SIAMJEPA: ON THE ROLE OF SIAMESE STUDENT ENCODERS IN JEPA

Anonymous authors
Paper under double-blind review

ABSTRACT

Joint Embedding Predictive Architectures (JEPAs) have emerged as a promising framework for self-supervised representation learning by predicting latent embeddings of masked regions rather than reconstructing pixels. Existing JEPA methods typically employ a single student encoder, leaving the role of Siamese student encoders largely unexplored. In this paper, we propose Siamese JEPA (SiamJEPA), a JEPA framework with masked Siamese student encoders and an exponential moving average (EMA) teacher, which can also be viewed as a JEPA formulation of the brain-inspired representation learning model PhiNet. We further introduce Random Shuffle Teacher (RST), which removes spatial correspondence in teacher targets to encourage semantic patch representations, and develop an RST-based semantic-to-spatial curriculum. Experiments on ImageNet show that Siamese student encoders effectively regularize the JEPA objective, improving representation separability and accelerating early-stage learning. Moreover, under RST, stronger Siamese regularization substantially increases class-discriminative information in individual patch tokens, suggesting that the semantic bias arises from the interaction between RST and the Siamese objective rather than from shuffling alone. SiamJEPA consistently outperforms comparable single-encoder JEPA variants under limited training budgets. With RST-based curriculum learning and a ViT-Base backbone, SiamJEPA achieves 73.3% linear-probing accuracy after 450 epochs using a substantially simpler masking strategy, compared with 72.9% for I-JEPA after 600 epochs. Under I-JEPA's linear-probing protocol, the same checkpoint achieves 74.2%, outperforming I-JEPA (72.9%) and DSeq-JEPA (73.5%). These results demonstrate that Siamese student encoders provide an effective inductive bias for predictive representation learning and can be further enhanced through semantic-to-spatial curriculum learning.

1 INTRODUCTION

Recently, representation learning has become a fundamental technique in computer vision, natural language processing, and robotics. Among various approaches, self-supervised learning (SSL) has emerged as one of the most successful paradigms for learning transferable representations from large-scale unlabeled data. Conceptually, several early studies introduced the principles of modern SSL, including information maximization (IMAX) and predictability maximization (PMAX) (Becker & Hinton, 1992; Zemel & Hinton, 1990; Schmidhuber & Prelinger, 1993). More recently, advances in deep neural networks, particularly Transformer architectures such as the Transformer (Vaswani et al., 2017) and the Vision Transformer (ViT) (Dosovitskiy et al., 2021), have enabled highly effective self-supervised representation learning methods.

A major family of SSL methods is based on joint embedding architectures, which learn representations by mapping different views of the same input into a shared latent space using deep neural networks. Representative examples include SimCLR (Chen et al., 2020a), BYOL (Grill et al., 2020), SimSiam (Chen & He, 2021), and DINO (Caron et al., 2021; Oquab et al., 2024; Siméoni et al., 2025). A central challenge of joint embedding methods is to prevent representation collapse. To address this issue, various techniques have been proposed, including contrastive learning with data augmentation (Chen et al., 2020a), stop-gradient operations, and exponential moving average (EMA) teacher networks (Grill et al., 2020; Chen & He, 2021). Together with large-scale engineering efforts (Oquab et al.,

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.

2024; Siméoni et al., 2025), these advances have established SSL as a cornerstone for training modern vision foundation models.

More recently, Joint Embedding Predictive Architectures (JEPAs) have emerged as a promising extension of joint embedding methods. Instead of directly aligning latent representations, JEPAs learn to predict masked latent representations from visible context. Representative examples include I-JEPA (Assran et al., 2023) and V-JEPA (Bardes et al., 2024; Assran et al., 2025). One of the key challenges in JEPA is preventing representation collapse while preserving informative latent representations. To this end, LeJEPA introduces Sketched Isotropic Gaussian Regularization (SIGReg) (Balestriero & LeCun, 2025), which regularizes the embedding space toward an isotropic Gaussian distribution and improves training stability. In contrast, most existing JEPA models, including I-JEPA and V-JEPA, follow the student–teacher paradigm of BYOL and DINO, relying on stop-gradient operations and exponential moving average (EMA) updates to avoid representation collapse.

