simulation uses image-goal model predictive control (MPC): the cross-entropy method (CEM) optimizes candidate action sequences by terminal latent distance, executes the first segment, and replans. Bridge-v2 uses offline expert-action recovery measured by recall and expert-action rank percentile.

**Baseline.** We compare with LWM Maes et al. (2026), V-JEPA2 Assran et al. (2025), DINO-WM Zhou et al. (2025), PLDM Sobal et al. (2025), C-JEPA Nam et al. (2026), and EB-JEPA Terver et al. (2026) using common trajectory splits, observations, actions, and goals (Appendix A.4). LWM Maes et al. (2026) shares LRC-JEPA's 5.5M-parameter encoder, predictive-state interface, inference architecture, and active parameter count, isolating the effect of representation disentanglement.

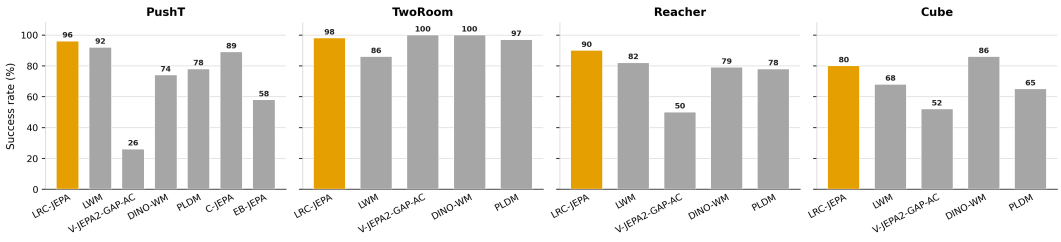



Figure 3: CEM planning success rates on four simulated environments. Values are averaged over three repetitions of 50 evaluation scenes.

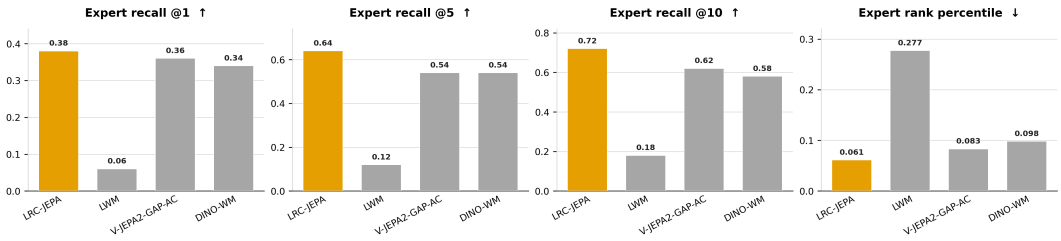



Figure 4: Expert action recall and ranking percentile of world models by CEM planning on Bridge-v2 offline dataset. Results averaged on 3 repetitions of 50 distinct scenes.

LRC-JEPA improves average success over parameter-matched LWM by 9pp across the four simulated environments (Fig. 3). It achieves the highest success on kinematics-intensive Push-T and Reacher and remains competitive on TwoRoom and Cube with pretrained DINO-WM, V-JEPA2-GAP-AC, and C-JEPA, despite training its encoder from scratch and planning by one compact state.

On Bridge-v2, LRC-JEPA leads all four offline metrics (Fig. 4), with recall@1/5/10 of $0.38/0.64/0.72$, versus $0.36/0.54/0.62$ for V-JEPA2-GAP-AC and $0.34/0.54/0.58$ for DINO-WM. Its active visual encoder has only $5.5$M parameters, compared with $303.9$M and $22.1$M, respectively (Table 1, with full planning time reported). Parameter-matched LWM reaches only $0.06/0.12/0.18$, supporting functional routing as the source of the gain rather than model scale.

Table 1: The number of active parameters of each world model encoder and the average full planning time over 50 different trajectories.

| **Model** | LRC-JEPA | LWM | V-JEPA2-GAP-AC | DINO-WM | EB-JEPA |
|---|---|---|---|---|---|
| **Num. of params.** | 5.5M | 5.5M | 303.9M | 22.1M | 1.1M |
| **Full planning time (s)** | 0.45 | 0.48 | 0.83 | 57.56 | 0.20 |

### 4.2 Physical Quantity Probing

**Evaluation Protocol.** We assess physical information accessible from LRC-JEPA's predictive latent $\mathbf{z}$ and context embedding $\mathbf{u}$ using ridge regression and MLP probes. Probes are trained on frozen, converged encoder outputs and evaluated on held-out data. Targets are simulator-provided physical quantities for simulated environments and pseudo-labels derived from SAM3 Carion et al. (2025) object masks and attributes for Bridge-v2. We evaluate LWM's single per-frame latent using the same protocol; fitting details appear in Appendix A.4.

**Comparison with baseline.** LRC-JEPA outperforms LWM in 28 of 32 Cube comparisons (Table 2), with the largest gains on kinematic and control-relevant quantities. For joint velocity, linear and MLP correlations rise from $0.111$ and $0.068$ to $0.392$ and $0.312$, respectively. Push-T and TwoRoom results appear in Appendix Tables 4 and 5.

Table 2: Physical quantity probes on Cube. Each entry reports MSE (↓) / Pearson correlation $r$ (↑). **Bold** marks the better predictive latent between LWM and LRC-JEPA; underlining marks the better representation between LRC-JEPA's $\mathbf{z}$ and $\mathbf{u}$, separately for each metric and probe type. EE: end-effector; quat.: quaternion.

| **Quantity** | **Linear** | | | **MLP** | | |
|---|---|---|---|---|---|---|
| | LWM | LRC-JEPA | | LWM | LRC-JEPA | |
| | $\mathbf{z}$ | $\mathbf{z}$ | $\mathbf{u}$ | $\mathbf{z}$ | $\mathbf{z}$ | $\mathbf{u}$ |
| Joint pos. | 0.293 / 0.698 | **0.274** / **0.723** | 0.572 / 0.500 | 0.311 / 0.692 | **0.305** / **0.696** | 0.444 / 0.653 |
| Joint vel. | 0.969 / 0.111 | **0.734** / **0.392** | 0.990 / 0.039 | 0.999 / 0.068 | **0.893** / **0.312** | 1.013 / 0.026 |
| EE pos. | 0.019 / 0.990 | **0.011** / **0.995** | 0.380 / 0.685 | 0.009 / 0.995 | **0.007** / **0.997** | 0.144 / 0.925 |
| EE yaw | 0.989 / 0.073 | **0.930** / **0.250** | 0.876 / 0.336 | **1.009** / 0.100 | 1.027 / **0.128** | 0.897 / 0.321 |
| Gripper | 0.077 / 0.960 | **0.061** / **0.969** | 0.193 / 0.897 | 0.038 / 0.981 | **0.035** / **0.982** | 0.084 / 0.957 |
| Block pos. | 0.007 / 0.996 | **0.006** / **0.997** | 0.120 / 0.933 | 0.006 / 0.997 | **0.005** / **0.998** | 0.017 / 0.991 |
| Block quat. | 0.678 / 0.053 | **0.653** / **0.164** | 0.672 / 0.073 | 0.694 / 0.048 | **0.692** / **0.063** | 0.694 / 0.094 |
| Block yaw | **1.010** / 0.056 | 1.012 / **0.060** | 0.971 / 0.189 | **1.046** / **0.066** | 1.052 / 0.065 | 0.968 / 0.230 |

**Information Allocation Across Streams.** Table 2 reveals the intended allocation: $\mathbf{z}$ better predicts joint position and velocity, end-effector and block positions, and gripper state, whereas $\mathbf{u}$ better predicts the more slowly varying end-effector and block yaw. Concatenation balances the two on most quantity probes (Appendix Table 6). These results support a bias toward control-relevant kinematics in $\mathbf{z}$ and persistent context in $\mathbf{u}$.

