Sub-JEPA: Subspace Gaussian Regularization for Stable End-to-End World Models

Kai Zhao[1], Dongliang Nie[1], Yuchen Lin[1], Zhehan Luo[2], Yixiao Gu[1], Deng-Ping Fan[3], Dan Zeng[1]

[1]Shanghai University [2]The University of Manchester [3]Nankai University

https://github.com/intcomp/sub-jepa

Abstract

Joint-Embedding Predictive Architectures (JEPAs) provide a simple framework for learning world models by predicting future latent representations. However, JEPA training is subject to a bias-variance tradeoff. Without sufficient structural constraints, excessive representational variance causes the model to collapse to trivial solutions. The recent LeWorldModel (LeWM) shows that this issue can be alleviated by simply constraining latent embeddings with an isotropic Gaussian prior. However, latent representations inherently lie on low-dimensional manifolds within a high-dimensional ambient space, and enforcing an isotropic Gaussian prior directly in this ambient space introduces an overly strong bias. In this work, we propose Sub-JEPA, which seeks a favorable operating point on the bias-variance frontier by applying Gaussian constraints in multiple random subspaces rather than in the original embedding space. This design relaxes the global constraint while preserving its anti-collapse effect, leading to a better balance between training stability and representation flexibility. Extensive experiments across four continuous-control environments demonstrate that Sub-JEPA consistently outperforms LeWM with very clear margins. Our method is simple yet effective, and can serve as a strong baseline for future JEPA-based world model research.

Keywords: World models, LeWorldModel, JEPA, subspace, Gaussian.

1 Introduction

World models (WM) [1, 2], predictive representations of how environments evolve under actions, have become critical building blocks of modern artificial intelligence. WM provides a principled route of learning compact representations of the observations and predicting future states under candidate actions. Such predictive models are attractive because they support planning in latent space while avoiding the cost of modeling every pixel in the observations [2, 3].

Among recent approaches, the Joint-Embedding Predictive Architecture (JEPA) has emerged as a particularly appealing formulation for latent world modeling [4]. JEPA learns an encoder that maps observations to latent representations and a dynamics model that predicts future latent states conditioned on the current representation and action [5]. This design shifts modeling capacity toward task-relevant abstractions and avoids the burden of full observation reconstruction [6].

Despite this conceptual simplicity, stable end-to-end training of JEPA-based models is subject to a bias-variance tradeoff [7, 5]: insufficient constraints lead to representation collapse, while excessive constraints suppress the richness of learned representations. Early methods attempted to mitigate collapse by introducing various regularization terms [8, 9] or heuristic training recipes [6, 10, 11], but these approaches are often complex, not end-to-end, and lack theoretical guidance. The LeWorldModel (LeWM) [12] pioneers stable end-to-end training of JEPA-based world models by introducing a simple bias that regularizes latent embeddings toward an isotropic Gaussian distribution.

However, the latent representations of natural control tasks typically lie on low-dimensional manifolds embedded within the high-dimensional ambient space [13, 14], and enforcing an isotropic Gaussian prior directly in this ambient space introduces an overly strong bias that mismatches the intrinsic geometry of the underlying dynamics [12].

In this work, we propose Sub-JEPA, which seeks a more favorable operating point on the bias-variance frontier by moving Gaussian regularization from the ambient space into low-dimensional subspaces. This simple design relaxes the global isotropic prior while preserving its anti-collapse effect, yielding a bias better matched to the intrinsic dimensionality of the underlying dynamics. Sub-JEPA retains the simplicity of the LeWM training recipe while introducing a more flexible structural prior over the latent space. Our main contributions are as follows:

- We identify that the global isotropic Gaussian prior in LeWM introduces an excessive bias for tasks with low intrinsic dimensionality.
- We propose Sub-JEPA, which relocates Gaussian regularization from the ambient embedding space to multiple low-dimensional row-orthonormal projected views, effectively relaxing the global prior

while preserving the anti-collapse effect.

- Extensive experiments across four continuous-control environments demonstrate that Sub-JEPA consistently outperforms LeWM by clear margins, with gains correlated with reductions in effective rank, confirming that subspace regularization better respects the intrinsic dimensionality of the underlying dynamics.

2 Related Work

Latent World Models. World models learn a compressed model of environment dynamics from observations, enabling agents to plan without repeated real-world interaction [1]. Early approaches predict in pixel space, which is computationally expensive and burdens the model with irrelevant visual details [4]. Autoregressive generative models such as IRIS [3] and DreamerV3 [2] couple the world model with an image decoder and achieve strong results in reward-driven settings, but reconstruction-based objectives can produce embeddings that are uninformative for control. Latent-space world models instead predict future embeddings directly, avoiding reconstruction cost while concentrating capacity on task-relevant structure [4]. Test-time planning via Model Predictive Control over such latent dynamics models has demonstrated strong performance across continuous-control and navigation tasks [15, 11, 9], but requires a well-structured, non-degenerate latent space.

JEPA and End-to-End Training. Joint-Embedding Predictive Architectures [4] train an encoder and a predictor jointly to anticipate the embedding of a future or masked view, without reconstructing raw pixels. This recipe has been instantiated in image representation learning with I-JEPA [6], in video modeling with V-JEPA [10], and in action-conditioned latent dynamics [5, 11]. The central challenge is representation collapse [7]: without explicit structural constraints the encoder can map all inputs to nearly identical embeddings, trivially minimizing the prediction loss while destroying useful structure. Existing latent world model solutions avoid collapse either by using pretrained frozen encoders such as DINOv2 [11], which limits representation flexibility, or by adding multi-term collapse-prevention losses, as in PLDM [9], which applies VICReg [8] and requires tuning several sensitive hyperparameters. LeWorldModel (LeWM) [12] achieves stable end-to-end training from raw pixels using only two loss terms: a next-embedding prediction objective and Gaussian regularizer.