Independently, brain-inspired representation learning methods such as PhiNets (Ishikawa et al., 2025) have been proposed. PhiNets is motivated by the biological circuitry of the hippocampus and neocortex, drawing inspiration from the temporal predictive hypothesis (Chen et al., 2024b) and the Complementary Learning Systems (CLS) theory (McClelland et al., 1995). Specifically, PhiNets employ Siamese *student* encoders together with an EMA-based target network. Within this framework, the Siamese student encoders are intended to model the temporal predictive hypothesis, while the student–teacher learning mechanism based on exponential moving average (EMA) can be interpreted as a computational analogue of fast and slow learning in the CLS theory. More recently, Transformer-based extensions of PhiNets have also been proposed (Yamada et al., 2025), which learn to predict the next frame in latent space and can be regarded as a JEPA method.

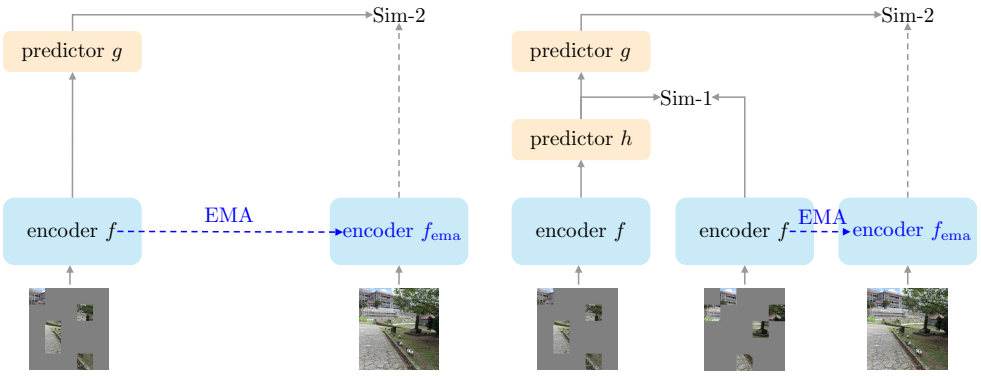
These approaches share architectural similarities with recent masked prediction methods, including SiamMAE (Gupta et al., 2023), CropMAE (Eymaël et al., 2024), and RSP (Jang et al., 2024), all of which employ Siamese encoders to learn from video data. However, unlike JEPA-based methods, these approaches are primarily designed for pixel- or feature-level reconstruction objectives rather than latent-space prediction. Although they have demonstrated strong performance on dense prediction tasks such as segmentation and pose estimation, their effectiveness for learning general-purpose visual representations remains less understood. More importantly, despite the increasing adoption of Siamese encoders in recent self-supervised learning methods, their role within latent predictive architectures such as JEPA has not been systematically investigated.

What role do Siamese student encoders play in Joint Embedding Predictive Architectures (JEPAs)?

In this paper, we investigate the role of Siamese student encoders in a JEPA framework. To this end, we propose SiamJEPA, a masked latent prediction architecture based on Siamese student encoders. Unlike existing JEPA models, SiamJEPA employs two independently masked views and learns to predict the latent representations of masked tokens from their unmasked counterparts without relying on pixel reconstruction. Furthermore, the proposed framework introduces a regularization parameter that controls the influence of the Siamese student encoders, with the conventional single-encoder JEPA recovered as a special case. Moreover, to efficiently train the SiamJEPA model, we propose the random shuffle teacher (RST) and the semantic-to-spatial curriculum learning. We evaluate SiamJEPA on the ImageNet linear-probing benchmark with a ViT-Base backbone (Dosovitskiy et al., 2021), comparing against JEPA-like methods and Masked Autoencoders (MAE) (He et al., 2022).