On Bridge-v2, SAM3-derived masks identify gripper, moving, static, and background objects for centroid and RGB probes. Predictive $\mathbf{z}$ better captures gripper and moving-object centroids, while context $\mathbf{u}$ generally better captures RGB values (Appendix Table 7), again separating motion-relevant information from persistent appearance.

### 4.3 Reconstruction Interventions

The training-only decoder makes information allocation interpretable. We reconstruct unaltered latent pairs, shuffle either stream within a trajectory, and replace $\mathbf{u}$ across trajectories (batch shuffling).

On Cube (Fig. 5), within-trajectory $\mathbf{u}$ shuffling causes little degradation (PSNR $29.26$ vs. $29.58$; SSIM $0.919$ vs. $0.923$), consistent with temporal invariance. Cross-trajectory $\mathbf{u}$ replacement

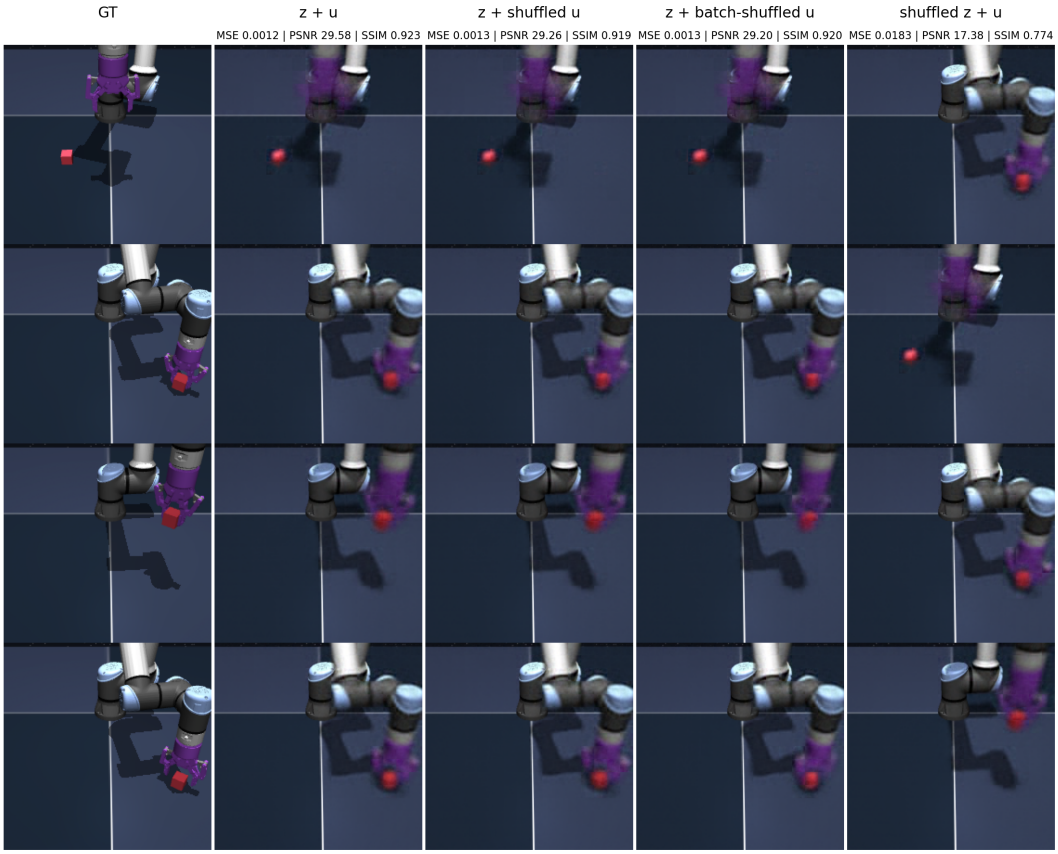



Figure 5: Reconstruction interventions on Cube environment. "Shuffled" denotes within-trajectory permutation, while "batch-shuffled" denotes replacement using context codes from other trajectories. Metrics calculated on the shown sample.

changes slowly varying gripper and block orientations, whereas shuffling $\mathbf{z}$ changes arm, gripper, and block locations and yields the lowest quality (PSNR $17.38$, SSIM $0.774$). These interventions corroborate the probes: $\mathbf{z}$ captures kinematics and $\mathbf{u}$ slowly varying context. Appendix Figs. 7 and 8 show further examples on Push-T and TwoRoom.

## 5 Ablation Study

Varying the number of residual-context queries $Q$ from 1 to 16 on Bridge-v2 primarily affects reconstruction: MSE improves monotonically, as do PSNR and SSIM (Fig. 6). Planning recall and $\mathbf{z}$ probes vary less consistently, although gripper-centroid probing improves with $Q$. Thus a small context bottleneck captures useful persistent information, while additional queries mainly improve reconstruction.

Furthermore, we train LRC-JEPA and its stop-gradient variant on the push-T and Bridge-v2 environments by applying a stop gradient operation on the residual connection from $z$ to the decoder. Table 3 shows that blocking this gradient reduces Push-T success from $96\%$ to $92\%$. The effect is substantially larger on Bridge-v2: recall@1/5/10 decreases from $0.38/0.64/0.72$ to $0.10/0.28/0.32$, while the rank percentile worsens from $0.061$ to $0.286$. Thus, affected by the entanglement of contextual nuisance, next latent prediction and SIGReg regularization alone do not recover the same planning-relevant representation, particularly in the visually diverse real-world setting. The residual connection and reconstruction gradients enforce separation of context from the predictive latent, achieving better planning and control. Appendix C provides the corresponding theoretical analysis.

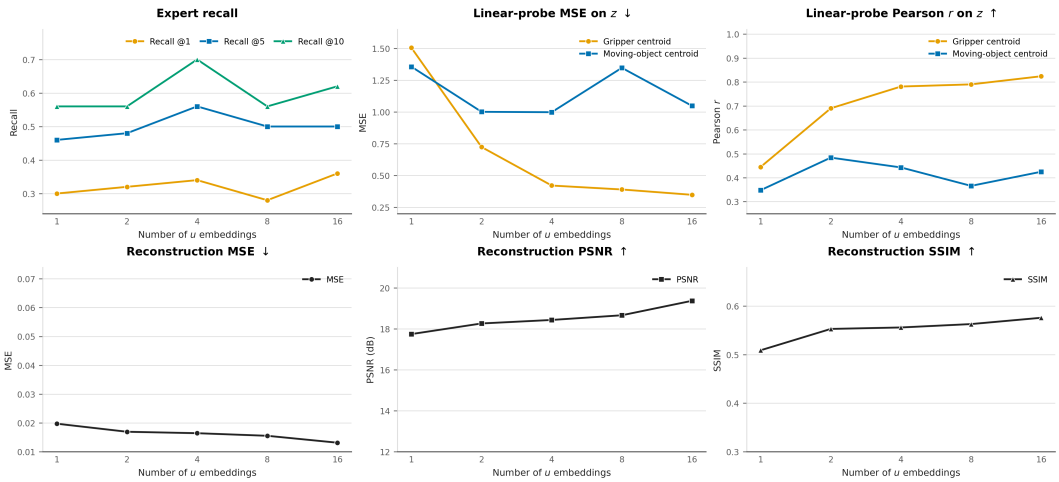



Figure 6: Effect of the number of residual-context queries $Q$ on Bridge-v2. We report expert-action recall, linear probes on $\mathbf{z}$, and reconstruction MSE, PSNR, and SSIM.