Anti-Collapse Regularization. A broad class of self-supervised methods prevents collapse by imposing structure on the embedding distribution. Contrastive methods such as SimCLR [16] and MoCo [17] push apart embeddings of different samples but require large batches or memory banks. Non-contrastive approaches such as BYOL [18] rely on teacher-student asymmetry with stop-gradient, while Barlow Twins [19] and VICReg [8] decorrelate embedding dimensions; Whitening-MSE [20] further enforces a uniform distribution on the unit sphere. Gaussian regularization, introduced in LeJEPA [21], takes a principled approach: random one-dimensional projections of the embeddings are tested for Gaussianity, and by the Cramér-Wold theorem [22] matching all projected marginals to a Gaussian implicitly enforces an isotropic Gaussian joint distribution. LeJEPA proves this prior to be the unique optimal embedding distribution for minimizing downstream prediction risk, placing the collapse-prevention objective on a rigorous theoretical footing.

Subspace Methods and Random Projections. Random projections underpin scalable dimensionality reduction via the Johnson-Lindenstrauss lemma [23] and enable tractable distribution matching through sliced Wasserstein distances [24], which reduce high-dimensional optimal transport to a sequence of one-dimensional comparisons. Gaussian regularization itself belongs to this family, sketching the embedding distribution with random directions to bypass the curse of dimensionality. In representation learning, orthogonal projections have been used to decorrelate and spread features [20]. Sub-JEPA extends this principle to Gaussian regularization by applying the regularizer over multiple fixed random row-orthonormal projections of the latent representation. Instead of enforcing a single global isotropic prior in the ambient space, the regularization is imposed independently on multiple low-dimensional projected views, yielding a more flexible inductive bias that better matches the intrinsic structure of the underlying dynamics.

3 Method

In this section, we formally introduce our subspace-regularized latent world model for learning from state-action trajectories without reward annotations. We first present the problem setting of latent dynamics learning, and then describe our proposed Sub-JEPA framework, which extends LeWorldModel (LeWM) [12] by replacing full-space Gaussian regularization with Gaussian regularization over multiple frozen row-orthonormal random projections.

3.1 Problem Setup

As illustrated in Figure 1, a latent world model consists of an encoder f and a predictor P. Given a raw observation o_t, the encoder maps it to a compact latent representation $\boldsymbol{z}_t = f(o_t) \in \mathbb{R}^D$. The predictor then estimates the next latent $\hat{\boldsymbol{z}}_{t+1} = P(\boldsymbol{z}_t, a_t)$ conditioned on the current latent and action a_t, and is trained to match the target latent $\boldsymbol{z}_{t+1} = f(o_{t+1})$ obtained by encoding the subsequent ob-

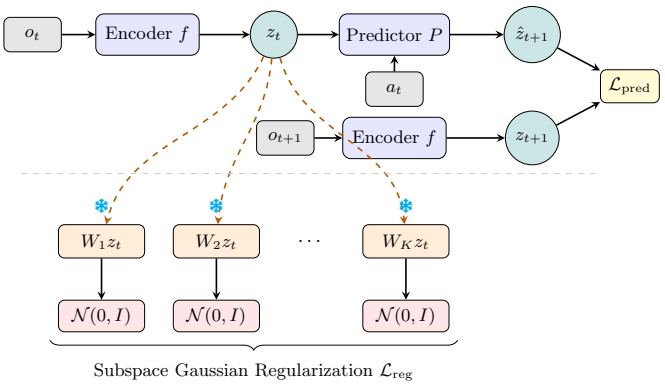

Figure 1: Overview of Sub-JEPA. Observations o_t and o_{t+1} are encoded by a shared encoder f into latents z_t and z_{t+1}. The predictor P maps (z_t, a_t) to $\hat{\boldsymbol{z}}_{t+1}$, trained with prediction loss $\mathcal{L}_{\text{pred}}$. Below the dashed line, z_t is projected onto K frozen (❄) row-orthonormal random projections. $\{W_k\}$; a subspace Gaussian regularization loss $\mathcal{L}_{\text{sub}}$ enforces $\mathcal{N}(0, I)$ in each subspace.

servation. This formulation captures environment dynamics entirely in latent space, without requiring pixel-level reconstruction.

We consider a fully offline and reward-free setting, where the model is trained only from pre-collected trajectories of observations and actions, without access to rewards or task labels, following the JEPA line of work [25, 12, 11]. The offline dataset consists of trajectories of length T:

$$\mathcal{D} = \{(o_{1:T}, a_{1:T})\}, \tag{1}$$

where o_t and a_t denote the observation (e.g. an RGB image of the environment), and action at time step t (e.g. continuous control signal), respectively.