Our contributions are summarized as follows:

- We propose SiamJEPA, a JEPA framework with Siamese student encoders.
- We introduce the Random Shuffle Teacher (RST) and a semantic-to-spatial curriculum that first learns spatially invariant semantic representations and restores spatial correspondence.
- We provide the first systematic study of Siamese student encoders in JEPA, showing that they effectively regularize the representation space and accelerate convergence over single-encoder JEPA.
- SiamJEPA is highly training-efficient, matching MAE with fewer epochs and achieving 74.2% ImageNet linear-probing accuracy under I-JEPA's protocol, outperforming I-JEPA (72.9%) and DSeq-JEPA (73.5%) with fewer training epochs (450 vs. 600).



(a) A JEPA architecture.

(b) SiamJEPA architecture (a.k.a., Phinet architecture (Yamada et al., 2025)).

Figure 1: JEPA and SiamJEPA architectures. Sim-1 is a loss function to align the Siamese encoders. The dashed line represents the StopGradient operator. In our implementation, the two masking sets are disjoint. Note that the SiamJEPA architecture is inspired by the brain-inspired representation learning called PhiNet.

2 Related Work

This section reviews Joint Embedding Predictive Architectures (JEPAs). Due to space limitations, we provide an extended discussion of related work on joint-embedding architectures and masked modeling in the Appendix.

Masked modeling without pixel reconstruction: Building on this idea, JEPAs were proposed to predict masked representations directly in latent space rather than reconstructing pixels (LeCun et al., 2022), and have since been extended from images (I-JEPA) (Assran et al., 2023) to videos (V-JEPA) (Assran et al., 2025) and, more recently, world-model learning via V-JEPA 2. Both I-JEPA and V-JEPA use an exponential moving average (EMA) teacher network to prevent collapse. LeJEPA (Balestriero & LeCun, 2025) instead introduces Sketched Isotropic Gaussian Regularization (SIGReg), encouraging encoder outputs toward an isotropic Gaussian by minimizing a sliced Wasserstein distance to a Gaussian reference. Recently, Wu et al. (2026) proposed VISReg, an alternative to SIGReg that separately regularizes the center, scale, and shape of the representation distribution. Latent prediction's key advantage is that it focuses on semantic information, often learning useful representations more efficiently than pixel reconstruction.

PhiNetv2 (Yamada et al., 2025) was proposed as a latent-prediction counterpart of Siamese masked architectures, predicting future latent representations rather than pixels and showing strong video representation-learning performance. Several questions remain open, however: its effectiveness for image representation learning is largely unexplored, Siamese encoders have received relatively little attention in JEPA-style architectures, and existing masked representation learning methods generally assume carefully designed context masking is essential. SiamJEPA is intrinsically a masked PhiNetv2 model, and this paper investigates the role of Siamese student encoders in a JEPA.

The idea behind JEPA is rooted in both classical and modern self-supervised learning. Earlier frameworks such as Information Maximization (IMAX) (Zemel & Hinton, 1990) and its predictive counterpart, Predictability Maximization (PMAX) (Schmidhuber & Prelinger, 1993), share JEPA's core principle that useful representations should capture information predictable from related observations. JEPA can thus be viewed as a modern realization of these ideas, enabled by transformer architectures and large-scale optimization.

Curriculum learning for JEPA: DSeq-JEPA (He et al., 2026) introduces a curriculum-like progression by sequentially predicting regions in order of visual importance; our curriculum instead operates across training stages, first encouraging spatially invariant semantic representations via RST, then restoring spatial correspondence to refine them. DSeq-JEPA builds on single-encoder I-JEPA with a curriculum over region order, whereas ours introduces a semantic-to-spatial curriculum within a Siamese JEPA – the two approaches are complementary rather than mutually exclusive. In principle, the region-level curriculum of DSeq-JEPA could be combined with our semantic-to-spatial curriculum, potentially leading to further improvements in representation quality.

3 JEPA WITH SIAMESE ENCODERS (SIAMJEPA)

In this section, we propose the SiamJEPA model, which is inspired by the brain-inspired representation learning method called PhiNet architecture with Transformer encoder (Yamada et al., 2025). Figure 1 shows the architecture of both JEPA and SiamJEPA architectures. The key difference from the I-JEPA model is the existence of an additional encoder in the student network.