Table 3: Effect of blocking reconstruction gradients at $\mathbf{z}$. Push-T reports planning success in percent. Bridge-v2 reports expert-action recall and rank percentile (pct.).

| **Model** | **Push-T ↑** | **Bridge recall@1/5/10 ↑** | **Bridge rank pct. ↓** |
|---|---|---|---|
| LRC-JEPA | 96.0 | 0.38/0.64/0.72 | 0.061 |
| w/ stop-gradient | 92.0 | 0.10/0.28/0.32 | 0.286 |

## 6 Conclusion and Limitations

LRC-JEPA resolves a capacity-allocation problem in compact JEPA world models by routing action-conditioned dynamics through a predictive state $\mathbf{z}$ and persistent visual context through reconstruction-only embeddings $\mathbf{u}$. When competing with a substantially larger pretrained model (4× for DINO-WM), LRC-JEPA demonstrates the efficiency of single-state latent planning (120× acceleration) and simultaneously realizes 38% control performance improvement (in terms of expert action rank percentile) on the real-world Bridge-v2 dataset. Theory, physical-state probes, and reconstruction interventions consistently support the intended functional separation.

Nevertheless, the approach assumes that nuisance context is approximately stable over the training horizon; rapidly changing but uncontrollable or irrelevant factors may be routed imperfectly. The decoder and context regularizer also add training-time computation and introduce sensitivity to bottleneck capacity and loss weights, although neither is used for planning. Finally, real-world data is evaluated through offline action recovery rather than closed-loop robot execution. Testing longer horizons, dynamic visual nuisances, and real-robot control is therefore an important direction for future work.

## AI use statement

During the manuscript preparation, we used generative AI tools to assist with manuscript revision, review, and reference checking and formatting. We also used generative AI for code and documentation editing, feedback on experimental methodology and result interpretation. The authors are responsible for independently verifying AI-assisted text, code, citations, and mathematical claims and take full responsibility for the final manuscript and released artifacts.

### Ethics statement

This work studies representation learning for visual world models using simulated control tasks and offline robot data. The Bridge-v2 evaluation measures offline action ranking and does not establish the safety or reliability of deployment on physical robots. Applying these models to physical systems requires additional validation and appropriate operational safeguards. Use of the datasets, third-party software, and pretrained models is subject to their respective licenses and access conditions. The license for our released code and checkpoints specifies the permitted uses of those artifacts and does not supersede third-party terms.

### Reproducibility statement

Appendix A describes the model architectures, training settings, data preprocessing, and evaluation protocols. The anonymous resources linked in the introduction provide code for training, model predictive control, and probe evaluation, together with pretrained checkpoints and their configurations. The release includes a checkpoint manifest containing file sizes and SHA-256 checksums, and the main README reports the checkpoint metrics and explains how to run the evaluations.

## References

Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *Journal of Machine Learning Research*, 19(50):1–34, 2018a. URL https://jmlr.org/papers/v19/17-646.html.

Alessandro Achille and Stefano Soatto. Information dropout: Learning optimal representations through noisy computation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40 (12):2897–2905, 2018b. doi: 10.1109/TPAMI.2017.2784440. URL https://doi.org/10.1109/TPAMI.2017.2784440.

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations*, 2017. URL https://arxiv.org/abs/1612.00410.

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

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

Randall Balestriero and Yann LeCun. LeJEPA: Provable and scalable self-supervised learning without the heuristics. *arXiv preprint arXiv:2511.08544*, 2025. URL https://arxiv.org/abs/2511.08544.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=xm6YD62D1Ub.

Adrien Bardes, Jean Ponce, and Yann LeCun. MC-JEPA: A joint-embedding predictive architecture for self-supervised learning of motion and content features. *arXiv preprint arXiv:2307.12698*, 2023. URL https://arxiv.org/abs/2307.12698.

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

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025. URL https://arxiv.org/abs/2511.16719.

Ricky T. Q. Chen, Xuechen Li, Roger Grosse, and David Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://arxiv.org/abs/1802.04942.

Jingyi Cui, Qi Zhang, Hongwei Wen, and Yisen Wang. A generalization theory for JEPA-based world models. *arXiv preprint arXiv:2606.27014*, 2026. URL https://arxiv.org/abs/2606.27014.

Emily L. Denton and Vighnesh Birodkar. Unsupervised learning of disentangled representations from video. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/2d2ca7eedf739ef4c3800713ec482e1a-Abstract.html.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2010.11929.

Frederik Ebert, Chelsea Finn, Sudeep Dasari, Annie Xie, Alex Lee, and Sergey Levine. Visual foresight: Model-based deep reinforcement learning for vision-based robotic control. *arXiv preprint arXiv:1812.00568*, 2018. URL https://arxiv.org/abs/1812.00568.

Chelsea Finn, Ian Goodfellow, and Sergey Levine. Unsupervised learning for physical interaction through video prediction. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/d9d4f495e875a2e075a1a4a6e1b9770f-Abstract.html.

Letian Fu, Long Lian, Renhao Wang, Baifeng Shi, Xudong Wang, Adam Yala, Trevor Darrell, Alexei A. Efros, and Ken Goldberg. Rethinking patch dependence for masked autoencoders. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=JT2KMuo2BV.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018. URL https://arxiv.org/abs/1803.10122.

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*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2555–2565. PMLR, 2019. URL https://proceedings.mlr.press/v97/hafner19a.html.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640:647–653, 2025. doi: 10.1038/s41586-025-08744-2. URL https://doi.org/10.1038/s41586-025-08744-2.

Nicklas Hansen, Hao Su, and Xiaolong Wang. TD-MPC2: Scalable, robust world models for continuous control. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Oxh5CstDJU.

Nicklas A. Hansen, Hao Su, and Xiaolong Wang. Temporal difference learning for model predictive control. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 8387–8406. PMLR, 2022. URL https://proceedings.mlr.press/v162/hansen22a.html.

Jun-Ting Hsieh, Bingbin Liu, De-An Huang, Li Fei-Fei, and Juan Carlos Niebles. Learning to decompose and disentangle representations for video prediction. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/496e05e1aea0a9c4655800e8a7b9ea28-Abstract.html.

David Klindt, Yann LeCun, and Randall Balestriero. When does LeJEPA learn a world model? *arXiv preprint arXiv:2605.26379*, 2026. URL https://arxiv.org/abs/2605.26379.

Chang Liu, Fei Suo, Yanzhou Jin, Zeyu Ping, Yusuke Iwasawa, Yutaka Matsuo, and Yaonan Zhu. Temporally centered SIGReg improves LeWorldModel representations for robot policy learning. *arXiv preprint arXiv:2607.26924*, 2026. URL https://arxiv.org/abs/2607.26924.

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*, 2026. URL https://arxiv.org/abs/2603.19312.

Heejeong Nam, Quentin Le Lidec, Lucas Maes, Yann LeCun, and Randall Balestriero. Causal-JEPA: Learning world models through object-level latent masking. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.11389.

NVIDIA. Cosmos world foundation model platform for physical AI. *arXiv preprint arXiv:2501.03575*, 2025. URL https://arxiv.org/abs/2501.03575.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=a68SUt6zFt.

Minting Pan, Xiangming Zhu, Yunbo Wang, and Xiaokang Yang. Iso-Dream: Isolating and leveraging noncontrollable visual dynamics in world models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 23178–23191, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/9316769afaaeeaad42a9e3633b14e801-Abstract-Conference.html.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4172–4182, 2023. URL https://arxiv.org/abs/2212.09748.

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. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/3e7cf447f21cd11c846463affefce665-Abstract-Conference.html.

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

Basile Terver, Randall Balestriero, Megi Dervishi, David Fan, Quentin Garrido, Tushar Nagarajan, Koustuv Sinha, Wancong Zhang, Mike Rabbat, Yann LeCun, and Amir Bar. A lightweight library for energy-based joint-embedding predictive architectures. *arXiv preprint arXiv:2602.03604*, 2026. URL https://arxiv.org/abs/2602.03604.

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. URL https://arxiv.org/abs/2605.31111.

Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. MoCoGAN: Decomposing motion and content for video generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1526–1535, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Tulyakov_MoCoGAN_Decomposing_Motion_CVPR_2018_paper.html.

Homer Rich Walke, Kevin Black, Tony Z. Zhao, Quan Vuong, Chongyi Zheng, Philippe Hansen-Estruch, Andre Wang He, Vivek Myers, Moo Jin Kim, Max Du, Abraham Lee, Kuan Fang, Chelsea Finn, and Sergey Levine. BridgeData V2: A dataset for robot learning at scale. In *Proceedings of the 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 1723–1736. PMLR, 2023. URL https://proceedings.mlr.press/v229/walke23a.html.

Haixu Wu, Zhiyu Yao, Jianmin Wang, and Mingsheng Long. MotionRNN: A flexible model for video prediction with spacetime-varying motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15435–15444, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Wu_MotionRNN_A_Flexible_Model_for_Video_Prediction_With_Spacetime-Varying_Motions_CVPR_2021_paper.html.

Amy Zhang, Rowan Thomas McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. Learning invariant representations for reinforcement learning without reconstruction. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=-2FCwDKRREu.

Xiangteng Zhang, Yang Guan, Bo Zhang, Hongyang Li, Ya-Qin Zhang, and Shengbo Eben Li. On the identifiability of controlled world models. *arXiv preprint arXiv:2607.22430*, 2026. URL https://arxiv.org/abs/2607.22430.

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. URL https://proceedings.mlr.press/v267/zhou25t.html.

# A Implementation Details

## A.1 Neural Networks

**LRC-JEPA architecture.** Our encoder is a ViT-Tiny trained from scratch with 12 layers, 3 attention heads, hidden dimension 192, and patch size 14. At $224 \times 224$ resolution, it produces 256 patch tokens and one class token. Both $d_z$ and $d_u$ are 192. The number of residual-context queries $Q$ is varied in ablations. The predictor has 6 causal Transformer layers, 16 attention heads, MLP width 2048, and dropout 0.1. The reconstruction decoder has 3 cross-attention layers, hidden dimension 256, 8 heads, and MLP width 1024. The resulting Bridge-v2 instantiation contains approximately 21.1 million parameters including the decoder. However, at planning time, only the predictive state $\mathbf{z}$, action encoder, and dynamics predictor are loaded, resulting in 5.5M active encoder parameters.

**V-JEPA2-GAP-AC baseline.** V-JEPA2-GAP-AC is our compact action-conditioned realization rather than the standard V-JEPA2-AC architecture. We freeze a pretrained V-JEPA2 ViT-L/16 encoder and process images at $256 \times 256$ resolution. The resulting 256 spatial patch representations are globally averaged to obtain one 1024-dimensional state vector per observation. We then train an action encoder and a six-layer causal action-conditioned predictor, using the same three-context/one-target temporal organization as LRC-JEPA. The predictor uses an internal width of 192, 16 attention heads, and MLP width 2048, and is trained for 10 epochs with batch size 192 and learning rate $5 \times 10^{-5}$. This baseline therefore tests whether a strong pretrained video representation remains effective when converted into the same compact global-state planning interface used by LRC-JEPA.

### A.2 Details on Capacity–complexity Stress Tests for Compact JEPA planning

**PCA compression.** For each checkpoint, we independently fit PCA to 1,000 clean predictive latents $z$ sampled without replacement from the TwoRoom evaluation dataset. At evaluation time, every encoded and autoregressively predicted latent is projected onto the leading $k$ components as $\Pi_k(\mathbf{z}) = \boldsymbol{\mu} + \mathbf{U}_k\mathbf{U}_k^\top(\mathbf{z} - \boldsymbol{\mu})$, ensuring that current, goal, and predicted representations occupy the same subspace. We evaluate $k \in \{32, 48, 64, 96, 128, 192\}$, where $k = 192$ is an exact no-op, on 50 trajectories with three evaluation seeds. The PCA basis is fixed across ranks and repetitions. Logistic trends are fitted to the binomial success counts using $p(k) = c\,\sigma[\beta(\log_2 k - \log_2 \kappa)]$; shaded regions show pointwise 95% intervals from 500 parametric-bootstrap refits. This intervention restricts the effective rank of the latent while retaining its 192-dimensional tensor shape, and therefore measures informational capacity rather than the computational cost of a natively lower-dimensional model.

**Static-pattern perturbation.** For each evaluation trajectory, we sample a trajectory-persistent $8 \times 8$ binary grayscale pattern, interpolate it by nearest-neighbor method to $224 \times 224$, and broadcast it across the RGB channels. The pattern differs across trajectories but remains fixed over time and is independent of the agent's actions. We blend the same pattern into both current and goal observations according to $\widetilde{\mathbf{x}}_\alpha = (1 - \alpha)\mathbf{x} + \alpha\mathbf{b}$, with $\alpha \in \{0, 0.05, 0.10, 0.15, 0.20, 0.30, 0.40, 0.50\}$. All opacity conditions use the same 50 trajectories and trajectory-specific planning seeds, with $\alpha = 0$ serving as the clean reference. We report the raw success rates with adjacent opacity conditions connected for readability.

### A.3 Details on Physical Quantity Probes

**Physical quantity probes.** All probes are trained on frozen representations extracted from individual frames; no encoder parameters are updated. We evaluate the predictive state $\mathbf{z}$, the mean-pooled residual context $\mathbf{u}$, and their concatenation $\mathbf{z} \oplus \mathbf{u}$. For the simulated environments, we randomly sample $10{,}000$ frames and use an $80/20$ train–test split. For Bridge-v2, object properties are obtained from the extracted segmentation masks; one valid moving, static, and background object is selected per frame, retaining $1{,}862$ of $2{,}000$ labeled frames, which are split $80/20$ by episode to prevent trajectory leakage. Each target and representation is evaluated independently using ridge regression with regularization $10^{-3}$ and a two-layer MLP with hidden widths $(256, 256)$, ReLU activations, and early stopping. Features and regression targets are standardized for probe fitting. We report target-normalized mean-squared error and Pearson correlation, averaging over dimensions for vector-valued quantities.

**Speed evaluation.** Since our eventual target is closed-loop deployment outside datacenter-scale infrastructure, inference-speed measurements are performed on a single consumer-class NVIDIA GeForce RTX 5090 GPU (32 GB). We use the same CEM population, elite count, iteration count, and horizon for all methods. Full planning time sums every CEM solve required by a trajectory and includes online current/goal image encoding, but excludes checkpoint loading, environment stepping, video generation, and result serialization. Reported values are averaged over 50 trajectories.

### A.4 Implementation Details

**Latent prediction and SIGReg.** For the implemented one-step objective, the causal predictor consumes $\mathbf{z}_{1:T}$ and is aligned with targets $\mathbf{z}_{2:T+1}$ according to Eq. 11. The target embeddings are not stop-gradient targets, and we do not use a target encoder or exponential-moving-average encoder. Instead, following LeJEPA and LWM Balestriero & LeCun (2025); Maes et al. (2026), we apply Sketched Isotropic Gaussian Regularization (SIGReg) Balestriero & LeCun (2025) to prevent collapse and stabilize training.

SIGReg is applied independently to the batch of latent representations at each time step and averaged over time. It matches random one-dimensional projections of the latent distribution to a standard Gaussian using the Epps–Pulley characteristic-function statistic. We use 1,024 random projections and 17 trapezoidal quadrature points over $[-5, 5]$.