3.2 Orthogonal Subspace Projection

Let $\boldsymbol{z} \in \mathbb{R}^D$ denote the latent representation produced by the encoder. We consider a collection of K linear projections that map $\boldsymbol{z}$ to low-dimensional representations in $\mathbb{R}^{d_s}$. Specifically, we introduce

$$\{\mathbf{P}_k \in \mathbb{R}^{d_s \times D}\}_{k=1}^{K}, \tag{2}$$

where $d_s < D$ denotes the subspace dimension. We set $d_s = \lfloor D/K \rceil$ to reduce the number of hyperparameters and maintain a simple parameterization. The projected embedding is defined as

$$\boldsymbol{z}^{(k)} = \mathbf{P}_k \boldsymbol{z} \in \mathbb{R}^{d_s}, \quad k = 1, \ldots, K. \tag{3}$$

Each projection matrix $\mathbf{P}_k$ is constructed by first sampling a random Gaussian matrix, followed by QR decomposition to obtain an orthonormal basis. The resulting orthogonal matrix is subsequently transposed to form the projection matrix $\mathbf{P}_k \in \mathbb{R}^{d_s \times D}$, yielding a row-orthonormal projection matrix. The orthogonalization ensures that the resulting projection directions are mutually orthogonal, preventing redundancy and preserving geometric structure during projection [26].

Freezing the projections prevents the regularizer itself from adapting to the evolving latent distribution [27, 28], ensuring that the Gaussian constraint is consistently applied across stable and independent subspaces throughout training.

3.3 Multi-Subspace Gaussian Regularization

Let $\mathbf{Z} \in \mathbb{R}^{N \times B \times D}$ denote the latent embedding tensor collected over temporal history length N and batch size B. We apply Gaussian regularization independently in each of the K row-orthonormal random projection subspaces. Concretely, let $\{\mathbf{P}_k \in \mathbb{R}^{d_s \times D}\}_{k=1}^{K}$ denote the orthogonal projection matrices introduced in Section 3.2. The projected tensor for the k-th subspace is obtained by applying $\mathbf{P}_k$ along the embedding dimension of $\mathbf{Z}$:

$$\mathbf{Z}^{(k)} = \mathbf{Z}\mathbf{P}_k^{\top} \in \mathbb{R}^{N \times B \times d_s}, \quad k = 1, \ldots, K, \tag{4}$$

where each slice $\mathbf{Z}^{(k)}_{n,b,:} \in \mathbb{R}^{d_s}$ is the d_s-dimensional subspace representation of the latent embedding at temporal index n and batch index b.

We then apply Gaussian regularization independently in each subspace. Following the formulation of LeWM, we sample M random unit vectors $\{u^{(m)}\}_{m=1}^{M} \subset \mathcal{S}^{d_s - 1}$, and compute one-dimensional projections by taking inner products along the last dimension:

$$z_{n,b}^{(k,m)} = \left\langle \mathbf{Z}^{(k)}_{n,b,:}, u^{(m)} \right\rangle, \quad n = 1, \ldots, N, \; b = 1, \ldots, B. \tag{5}$$

Here, the superscript (k, m) indicates that the scalar projection is computed in the k-th subspace along the m-th random direction, while the subscripts (n, b) denote the temporal index and batch index, respectively. Therefore, $z_{n,b}^{(k,m)}$ is a scalar value obtained from the latent embedding at temporal index n and batch index b.

For each (k, m), all projections $\{z_{n,b}^{(k,m)}\}_{n=1,b=1}^{N,B}$ are scalar samples. We evaluate the Epps–Pulley [29] normality statistic on this sample set:

$$T^{(k,m)} = T\left(\{z_{n,b}^{(k,m)}\}_{n=1,b=1}^{N,B}\right). \tag{6}$$

The regularization objective averages over M random directions and K subspaces:

$$\mathcal{L}_{\text{reg}} = \frac{1}{KM} \sum_{k=1}^{K} \sum_{m=1}^{M} T^{(k,m)}. \tag{7}$$

The overall training objective is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{pred}}(\hat{\boldsymbol{z}}_{t+1}, \boldsymbol{z}_{t+1}) + \lambda \mathcal{L}_{\text{reg}}(\mathbf{Z}), \tag{8}$$

Here, $\mathcal{L}_{\text{pred}}$ denotes the latent-state prediction error (e.g., mean squared error), and λ is the regularization weight.

4 Experiments

4.1 Experimental Setup

Environments. We evaluate Sub-JEPA on four continuous-control benchmarks: Two-Room [11], a 2D navigation task with low intrinsic dimensionality; Reacher [30], a two-link planar reaching task; PushT [31], a 2D block-pushing manipulation task; and OGB-Cube [32], a visually rich 3D manipulation environment. All tasks have continuous action spaces, and all models are trained directly from raw RGB observations.

Baselines. We compare Sub-JEPA against three representative baselines. LeWM [12] is our primary baseline and the direct foundation of our method. It imposes a global Gaussian constraint in the ambient embedding space. PLDM [9] is an end-to-end pixel-based world model trained with multiple heuristic objectives. DINO-WM [11] uses a frozen pretrained DINOv2 [33] visual encoder to mitigate representation collapse. For a fair comparison with pixel-only end-to-end world models, we report DINO-WM without proprioceptive inputs as the main reference and include the proprioceptive variant only as an additional upper-bound reference when available.

Implementation Details. Sub-JEPA uses the same encoder, latent predictor, optimizer, training schedule, and loss-weight tuning protocol as LeWM, except that full-space Gaussian regularization is replaced by Multi-Subspace Gaussian regularization (Section 3.3). The embedding dimension is set to $D = 192$ for all experiments. We use $K = 32$ for all environments except PushT where we instead use $K = 16$ for better performance. The values are selected through ablation studies on a held-out validation set, as described in Section 4.3. The projection matrices $\{\mathbf{P}_k\}$ are orthogonally initialized and kept frozen throughout training, as described in Section 3.2.

4.2 Planning Performance

4.2.1 Success Rates

We first report the planning success rates. Overall, Sub-JEPA consistently outperforms LeWM [12] across all four environments, demonstrating that relocating the Gaussian regularization into subspaces yields more powerful latent representations.