3.1 SIAMESE STUDENT ENCODERS

We employ Siamese student encoders with masking augmentation applied independently to each encoder. Let $\boldsymbol{X} \in \mathbb{R}^{m \times d_{\text{in}}}$ denote the patchified input image and m is the number of image patches. The outputs of the student and teacher vision transformer encoders are given by

$$\boldsymbol{H}^{(1)} = f(\text{Mask}(\boldsymbol{X}, M_1)),\ \boldsymbol{H}^{(2)} = f(\text{Mask}(\boldsymbol{X}, M_2)),\ \boldsymbol{Y} = f_{\text{ema}}(\boldsymbol{X}),$$

where $\boldsymbol{H}^{(1)} \in \mathbb{R}^{N \times (m_1+1) \times d}, \boldsymbol{H}^{(2)} \in \mathbb{R}^{N \times (m_2+1) \times d}, \boldsymbol{Y} \in \mathbb{R}^{N \times (m+1) \times d}$ are the output representations, N is the batch size, $m_1 = |M_1|$ and $m_2 = |M_2|$ are the number of unmasked tokens, M_1 and M_2 are the masking indices, and d is the embedding dimension. Here, $f(\cdot)$ denotes the Siamese student encoder with shared weights, while $f_{\text{ema}}(\cdot)$ denotes the teacher encoder whose parameters are updated using an exponential moving average (EMA). The function $\text{Mask}(\boldsymbol{X}, M)$ denotes the masking operator. In our implementation, the two masking sets are disjoint, i.e., $M_1 \cap M_2 = \emptyset$. This non-overlapping masking strategy is crucial for preventing shortcut learning in SiamJEPA. Positional embeddings are added to the input patches before they are fed into the encoders.

We further decompose the encoder outputs as $\boldsymbol{H}^{(1)} = [\boldsymbol{H}^{(1)}_{\text{cls}}; \boldsymbol{H}^{(1)}_{\text{patch}}], \boldsymbol{H}^{(2)} = [\boldsymbol{H}^{(2)}_{\text{cls}}; \boldsymbol{H}^{(2)}_{\text{patch}}], \boldsymbol{Y} = [\boldsymbol{Y}_{\text{cls}}; \boldsymbol{Y}_{\text{patch}}]$, where $[\cdot, ; \cdot]$ denotes token-wise concatenation, $\boldsymbol{H}_{\text{cls}} \in \mathbb{R}^{N \times 1 \times d}$ denotes the CLS token representation, and $\boldsymbol{H}_{\text{patch}} \in \mathbb{R}^{N \times m \times d}$ denotes the patch token representations.

3.2 PREDICTOR NETWORKS

We employ two predictor networks, denoted by $h(\cdot)$ and $g(\cdot)$. The predictor $h(\cdot)$ is responsible for aligning the global representations produced by the Siamese student encoders, whereas the predictor g predicts the latent representations of the masked tokens. This design follows the key idea of the PhiNet architecture (Ishikawa et al., 2025; Yamada et al., 2025).

For the predictor $h(\cdot)$, we employ the linear model as $h(\boldsymbol{H}) = \boldsymbol{H}\boldsymbol{W}$, where $\boldsymbol{W} \in \mathbb{R}^{d \times d}$ is a linear transform to help aligning the two Siamese student encoders.

For the predictor $g(\cdot)$, we adopt the probabilistic predictor architecture of Jang et al. (2024), which is inspired by Denton & Fergus (2018) and Hafner et al. (2021). Specifically, we define the posterior and prior distributions as

$$\text{Posterior:}\quad \boldsymbol{Z}^{(1)} \sim q(\boldsymbol{Z}^{(1)} \mid h(\boldsymbol{H}^{(1)}_{\text{cls}}), \boldsymbol{H}^{(2)}_{\text{cls}}),\ \ \text{Prior:}\quad \widehat{\boldsymbol{Z}}^{(1)} \sim p(\widehat{\boldsymbol{Z}}^{(1)} \mid h(\boldsymbol{H}^{(1)}_{\text{cls}})).$$