**Temporal context regularization.** We use a temporal adaptation of VICReg Bardes et al. (2022) to encourage the residual context to encode static or slowly varying information complementary to

the predictive latent. We first flatten the $Q$ context tokens:

$$\mathbf{v}_{bt} = \mathrm{vec}(\mathbf{u}_{bt}) \in \mathbb{R}^{d_u Q}, \qquad \overline{\mathbf{v}}_b = \frac{1}{T}\sum_{t=1}^{T} \mathbf{v}_{bt}. \tag{14}$$

The invariance term suppresses within-clip temporal variation:

$$\mathcal{L}_{\mathrm{inv}} = \frac{1}{BTd_uQ}\sum_{b,t} |\mathbf{v}_{bt} - \overline{\mathbf{v}}_b|_2^2 . \tag{15}$$

To prevent a constant context code, the variance and covariance terms are computed across clip-level means:

$$\mathcal{L}_{\mathrm{var}} = \frac{1}{d_uQ}\sum_{j} \max\left(0, , \gamma - \sqrt{\mathrm{Var}_b[\overline{v}_{b,j}] + \epsilon}\right), \tag{16}$$

$$\mathcal{L}_{\mathrm{cov}} = \frac{1}{d_uQ}\sum_{j\neq k} C_{jk}^{,2}, \qquad C = \mathrm{Cov}_b(\overline{\mathbf{v}}_b). \tag{17}$$

The complete context regularizer is

$$\mathcal{L}_u = \alpha\mathcal{L}_{\mathrm{inv}} + \beta\mathcal{L}_{\mathrm{var}} + \delta\mathcal{L}_{\mathrm{cov}}, \tag{18}$$

with $(\alpha, \beta, \delta) = (25, 25, 1)$, $\gamma = 1$, and $\epsilon = 10^{-4}$. Thus, $\mathcal{L}_{\mathrm{inv}}$ encourages stability within a trajectory, whereas $\mathcal{L}_{\mathrm{var}}$ and $\mathcal{L}_{\mathrm{cov}}$ preserve variation across trajectories and prevent dimensional collapse.

**Patch reconstruction.** The reconstruction loss in Eq. 12 is averaged over the batch, time, patch, and pixel dimensions. In the full model, neither $\mathbf{z}_t$ nor $\mathbf{u}_t$ is detached from this loss. Reconstruction therefore directly shapes the visual encoder and both latent streams, rather than serving only as a visualization objective. The loss weights are $(\lambda_z, \lambda_u, \lambda_{\mathrm{rec}}) = (0.05, 0.05, 0.2)$.

**LRC-JEPA training.** Training clips contain three context observations and one prediction target. Consecutive visual observations are separated by five low-level environment steps, whose actions are concatenated and z-score normalized. Frames are normalized using ImageNet statistics. The main simulation models are trained for 10 epochs with batch size 128 and learning rate $5 \times 10^{-5}$; Bridge-v2 is trained for 20 epochs with batch size 192 and learning rate $5 \times 10^{-4}$. $90/10$ training–validation split is used for all datasets.

Optimization uses AdamW with weight decay $10^{-3}$, linear warm-up followed by cosine annealing, bfloat16 mixed-precision training, and gradient clipping at 1.0. The decoder has a separate optimizer parameter group, allowing its learning rate to be varied independently; the main setting uses the same learning rate for the decoder and the remaining network.

**Latent Planning.** After training, the context encoder and reconstruction decoder are not needed for control. Given an observed history and a candidate action sequence, the predictor is rolled out autoregressively in the $\mathbf{z}$ space. For an image goal $\mathbf{i}_g$, the terminal cost is

$$\mathcal{C}\left(\mathbf{a}_{t:t+K-1}\right) = \|\widehat{\mathbf{z}}_{t+K} - \mathbf{z}_g\|_2^2 , \qquad \mathbf{z}_g = g_z(\mathrm{Enc}_\theta(\mathbf{i}_g)_{\mathrm{cls}}) . \tag{19}$$

Candidate actions are optimized using the Cross-Entropy Method (CEM). In closed-loop control, only the first part of the optimized sequence is executed before obtaining a new observation and replanning. Because planning operates on a single $d_z$-dimensional state per frame rather than on hundreds of spatial tokens, it retains the computational efficiency of compact JEPA-style world models.

**Planning protocol.** For online control, we perform latent-space MPC using CEM. The planning horizon contains five action tokens, with each token representing five low-level actions. CEM evaluates 300 candidate sequences for 30 optimization iterations and retains the best 30 candidates as elites. Candidate sequences are scored by the squared distance between the terminal predicted state $\mathbf{z}$ and the encoded goal state. The controller executes five action tokens before replanning, with

an overall budget of 50 low-level environment steps. Current and goal observations are encoded once per MPC solve and their representations are reused across CEM iterations, since they are independent of the candidate actions. Bridge-v2 uses the same latent planning objective in an offline start–goal evaluation. The demonstrated action sequence is ranked against 300 random candidate sequences using the predicted terminal latent distance to the goal, and recall at ranks 1, 5, and 10 and normalized expert rank are reported.

### A.5 Relation to Prior Designs

LRC-JEPA is most directly related to LWM Maes et al. (2026), from which it retains compact action-conditioned prediction and SIGReg-based end-to-end training. LWM, however, represents each frame using only one predictive latent and entangles control-relevant and residual contextual information. DINO-WM Zhou et al. (2025) instead predicts frozen DINOv2 spatial patch features, making its planning-time dynamics substantially more expensive than global-latent prediction. Standard V-JEPA2-AC Assran et al. (2025) similarly learns a large action-conditioned predictor over frozen, pretrained spatial feature maps, whereas our V-JEPA2-GAP-AC realization spatially pools these features before learning the action-conditioned dynamics.

LRC-JEPA differs from these approaches by training a compact visual state from scratch while retaining spatial and appearance information in a separate, reconstruction-only context bottleneck. MC-JEPA Bardes et al. (2023) jointly learns motion and content through shared representation and optical-flow objectives, whereas LRC-JEPA assigns the two streams asymmetric world-model roles: only $\mathbf{z}$ is predicted and planned through, while $\mathbf{u}$ supplies temporally regularized residual context. The reconstruction decoder is related to cross-attention masked decoders Fu et al. (2025), but queries the compact $[\mathbf{z}_t; \mathbf{u}_t^{\mathrm{dec}}]$ representation rather than the complete set of visible patch tokens.

## B Theoretical Analysis for LWM

### B.1 Definitions and assumptions

**Definition 1** (Optimal-representation criteria, Achille & Soatto, 2018a)**.** *Let $X$ be the input, $Y$ the task, $N$ a nuisance, and $Z$ a stochastic representation of $X$ with coordinates $Z_1, \ldots, Z_d$. Then $Z$ is* sufficient *for $Y$ if $I(Y;Z) = I(Y;X)$;* minimal *if it minimizes $I(Z;X)$ among sufficient representations;* invariant *to $N$ if $I(Z;N) = 0$; and* disentangled *if, among minimal sufficient representations, it minimizes*

$$\mathrm{TC}(Z) \;:=\; \sum_{i=1}^{d} H(Z_i) - H(Z) \;=\; D_{\mathrm{KL}}\Big(q(\boldsymbol{z}) \,\Big\|\, \textstyle\prod_{i=1}^{d} q(z_i)\Big) \;\geq\; 0, \tag{20}$$

*which is zero if $Z_1, \ldots, Z_d$ are mutually independent.*

**Remark 1** (Tightness residual in Theorem 1(a))**.** *The nuisance attains the bound up to a residual $\epsilon := I(Z;Y \mid N) - I(X;Y) \in [0, H(Y \mid X)]$, which is zero whenever $Y$ is a deterministic function of $X$ (Achille & Soatto, 2018a, Prop. 3.1).*

**Remark 2** (The task variable)**.** *LWM has no external label. We take the training task to be next-latent prediction, $Y = \boldsymbol{z}_{t+1}$, so sufficiency is the fixed-point property that $\boldsymbol{z}_t$ retains what is needed to predict $\boldsymbol{z}_{t+1}$ given $\boldsymbol{a}_t$. Squared error $\mathcal{L}_{\mathrm{pred}}$ is a surrogate for this property, not a proof of information-theoretic sufficiency. The minimality and* TC *are defined on the marginal of $Z$.*