For a fair comparison, LeWM and Sub-JEPA are evaluated across six random seeds (reporting mean ± std). Since official checkpoints for PLDM [9] and DINO-WM [11] are unavailable, we cite their success rates directly from LeWM under the exact same protocol.

As shown in Table 1, the improvement is most prominent on Two-Room [11]. LeWM is known to struggle here because enforcing isotropic Gaussianity over the entire high-dimensional embedding introduces an overly restrictive bias for this low-intrinsic-dimension task. By contrast, Sub-JEPA regularizes lower-dimensional orthogonal views. This relaxes the global constraint and better preserves task-relevant structure while maintaining the anti-collapse benefits.

On Reacher [30] and PushT [31], Sub-JEPA achieves strong performance, consistently outperforming PLDM and surpassing DINO-WM without relying on a pretrained visual backbone. On OGB-Cube [32], Sub-JEPA substantially improves upon LeWM. Although DINO-WM still leads in visually complex 3D manipulation tasks (where pretrained features excel), Sub-JEPA provides a highly effective, simple end-to-end alternative without relying on external pretraining.

4.2.2 Success Rates w.r.t. Effective Rank

To verify whether our subspace regularization yields more compact latent representations as motivated, we analyze the effective rank [34] of the learned embeddings. Effective rank measures how many principal directions are effectively used by the latent embeddings.

For each environment, we uniformly sample $N = 2000$ observations from the evaluation set and use the same sampled observations for LeWM and Sub-JEPA. Each observation is fed through the trained encoder with weights frozen, and the projected `[CLS]` embedding is used as the latent representation. Let $Z \in \mathbb{R}^{N\times D}$ denote the centered latent matrix, and let $\{\lambda_i\}_{i=1}^{D}$ be the eigenvalues of its empirical covariance matrix. We normalize the covariance spectrum as $p_i = \lambda_i / \sum_j \lambda_j$ and define the effective rank as

$$r_{\text{eff}} = \exp\left(-\sum_i p_i \log p_i\right). \tag{9}$$

As shown in Figure 2, the effective-rank reduction from LeWM to Sub-JEPA is strongly and directly correlated with planning improvement across all four environments. Two-Room [11] and OGB-Cube [32] exhibit the largest rank compressions and simultaneously the largest success-rate gains, while PushT [31] and Reacher [30] show smaller rank gaps and more moderate improvements.

This direct correspondence strongly supports our hypothesis: when task-relevant features lie on a low-dimensional manifold, the full-space Gaussian prior in LeWM forces the latent space into unnecessarily high rank, imposing a bias that hinders task-aligned representation learning. Subspace-wise regularization relaxes this bias and allows the latent geometry to contract toward the task's intrinsic dimensionality—the greater this contraction, the larger the planning gain.

4.3 Ablation Studies

We conduct ablation studies to validate the design choices of Sub-JEPA and to understand the bias-variance tradeoff

Table 1: Planning success rate (%, higher is better) across four continuous-control environments. Sub-JEPA consistently improves upon LeWM across all tasks, with the largest gain on Two-Room [11], where the full-space Gaussian prior of LeWM is most mismatched to the low intrinsic dimensionality of the task.

Method	Two-Room [11]	Reacher [30]	PushT [31]	OGB-Cube [32]
PLDM [9]	97.00	78.00	78.00	65.00
DINO-WM (w/o proprio.) [11]	100.00	79.00	74.00	86.00
DINO-WM (w/ proprio.) [11]	100.00	–	92.00	–
LeWM [12]	$84.33_{\pm 4.23}$	$82.67_{\pm 4.42}$	$84.67_{\pm 6.53}$	$67.33_{\pm 5.01}$
Sub-JEPA (Ours)	$95.00_{\pm 2.76}$	$84.00_{\pm 4.00}$	$89.00_{\pm 5.33}$	$76.33_{\pm 5.99}$

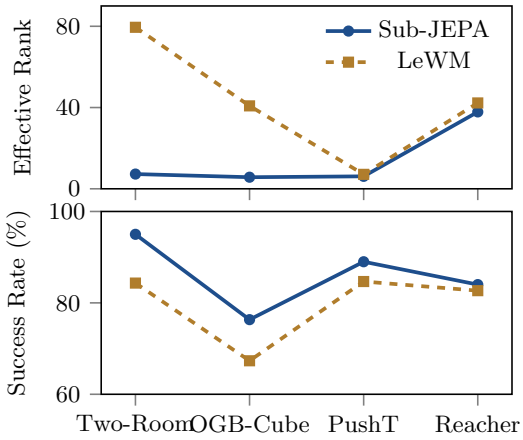

Figure 2: Effective rank and planning success rate across four environments. Top: effective rank of latent representations. Bottom: planning success rate. Larger effective-rank reductions from LeWM to Sub-JEPA broadly coincide with larger planning gains, suggesting that Multi-Subspace Gaussian regularization improves performance by reducing spurious high-rank variation in the latent space.

in subspace design.

4.3.1 Number of subspaces K

We first study the effect of the number of subspaces K. Specifically, we set $D = 192$ and vary $K \in \{1, 2, 4, 8, 16, 32, 64\}$, which corresponds to subspace dimensions $d_s = \{192, 96, 48, 24, 12, 6, 3\}$. When $K = 1$ and $d_s = D$, the regularizer reduces to a full-space Gaussian regularization with an orthogonal transformation, making it the closest counterpart to LeWM [12] within our implementation, whereas larger K values decompose the same latent space into progressively smaller subspaces.