The posterior incorporates information from both Siamese branches, whereas the prior is conditioned only on the representation of the first branch. We also denote the flipped version of $h(\boldsymbol{H}^{(1)}_{\text{cls}})$ and $\boldsymbol{H}^{(2)}_{\text{cls}}$ as $\boldsymbol{Z}^{(2)}$ and $\widehat{\boldsymbol{Z}}^{(2)}$, respectively. Consequently, the latent variable $\boldsymbol{Z}$ models the uncertainty of the latent representation given the available context. $q(\cdot)$ and $p(\cdot)$ denote the posterior and prior distributions, respectively, both parameterized by two-layer neural networks. Note that before constructing these distributions, we first apply a projector head with batch normalization.

For the predictor $g(\cdot)$, we employ a transformer model as

$$\widehat{\boldsymbol{Y}}^{(1)} = g(h(\boldsymbol{H}^{(1)}), \boldsymbol{Z}^{(1)}) \in \mathbb{R}^{N \times (m+1) \times d} \ \text{ and } \ \widehat{\boldsymbol{Y}}^{(2)} = g(h(\boldsymbol{H}^{(2)}), \boldsymbol{Z}^{(2)}) \in \mathbb{R}^{N \times (m+1) \times d},$$

where $\widehat{\boldsymbol{Y}}^{(1)} = [\widehat{\boldsymbol{Y}}^{(1)}_{\text{cls}}; \widehat{\boldsymbol{Y}}^{(1)}_{\text{patch}}] \in \mathbb{R}^{N \times (m+1) \times d}$ and $\widehat{\boldsymbol{Y}}^{(2)} = [\widehat{\boldsymbol{Y}}^{(2)}_{\text{cls}}; \widehat{\boldsymbol{Y}}^{(2)}_{\text{patch}}] \in \mathbb{R}^{N \times (m+1) \times d}$.

3.3 LOSS FUNCTIONS

We optimize two complementary objectives: (Sim-1) aligning the global representations of the Siamese student encoders, and (Sim-2) predicting the latent representations of the masked tokens.

Specifically, Sim-1 is optimized using the Kullback–Leibler (KL) divergence, whereas Sim-2 is optimized using the normalized mean squared error (NMSE).

Sim-1:

$$\mathrm{KL}^{(1)} = \mathrm{KL}(q(\boldsymbol{Z}|[h(\boldsymbol{H}_{\text{cls}}^{(1)}), \boldsymbol{H}_{\text{cls}}^{(2)}])\|p(\boldsymbol{Z}|h(\boldsymbol{H}_{\text{cls}}^{(1)}))),$$
$$\mathrm{KL}^{(2)} = \mathrm{KL}(q(\boldsymbol{Z}|[h(\boldsymbol{H}_{\text{cls}}^{(2)}), \boldsymbol{H}_{\text{cls}}^{(1)}])\|p(\boldsymbol{Z}|h(\boldsymbol{H}_{\text{cls}}^{(2)}))).$$

The posterior distribution is conditioned on both views, whereas the prior distribution is conditioned on only one view. Therefore, minimizing the KL divergence encourages the prior distribution inferred from a single view to approximate the posterior distribution inferred from both views. Consequently, the latent variable obtained from one view becomes predictive of the latent representation that would otherwise require information from both views, thereby encouraging the encoder to learn representations that are robust to view variations.

We also found that stopping the gradient through the prior branch consistently improves optimization stability and downstream performance. Therefore, we employ the following objective:

$$\mathrm{KL}_{\text{sg}}^{(1)} = \mathrm{KL}(q(\boldsymbol{Z}|[h(\boldsymbol{H}_{\text{cls}}^{(1)}), \boldsymbol{H}_{\text{cls}}^{(2)}])\|\mathrm{sg}(p(\boldsymbol{Z}|h(\boldsymbol{H}_{\text{cls}}^{(1)})))),$$
$$\mathrm{KL}_{\text{sg}}^{(2)} = \mathrm{KL}(q(\boldsymbol{Z}|[h(\boldsymbol{H}_{\text{cls}}^{(2)}), \boldsymbol{H}_{\text{cls}}^{(1)}])\|\mathrm{sg}(p(\boldsymbol{Z}|h(\boldsymbol{H}_{\text{cls}}^{(2)})))),$$

where $\mathrm{sg}(\cdot)$ is the stopgradient operator. This heuristic was originally introduced empirically by Yamada et al. (2025).