**Assumption 1** (Noisy surrogate encoder)**.** *If the encoder $\mathrm{enc}_\theta$ is deterministic, $I(Z;X)$ as written is infinite. We analyze the standard noisy surrogate $\tilde{\boldsymbol{z}} = \mathrm{enc}_\theta(\boldsymbol{o}) + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_d)$, at fixed $\sigma > 0$, and recover the deterministic map as $\sigma \to 0^+$ (Alemi et al., 2017; Achille & Soatto, 2018b). Then $I(Z;X) = H(Z) - H(\varepsilon)$. Matching $q(\boldsymbol{z})$ to $\mathcal{N}(0, I_d)$ fixes the scalar mutual information $I(Z;X)$ at its Gaussian value, but leaves unspecified how this information budget is allocated across coordinates of $X$.*

**Assumption 2** (Single-factor Gaussian controlled world)**.** *Let $s \in \mathbb{R}^n$ be the unobserved state and $X = g(s)$ with $g$ injective on the support of $s$. Proposition 1(b) invokes the setting in which action-conditioned latent prediction plus an isotropic-Gaussian constraint identifies $s$ up to $O(n)$ (Klindt et al., 2026; Zhang et al., 2026): $s \sim \mathcal{N}(0, I_n)$; stationary linear-Gaussian dynamics $s_{t+1} = As_t +$*

$B\boldsymbol{a}_t + \xi_t$ *with* $s_{t+1} \sim \mathcal{N}(0, I_n)$*; jointly Gaussian* $(s_t, \boldsymbol{a}_t)$*; spectral separation of the predictable signal; and non-degenerate conditional action variation given* $s_t$*.*

**Assumption 3** (Population optimality and realizability)**.** *The deterministic predictive encoder has dimension* $d_z = d_s$*. The encoder and predictor jointly attain a global minimum of the population one-step LWM objective at which the ideal population SIGReg loss vanishes. The encoder class contains a map satisfying* $f(g(s)) = Qs$ *for some* $Q \in O(d_s)$*. The predictor is continuous, and its class can realize the conditional mean* $\mathbb{E}[Z_{t+1} \mid Z_t, a_t]$ *for every admissible encoder.*

### B.2 The decomposition lemma

**Lemma 1** (Exact information decomposition; Chen et al., 2018, cf. Achille & Soatto, 2018a;b)**.** *Let* $k$ *index a training set of size* $K$ *with* $p(k) = 1/K$*,* $q(\boldsymbol{z} \mid k)$ *the (surrogate) encoder,* $q(\boldsymbol{z}) = \sum_k q(\boldsymbol{z} \mid k)p(k)$ *the aggregate, and* $p(\boldsymbol{z}) = \prod_i p(z_i)$ *a factorized prior. Then*

$$\mathbb{E}_{p(k)}\big[D_{\mathrm{KL}}(q(\boldsymbol{z} \mid k) \,\|\, p(\boldsymbol{z}))\big] = \underbrace{I_q(Z;k)}_{\text{(i) rate / minimality}} + \underbrace{\mathrm{TC}(Z)}_{\text{(ii) total correlation}} + \underbrace{\sum_{i=1}^{d} D_{\mathrm{KL}}\big(q(z_i) \,\|\, p(z_i)\big)}_{\text{(iii) coordinate-wise matching}}, \tag{21}$$

*an exact split into three non-negative terms. Term* (i) *is the index-code mutual information, the empirical counterpart of* $I(Z;X)$*;* (ii) *is equation 20;* (iii) *depends only on coordinate marginals. The three are separately controllable:* (iii) *can vanish while* (i) *takes any compatible value and while* (ii) *remains positive (correlated or copula-dependent coordinates with Gaussian margins).*

*Proof.* Write $q(\boldsymbol{z}, k) = q(\boldsymbol{z} \mid k)p(k)$ and $q(\boldsymbol{z}) = \sum_k q(\boldsymbol{z} \mid k)p(k)$. Insert $q(\boldsymbol{z})$ and $\prod_i q(z_i)$ inside the logarithm:

$$\begin{aligned}
\mathbb{E}_{p(k)}\big[D_{\mathrm{KL}}(q(\boldsymbol{z} \mid k) \,\|\, p(\boldsymbol{z}))\big] &= \mathbb{E}_{q(\boldsymbol{z},k)}\left[\log \frac{q(\boldsymbol{z} \mid k)}{p(\boldsymbol{z})}\right] \\
&= \mathbb{E}_{q(\boldsymbol{z},k)}\left[\log \frac{q(\boldsymbol{z},k)}{q(\boldsymbol{z})p(k)} + \log \frac{q(\boldsymbol{z})}{\prod_i q(z_i)} + \log \frac{\prod_i q(z_i)}{\prod_i p(z_i)}\right] \\
&= \underbrace{D_{\mathrm{KL}}\big(q(\boldsymbol{z},k) \,\|\, q(\boldsymbol{z})p(k)\big)}_{\text{(i)}} + \underbrace{D_{\mathrm{KL}}\Big(q(\boldsymbol{z}) \,\Big\|\, \textstyle\prod_i q(z_i)\Big)}_{\text{(ii)}} \\
&\quad + \underbrace{\sum_i D_{\mathrm{KL}}\big(q(z_i) \,\|\, p(z_i)\big)}_{\text{(iii)}}.
\end{aligned}$$

The factors telescope, so the identity is exact; each summand is a KL, hence non-negative. Over sufficiently expressive stochastic encoder famillies, separate controllability follows because the terms are functionals of $q(\boldsymbol{z}, k)$, of the dependence in $q(\boldsymbol{z})$, and of the coordinate marginals, respectively. □

**Remark 3** (What SIGReg targets)**.** *LWM does* not *optimize the left-hand side of equation 21. That KL includes the rate and is the factorized-prior regularizer of Achille & Soatto (2018b). SIGReg is a statistic of the aggregate* $q(\boldsymbol{z})$*: Epps–Pulley tests of random projections against* $\mathcal{N}(0,1)$ *(Balestriero & LeCun, 2025; Maes et al., 2026). Its population minimizer* $q(\boldsymbol{z}) = \mathcal{N}(0, I_d)$ *makes* (ii) *and* (iii) *vanish together and leaves* (i) *unconstrained.*

### B.3 Proof of Proposition 1(Sufficiency)

For LWM, SIGReg averages the Epps–Pulley normality statistic over random one-dimensional projections. The SIGReg objective is zero when almost every projection of $Z$ is standard normal; continuity of the characteristic function extends this to all directions, and the Cram'er–Wold theorem then implies $q(\boldsymbol{z}) = \mathcal{N}(0, I_d)$. Hence the learned latent variables are mutually independent in any orthonormal basis. Under Assumptions 2 and 3, the deterministic predictive representation satisfies $Z_t = Qs_t$ almost surely for some orthogonal transformation $Q \in O(d_s)$. Consequently, $s_t = Q^\top Z_t$, establishing state sufficiency. □

### B.4 Proof of Proposition 1(Minimality)

The objective $\mathcal{L}_{\text{LeWM}}$ targets unsupervised latent representation learning through $\mathcal{L}_{\text{pred}}$ and SIGReg. While $\mathcal{L}_{\text{pred}}$ encourages retention of transition-relevant information, SIGReg depends only on the aggregate $q(\boldsymbol{z})$ and is therefore insensitive to conditional changes that alter $I(Z;X)$; its Gaussian target primarily prevents collapse. Hence, $\mathcal{L}_{\text{LeWM}}$ neither explicitly targets nor enforces minimality. □