Results are reported in Table 2. The effect of K is task-dependent. On Two-Room [11], performance improves as K increases from small to moderate values, peaking around $K = 32$, after which it slightly degrades. This is consistent with the low-dimensional structure of the task: relaxing the full-space Gaussian constraint allows the model to capture the underlying navigation dynamics more faithfully. A similar trend is observed on OGB-Cube [32], where larger K consistently improves performance and achieves the best result at $K = 32$. Reacher [30] also benefits slightly from moderate increases in K, suggesting that subspace-wise regularization remains effective even in visually complex or continuous-control settings.

Table 2: Planning success rate (%) for different numbers of subspaces K under fixed total latent dimensionality $D = 192$, where $d_s = 192/K$. Gray cells indicate configurations that outperform the LeWM baseline under the same evaluation protocol. Subspace-wise regularization improves performance across a broad range of configurations.

K	Two-Room	Reacher	PushT	OGB-Cube
LeWM	$84.33_{\pm 4.23}$	$82.67_{\pm 4.42}$	$84.67_{\pm 6.53}$	$67.33_{\pm 5.01}$
1	$87.00_{\pm 4.43}$	$82.00_{\pm 6.32}$	$84.67_{\pm 5.47}$	$68.00_{\pm 6.20}$
2	$91.33_{\pm 4.42}$	$80.67_{\pm 6.39}$	$88.00_{\pm 7.04}$	$68.00_{\pm 5.40}$
4	$92.00_{\pm 2.83}$	$83.67_{\pm 3.88}$	$88.67_{\pm 4.76}$	$70.00_{\pm 6.45}$
8	$91.00_{\pm 3.29}$	$82.68_{\pm 4.38}$	$85.67_{\pm 2.94}$	$70.33_{\pm 5.89}$
16	$93.67_{\pm 4.27}$	$81.00_{\pm 2.10}$	$89.00_{\pm 5.33}$	$69.00_{\pm 8.69}$
32	$95.00_{\pm 2.76}$	$84.00_{\pm 4.00}$	$28.00_{\pm 5.22}$	$76.33_{\pm 5.99}$
64	$92.00_{\pm 2.34}$	$74.33_{\pm 7.53}$	$9.00_{\pm 5.33}$	$67.67_{\pm 6.12}$

A different trend is observed on PushT [31]. In PushT, performance collapses at $K = 32$, where each subspace contains only six dimensions. This suggests that excessively aggressive partitioning can make individual subspaces too low-dimensional to provide stable and informative normality estimates [35], particularly for manipulation tasks requiring tightly coupled object–agent interactions.

Overall, K controls a trade-off between flexibility and statistical reliability. Increasing K relaxes the global Gaussian prior and enables more expressive latent structure, while excessively small subspaces may weaken the effectiveness of the regularization signal.

4.3.2 Joint effect of K and d_s

In this ablation, we relax the constraint $d_s = D/K$ and independently vary both K and d_s to study their joint

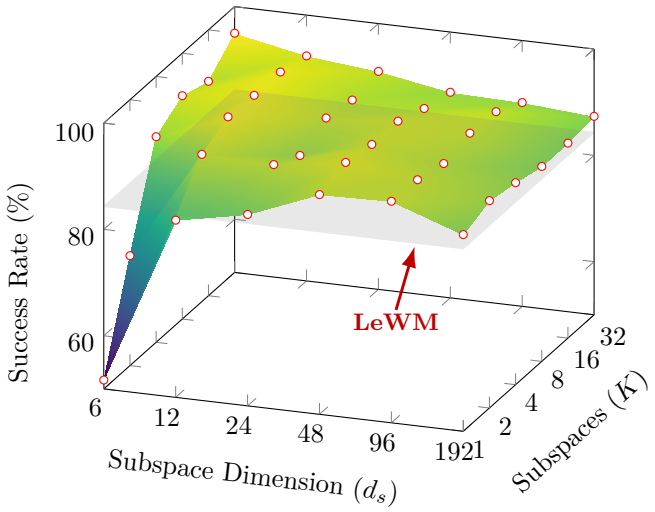

Figure 3: Success rate on Two-Room [11] is shown as a function of K and d_s, with the baseline success rate of LeWM [12] shown as a flat reference plane. Sub-JEPA outperforms LeWM across a broad mid-range of configurations.

effect on planning performance.

As shown in Figure 3, performance is stable across a broad mid-range of (K, d_s) configurations, but degrades sharply when subspaces become too narrow (small d_s), where the normality signal grows unreliable. This reveals a bias-variance tradeoff in subspace design: larger K relaxes the global prior but requires sufficient d_s to maintain regularization stability.

4.3.3 Justification of frozen orthogonal projection

We use orthogonally initialized frozen projection matrices and ablate this choice against two alternatives: (1) randomly initialized frozen projection, where the projection matrices are sampled from a standard Gaussian and kept fixed; and (2) trainable projection with soft orthogonality regularization, where the matrices are initialized orthogonally but updated during training subject to an orthogonality penalty.

As shown in Table 3, orthogonal frozen projection consistently achieves the best performance across all environments. This confirms that the projection design is not merely an implementation detail, but a key component of Sub-JEPA. By preserving geometric isometry, the orthogonal frozen projection ensures that each subspace receives a balanced and non-redundant view of the latent representation, allowing the subspace-level normality constraints to act uniformly [23].