## C Theoretical Analysis for Proposed Method

### C.1 Optimality of LRC-JEPA

**Assumption 4** (Two-factor Gaussian controlled world)**.** *Observations factor as $X_t = g(s_t, v)$ with $g$ injective on the joint support. The dynamic state $s_t \in \mathbb{R}^{d_s}$ follows linear-Gaussian controlled dynamics, with $s_t \sim \mathcal{N}(0, I_{d_s})$, spectral separation such that the predictable signal spans $s$, and action excitation. The appearance factor $v \in \mathbb{R}^{d_v}$ is constant within each trajectory, independent of $(s_{1:T}, \boldsymbol{a}_{1:T})$, and non-degenerate across trajectories; $V$ is a nuisance for the prediction task in the sense of Theorem 1, $I(Y;V) = 0$.*

**Assumption 5** (Global minimizer of $\mathcal{L}_{LRC}$)**.** *Suppose the representation rate is minimum at convergence, enforced by architectural capacity controls $d_z = d_s, Qd_u = d_v$. At the global optimum of $\mathcal{L}_{LRC}$, $\mathcal{L}_u = 0$ and the context embedding $u_t$ is constant within a trajectory while retaining cross-trajectory variance (the temporal-VICReg population minimizer).*

With above anlaysis, we are ready to prove the our key theorem:

*Proof of Theorem 2.* We first show joint realizability of all loss terms, then use the context and reconstruction constraints to separate nuisance and state information, and finally invoke SIGReg and identifiability to establish invariance, minimality, coordinate disentanglement, and state recovery.

Consider $\boldsymbol{z}_t = Qs_t$ and $\boldsymbol{u}_t = \rho(v)$, where $Q \in O(n)$ and $\rho$ is injective and realizable by the context head. With predictor $\widehat{\boldsymbol{z}}_{t+1} = QAQ^\top \boldsymbol{z}_t + QB\boldsymbol{a}_t$ and a decoder recovering $(s_t, v)$ from $(\boldsymbol{z}_t, \boldsymbol{u}_t)$ before applying $g$, the objective simultaneously attains the minimum prediction error for $\mathcal{L}_{\text{pred}}$, $\mathcal{L}_{\text{SIG}} = 0$, the minimum of $\mathcal{L}_u$, and $\mathcal{L}_{\text{rec}} = 0$. Hence every global minimizer attains the minimum of each term individually.

**Step 1: Invariance restricts $U_t$ to shared context.** In particular, $\mathcal{L}_{\text{SIG}} = 0$ implies $q(\boldsymbol{z}) = \mathcal{N}(0, I_n)$ and $\text{TC}(Z) = 0$ by Appendix B.3, *proving (iv) disentanglement in the Theorem 2*. Minimizing $\mathcal{L}_u$ makes $\boldsymbol{u}_t$ a.s. constant along each trajectory. Conditional on almost every context $v$, the stationary state process is ergodic, $\omega(g(\cdot, v))$ is a.e. constant in $s$; thus $\boldsymbol{u}_t = \rho(v)$.

**Step 2: Compactness makes sufficiency.** Since $U$ is a deterministic function of $V$,

$$H(U) = I(U;V) = H(V) - H(V \mid U). \tag{22}$$

Perfect reconstruction gives

$$H(V, S_t \mid U, Z_t) = 0. \tag{23}$$

Therefore,

$$\begin{aligned} H(V, S_t \mid U) &= I(V, S_t; Z_t \mid U) && (24)\\ &\le H(Z_t \mid U) && (25)\\ &\le H(Z_t). && (26) \end{aligned}$$

Because $S_t$ is independent of $(V, U)$,

$$H(V, S_t \mid U) = H(V \mid U) + H(S_t). \tag{27}$$

Thus,

$$H(Z_t) \ge H(V \mid U) + H(S_t). \tag{28}$$

Combining equation 22 and equation 28 yields

$$\mathcal{R} = H(U) + \sum_{t=1}^{T} H(Z_t) \tag{29}$$

$$\geq H(V) + \sum_{t=1}^{T} H(S_t) + (T-1)H(V \mid U). \tag{30}$$

Since $T > 1$, every minimum-rate solution of equation 30 must satisfy

$$H(V \mid U) = 0. \tag{31}$$

Hence $U$ is sufficient for $V$.

Exact reconstruction then gives $D_\psi(\boldsymbol{z}_t, \rho(v)) = X_t$ a.s., and injectivity of $g$ implies $s_t = m(\boldsymbol{z}_t, \rho(v))$ for some measurable $m$. Therefore the state must be recoverable from $\boldsymbol{z}_t$ once $v$ is fixed. Since $U_t$ is sufficient for $V$, $Z_t$ must be sufficient for $S_t$ conditioned on $U_t$, *proving (i) sufficiency in the Theorem 2*.

**Step 3: Sufficiency leads to minimality and invariance** Combining this routing constraint with $s \perp v$, $s \sim \mathcal{N}(0, I_n)$, $q(\boldsymbol{z}) = \mathcal{N}(0, I_n)$, and Assumption 1 yields $I(Z; V) = 0$, *proving (iii) invariance in the Theorem 2*. Consequently, $\boldsymbol{z}_t$ is a.e. a function of $s_t$ alone and $I(Z; X) = I(Z; S)$. Since the predictable signal spans $s$ by Assumption 5, every prediction-sufficient representation $Z'$ satisfies $I(Z'; X) \geq I_\sigma(S; X)$, while $Z$ attains this lower bound; hence $Z$ is minimal, *proving (ii) minimality in the Theorem 2*.

In fact, $\boldsymbol{z}$ is an isotropic Gaussian representation of $s$ that minimizes next-latent MSE under the assumed linear-Gaussian controlled dynamics. By Klindt et al. (2026) and Zhang et al. (2026), $\boldsymbol{z}_t$ is equivalent to $\mathbf{s}_t$ up to an orthogonal transformation. □

## D Additional Results

We assess physical information accessible from LRC-JEPA's and LWM's predictive latent $\boldsymbol{z}$ using ridge regression and MLP probes. Probes are trained on frozen, converged encoder outputs and evaluated on held-out data. More details on probe fitting are provided in Appendix A.4 . Table 4 and 5 report probing results on the two simulated environments, Push-T and TwoRoom, where the ground truth quantities are provided by the simulator. Table 6 shows probing results with the concatenation of the two embeddings ($\boldsymbol{z}$ and $\boldsymbol{u}$) on Cube environment. Table 7 and 8 further analyze LRC-JEPA's predictive latents and context embeddings on the real-world Bridge-v2 dataset, where $Q = 2$ embeddings are used for context representation.

Table 4: Physical quantity probes on the simulated Push-T environment comparing LRC-JEPA and LWM. Each entry reports the mean-squared error (MSE; lower is better) and the Pearson correlation coefficient ($r$; higher is better). Both probe heads are fitted using the latent representation $z$.

| **Quantity** | **Model** | $z$ **Linear** | | $z$ **MLP** | |
|---|---|---|---|---|---|
| | | **MSE** ↓ | $r$ ↑ | **MSE** ↓ | $r$ ↑ |
| Agent position | LRC-JEPA | 0.066 | 0.967 | 0.022 | 0.989 |
| | LWM | 0.044 | 0.978 | 0.018 | 0.991 |
| Block position | LRC-JEPA | 0.021 | 0.989 | 0.005 | 0.997 |
| | LWM | 0.021 | 0.990 | 0.006 | 0.997 |
| Block angle | LRC-JEPA | 0.175 | 0.905 | 0.049 | 0.974 |
| | LWM | 0.184 | 0.900 | 0.055 | 0.971 |

Table 5: Physical quantity probes on the simulated TwoRoom environment comparing LRC-JEPA and LWM. Each entry reports the mean-squared error (MSE; lower is better) and the Pearson correlation coefficient ($r$; higher is better). Both probe heads are fitted using the latent representation $z$.

| **Quantity** | **Model** | $z$ **Linear** | | $z$ **MLP** | |
|---|---|---|---|---|---|
| | | **MSE** $\downarrow$ | $r \uparrow$ | **MSE** $\downarrow$ | $r \uparrow$ |
| Agent position | LRC-JEPA | 0.002 | 0.999 | 0.000 | 1.000 |
| | LWM | 0.008 | 0.996 | 0.001 | 1.000 |

Table 6: Physical quantity probes using the concatenated $\boldsymbol{z} \oplus \boldsymbol{u}$ representation of LRC-JEPA on the Cube environment. Each entry reports MSE / Pearson correlation $r$ (lower / higher is better). Here, $\oplus$ denotes concatenation.

| **Quantity** | **Linear** | **MLP** |
|---|---|---|
| Joint position | 0.272 / 0.731 | 0.306 / 0.711 |
| Joint velocity | 0.746 / 0.397 | 1.002 / 0.088 |
| End-effector position | 0.009 / 0.996 | 0.019 / 0.991 |
| End-effector yaw | 0.891 / 0.323 | 0.923 / 0.299 |
| Gripper | 0.058 / 0.970 | 0.060 / 0.969 |
| Block position | 0.004 / 0.998 | 0.011 / 0.995 |
| Block quaternion | 0.673 / 0.160 | 0.711 / 0.062 |
| Block yaw | 0.989 / 0.173 | 1.031 / 0.150 |

Table 7: Object-centric physical quantity probes using $\boldsymbol{z}$ and mean-pooled ($Q = 2$) $\boldsymbol{u}$ on Bridge-v2. Each entry reports MSE / Pearson correlation $r$. One valid moving, static, and background object is selected per frame. Bold marks the better representation for each probe type and metric.

| | $z$ | | $u$ | |
|---|---|---|---|---|
| Probe target | Linear | MLP | Linear | MLP |
| Gripper centroid | **0.724 / 0.690** | **0.604 / 0.735** | 1.193 / 0.321 | 0.858 / 0.402 |
| Moving-object centroid | **1.001 / 0.484** | **0.879 / 0.473** | 1.528 / 0.225 | 0.995 / 0.323 |
| Moving-object RGB | 2.440 / 0.131 | **1.141 / 0.269** | **1.914 / 0.208** | 1.179 / 0.265 |
| Static-object centroid | 1.841 / **0.252** | 1.277 / 0.259 | **1.508** / 0.246 | **1.239 / 0.267** |
| Static-object RGB | 2.324 / 0.036 | 1.208 / 0.326 | **2.035 / 0.137** | **0.999 / 0.413** |
| Background centroid | 1.241 / **0.425** | **0.554 / 0.609** | **1.211** / 0.337 | 0.684 / 0.489 |

Table 8: Object-centric physical quantity probes using the concatenated $\boldsymbol{z} \oplus \boldsymbol{u}$ representation on Bridge-v2. Each entry reports MSE / Pearson correlation $r$ (lower / higher is better) for linear and MLP probes. Here, $\boldsymbol{u}$ is mean-pooled ($Q = 2$) before concatenation. One valid moving, static, and background object is selected per frame.

| Probe target | Linear | MLP |
|---|---|---|
| Gripper centroid | 10.829 / 0.299 | 0.703 / 0.690 |
| Moving-object centroid | 15.260 / 0.144 | 0.888 / 0.460 |
| Moving-object RGB | 13.374 / 0.165 | 1.277 / 0.258 |
| Static-object centroid | 13.174 / $-0.018$ | 1.219 / 0.289 |
| Static-object RGB | 11.721 / 0.204 | 1.200 / 0.347 |
| Background centroid | 4.750 / 0.156 | 0.598 / 0.571 |

### D.1 RECONSTRUCTION VISUALIZATION

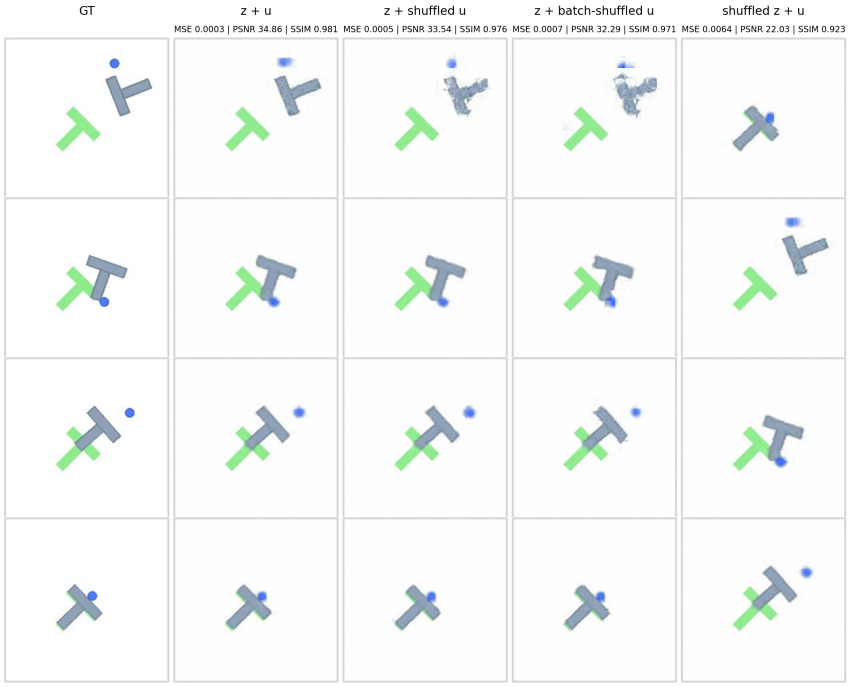



Figure 7: Push-T reconstructions. "Shuffled" permutes frames within a trajectory; "batch-shuffled" exchanges context across trajectories.

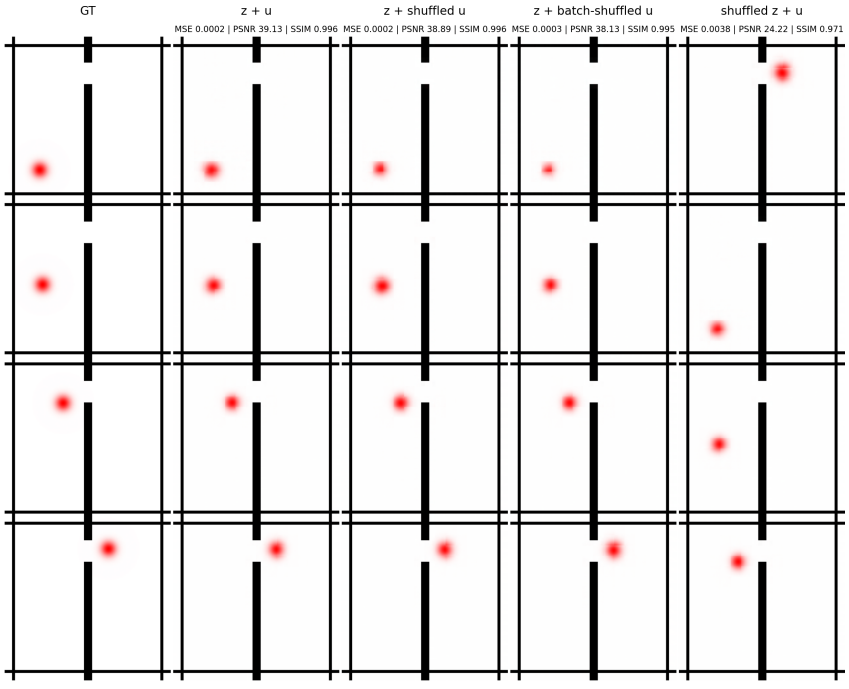



Figure 8: TwoRoom reconstructions under the same within-trajectory and across-trajectory interventions.