Random frozen projection performs substantially worse, especially on PushT [31]. Without orthogonality, different subspaces may contain redundant or unevenly scaled information, weakening the intended projection of the full latent space. Trainable projection with soft orthogonality regularization also underperforms the frozen orthogonal variant. We attribute this to co-adaptation between the encoder and the projection matrices: as training progresses, the learned projections can align with directions that reduce the effective strength of the regularizer, thereby weakening the anti-collapse effect of Multi-Subspace Gaussian regularization.

Table 3: Planning success rate (%) across environments for different projection strategies. Orthogonal frozen projection gives the most stable and effective subspace projection.

Projection	Two-Room	Reacher	PushT	OGB-Cube
Ortho init ❄	**95.00**$_{\pm 2.76}$	**84.00**$_{\pm 4.00}$	**89.00**$_{\pm 5.33}$	**76.33**$_{\pm 5.99}$
Rand init ❄	53.00$_{\pm 8.44}$	68.00$_{\pm 5.29}$	13.33$_{\pm 5.61}$	61.00$_{\pm 5.55}$
Ortho reg 🔥	61.67$_{\pm 5.16}$	82.67$_{\pm 4.68}$	57.00$_{\pm 9.27}$	70.33$_{\pm 7.09}$

4.4 Physical State Probing

Planning success evaluates whether a learned world model supports goal-directed control, but it does not directly characterize the physical information contained in the latent representation. We therefore evaluate the physical decodability of the learned representations on PushT [31]. Following the probing protocol of LeWM [12], we freeze the encoder and train lightweight probes from latent embeddings to ground-truth physical variables. We consider three task-relevant quantities: agent location, block location, and block angle. For each quantity, we train both a linear probe and a small MLP probe. The linear probe measures whether the variable is directly accessible from the representation, while the MLP probe tests whether the information is retained in a possibly nonlinear form. We report mean squared error (MSE) and Pearson correlation coefficient r, averaged over six random seeds.

Table 4 reports the physical latent probing results on PushT. Overall, Sub-JEPA matches or outperforms LeWM across most properties and probe types, suggesting that Multi-Subspace Gaussian regularization preserves physically meaningful latent structure while improving nonlinear recoverability.

An interesting exception arises for block angle prediction under the linear probe, where Sub-JEPA slightly underperforms LeWM. We hypothesize that rotation is an inherently non-linear quantity—subspace projection may fragment the angular structure across independent subspaces, making it less linearly decodable. Consistent with this view, the gap closes under the MLP probe, suggesting that a nonlinear decoder can compensate for this fragmentation. This pattern highlights a nuanced tradeoff: subspace regularization improves the linear decodability of translational quantities while potentially complicating that of rotational ones.

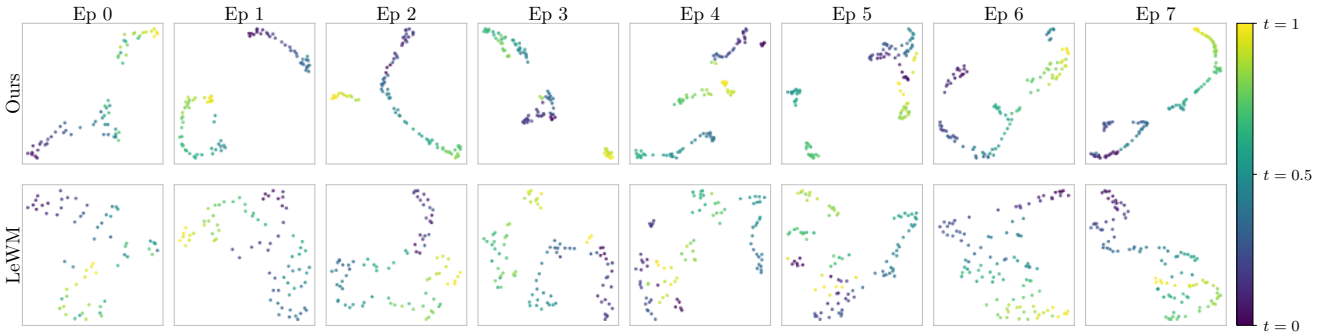

Figure 4: Visualization of latent embeddings extracted from consecutive observations in representative Two-Room [11] episodes. Each column corresponds to one episode, with Sub-JEPA shown in the top row and LeWM shown in the bottom row. Each point denotes one time step, and colors indicate the normalized timestamp. Sub-JEPA produces more temporally coherent latent trajectories, while LeWM exhibits less organized temporal evolution in several episodes.

Table 4: Physical latent probing results on PushT.

Property	Model	Linear		MLP	
		MSE ↓	r ↑	MSE ↓	r ↑
Agent Loc.	PLDM	0.090	0.955	0.014	0.993
	LeWM	0.052	0.974	0.004	0.998
	Ours	**0.048**	**0.976**	**0.002**	**0.999**
Block Loc.	PLDM	0.122	0.938	0.011	0.994
	LeWM	0.029	0.986	**0.001**	0.999
	Ours	**0.024**	**0.988**	**0.001**	**1.000**
Block Angle	PLDM	0.446	0.745	0.056	0.972
	LeWM	**0.187**	**0.902**	0.022	0.989
	Ours	0.218	0.884	**0.021**	**0.990**

4.5 Latent Trajectories

To verify that Sub-JEPA learns representations better matched to the task geometry, we visualize latent trajectories on Two-Room [11], where the low intrinsic dimensionality makes the mismatch with the full-space Gaussian prior most pronounced. For each episode, consecutive observations are encoded and the `[CLS]` embeddings are projected to 2D via UMAP [36], colored by normalized temporal index.

As shown in Figure 4, Sub-JEPA consistently produces temporally coherent latent trajectories across episodes, with temporally adjacent observations forming well-organized paths. LeWM, by contrast, exhibits less regular temporal structure in several episodes, suggesting that the full-space Gaussian prior distorts latent geometry when task dynamics are intrinsically low-dimensional.

4.6 Temporal Latent Path Straightening

Following LeWM [12], we examine latent trajectory geometry via temporal path straightening, which measures how linearly dynamics evolve in latent space [37, 38]—a property particularly relevant for latent world models where planning quality depends on rollout regularity rather than pixel reconstruction [1, 15, 4].

Given latent embeddings $z_{1:T} \in \mathbb{R}^{B\times T\times D}$, we define the temporal velocity as $v_t = z_{t+1} - z_t$, and measure straightness via the mean cosine similarity between consecutive velocities [12]:

$$S_{\text{straight}} = \frac{1}{B(T-2)} \sum_{i=1}^{B} \sum_{t=1}^{T-2} \frac{\left\langle v_t^{(i)}, v_{t+1}^{(i)} \right\rangle}{\left\| v_t^{(i)} \right\| \left\| v_{t+1}^{(i)} \right\|}. \tag{10}$$

Higher values indicate smoother and more linearly evolving latent trajectories.

As shown in Figure 6, Sub-JEPA consistently produces straighter latent trajectories than LeWM on both PushT [31] and OGB-Cube [32], emerging naturally without explicit optimization. This suggests that subspace-wise regularization reduces geometric distortion relative to full-space SIGReg, offering a geometric explanation for the planning gains in Table 1.

4.7 Open-loop Rollout Visualization

To evaluate long-horizon stability, we compare open-loop rollouts of Sub-JEPA and LeWM [12] on Two-Room [11]. Each model is conditioned on three context frames at $t = \{0, 5, 10\}$ and autoregressively predicts future states from actions alone, decoded into RGB over 20 steps up to $t = 95$. As shown in Figure 5, both models produce accurate short-horizon predictions, but diverge at longer horizons. LeWM exhibits increasing spatial drift and structural distortion, while Sub-JEPA maintains scene geometry more consistently, yielding lower reconstruction error at distant steps.

This suggests that subspace-wise regularization improves recursive latent stability: by replacing a single global Gaussian constraint with structured subspace constraints, Sub-JEPA better supports coherent long-term state propagation.

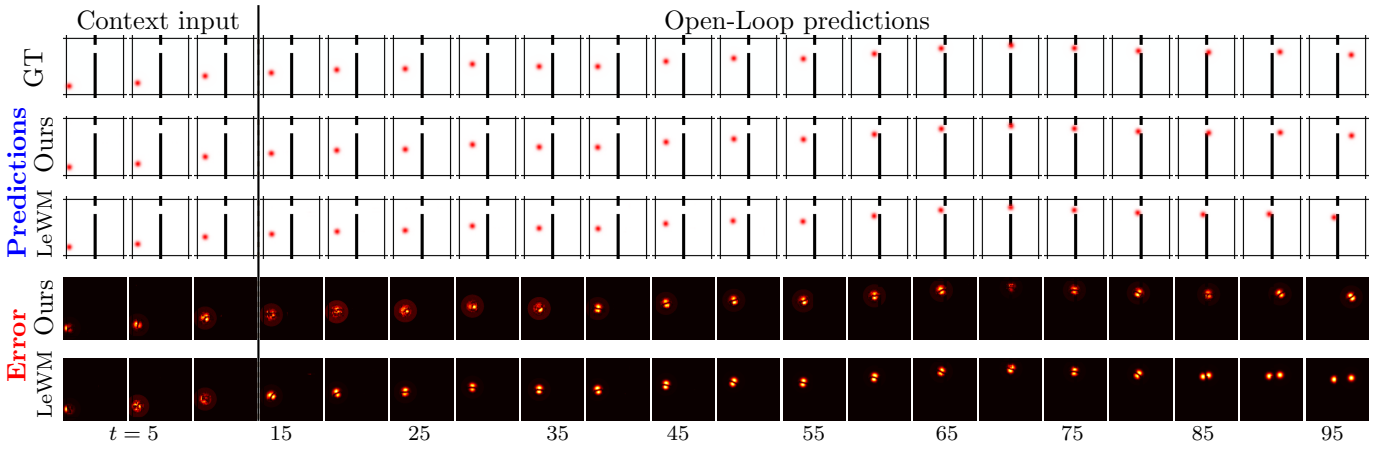

Figure 5: Open-loop rollout comparison between Sub-JEPA and LeWM on Two-Room [11]. Models are conditioned on three context frames ($t = \{0, 5, 10\}$) and then recursively predict future observations using only actions. Rows 2–3 show predicted trajectories for Sub-JEPA and LeWM, respectively. Rows 4–5 show absolute pixel errors relative to the ground truth. Sub-JEPA exhibits more accurate long-horizon predictions and reduced cumulative drift.

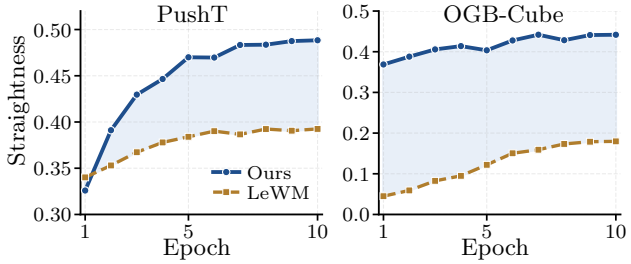

Figure 6: Temporal latent path straightening over training, measured by mean cosine similarity between consecutive latent velocity vectors (higher = straighter trajectories). Sub-JEPA achieves consistently higher temporal straightness than LeWM on both environments, suggesting that Multi-Subspace Gaussian regularization implicitly encourages smoother latent dynamics.

Conclusion

We presented Sub-JEPA, a simple yet effective extension of LeWorldModel that relocates Gaussian regularization from the ambient embedding space into random orthogonal subspaces. Our key insight is that the isotropic Gaussian prior enforced by LeWM operates in the full high-dimensional latent space, which introduces an excessive bias when the underlying task dynamics reside on a low-dimensional manifold. Extensive experiments demonstrate that Sub-JEPA consistently outperforms LeWM across four tasks, that physical state information is better preserved and more nonlinearly decodable in the learned representations, and that the latent trajectories are substantially more temporally coherent.

References

[1] David Ha and Jürgen Schmidhuber. Recurrent World Models Facilitate Policy Evolution. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *NeurIPS*, volume 31. Curran Associates, Inc., 2018.

[2] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640(8059):647–653, 2025.

[3] Vincent Micheli, Eloi Alonso, and François Fleuret. Transformers are Sample-Efficient World Models. In *The Eleventh ICLR*, 2023.

[4] Yann LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022.

[5] Vlad Sobal, Jyothir S V, Siddhartha Jalagam, Nicolas Carion, Kyunghyun Cho, and Yann LeCun. Joint Embedding Predictive Architectures Focus on Slow Features, 2022.

[6] 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 *CVPR*, pages 15619–15629, 2023.

[7] Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding Dimensional Collapse in Contrastive Self-supervised Learning. In *ICLR*, 2022.

[8] Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning. In *ICLR*, 2022.

[9] Vlad Sobal, Wancong Zhang, Kyunghyun Cho, Randall Balestriero, Tim G. J. Rudner, and Yann LeCun. Stress-Testing Offline Reward-Free Reinforcement Learning: A Case for Planning with Latent Dynamics Models. In *7th Robot Learning Workshop: Towards Robots with Human-Level Abilities*, 2025.

[10] 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. *TMLR*, 2024. Featured Certification.

[11] Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. DINO-WM: World Models on Pre-trained Visual Features enable Zero-shot Planning. In *ICML*, 2025.

[12] 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*, 2026.

[13] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE TPAMI*, 35(8):1798–1828, 2013.

[14] Joshua B Tenenbaum, Vin de Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[15] Nicklas Hansen, Xiaolong Wang, and Hao Su. Temporal Difference Learning for Model Predictive Control. In *ICML*, 2022.

[16] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607, 2020.

[17] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum Contrast for Unsupervised Visual Representation Learning. In *CVPR*, pages 9726–9735, 2020.

[18] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. In *NeurIPS*, 2020.

[19] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *ICML*, pages 12310–12320, 2021.

[20] Aleksandr Ermolov, Aliaksandr Siarohin, Enver Sangineto, and Nicu Sebe. Whitening for self-supervised representation learning. In *ICML*, pages 3015–3024. PMLR, 2021.

[21] Randall Balestriero and Yann LeCun. LeJEPA: Provable and Scalable Self-Supervised Learning Without the Heuristics, 2025.

[22] H. Cramér and H. Wold. Some Theorems on Distribution Functions. *Journal of the London Mathematical Society*, s1-11(4):290–294, 10 1936.

[23] N. Halko, P. G. Martinsson, and J. A. Tropp. Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions. *SIAM Review*, 53(2):217–288, 2011.

[24] Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. Sliced and Radon Wasserstein Barycenters of Measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45, January 2015.

[25] Mido 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, 2025.

[26] William B. Johnson and Joram Lindenstrauss. Extensions of Lipschitz mappings into Hilbert space. *Contemporary mathematics*, 26:189–206, 1984.

[27] Ali Rahimi and Benjamin Recht. Random Features for Large-Scale Kernel Machines. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *NeurIPS*, volume 20. Curran Associates, Inc., 2007.

[28] A Saxe, J McClelland, and S Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *ICLR*, 2014.

[29] T. W. Epps and Lawrence B. Pulley. A Test for Normality Based on the Empirical Characteristic Function. *Biometrika*, 70(3):723–726, 1983.

[30] Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, Timothy Lillicrap, and Martin Riedmiller. DeepMind Control Suite, 2018.

[31] Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 44(10-11):1684–1704, 2025.

[32] Seohong Park, Kevin Frans, Benjamin Eysenbach, and Sergey Levine. OGBENCH: BENCHMARKING OFFLINE GOAL-CONDITIONED RL. In *ICLR*, pages 57515–57560, 2025.

[33] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, and Alaaeldin El-Nouby *et al.* DINOv2: Learning Robust Visual Features without Supervision. *TMLR*, 2024. Featured Certification.

[34] Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *2007 15th European signal processing conference*, pages 606–610. IEEE, 2007.

[35] Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep Variational Information Bottleneck. In *ICLR*, 2017.

[36] John Healy and Leland McInnes. Uniform manifold approximation and projection. *Nature Reviews Methods Primers*, 4(1):82, 2024.

[37] Olivier J. Hénaff, Robbe L. T. Goris, and Eero P. Simoncelli. Perceptual Straightening of Natural Videos. *Nature Neuroscience*, 22(6):984–991, 2019.

[38] Christian Internò, Robert Geirhos, Markus Olhofer, Sunny Liu, Barbara Hammer, and David Klindt. AI-Generated Video Detection via Perceptual Straightening. In *NeurIPS*, 2026.