This probabilistic formulation was originally introduced for future-frame prediction (Denton & Fergus, 2018). In this work, we adopt the same principle to encourage consistency between two augmented views. By requiring the prior inferred from a single view to match the posterior inferred from both views, the encoder is encouraged to capture information that is shared across views while discarding view-specific variations. This complements the representation learning objective of SiamJEPA and promotes more invariant latent representations.

Following PhiNet v2 (Yamada et al., 2025), we adopt the KL divergence for the Siamese student encoder. Although other objectives, such as cosine similarity or mean squared error in latent space, could also be considered, the probabilistic formulation provides an intuitive interpretation of latent consistency and demonstrated strong empirical performance in our preliminary experiments.

Sim-2:

$$\mathrm{MSE}^{(\ell)} = \frac{1}{|\bar{M}|}\|\mathrm{Mask}(\boldsymbol{Y}_{\text{patch}}^{(\ell)} - \widehat{\boldsymbol{Y}}_{\text{patch}}^{(\ell)}, \bar{M})\|_{\text{Frob}}^2,\ \ell \in \{1, 2\}, \tag{1}$$

where $\|\cdot\|_{\text{Frob}}$ is the Frobenius norm and $\bar{M} = M \setminus (M_1 \cup M_2)$ and $M = \{1, 2, \ldots, m\}$ as the set of index of patches.

Final loss function is given as

$$L = \frac{1}{2}(\mathrm{MSE}^{(1)} + \mathrm{MSE}^{(2)}) + \frac{\lambda_{\text{KL}}}{2}(\mathrm{KL}_{\text{sg}}^{(1)} + \mathrm{KL}_{\text{sg}}^{(2)}).$$

Note that if we set $\lambda_{\text{KL}} = 0$, the SiamJEPA model is similar to that of single encoder JEPA model. Since our model uses the special decoder depends on the posterior distribution, we name the SiamJEPA model with $\lambda_{\text{KL}} = 10^{-5}$ as a JEPA-like method.

3.4 Random Shuffle Teacher (RST)

A key objective of JEPA is to align student and teacher representations at the patch level. While this patch-wise prediction encourages the model to capture spatially structured information, it may also allow the prediction task to rely heavily on spatial correspondence between student and teacher tokens. To encourage each patch representation to capture more global semantic information, we propose the *Random Shuffle Teacher* (RST), which randomly permutes the teacher patch tokens before computing the prediction objective.

The key intuition behind RST is that random permutation breaks the direct spatial correspondence between student and teacher tokens. Consequently, the student cannot reliably predict the target

Table 1: ImageNet linear probing performance comparison. The results demonstrate that enforcing consistency between the Siamese student encoders improves the quality of the learned representations, resulting in higher ImageNet linear probing performance. Note that MAE is evaluated using the CLS token, whereas SiamJEPA is evaluated using mean pooling. † We conduct the experiments using the official implementation available at `https://github.com/facebookresearch/mae`. Results for the original I-JEPA and Dseq-JEPA are taken from (Assran et al., 2023) and (He et al., 2026) (600 training epochs), respectively. ‡ Same checkpoint as the row above (total epoch 450), evaluated with the I-JEPA linear-probing protocol (avg-pooled/4-layer-concat features, best of a 12-head grid) instead of 12th-layer mean pooling.

Method	Epochs	**Top-1 Acc. (%)**
MAE†	400	61.9
MAE (He et al., 2022)	1600	68.0
CAE (Chen et al., 2024a)	1600	70.4
I-JEPA (Assran et al., 2023)	600	72.9
Dseq-JEPA (He et al., 2026)	600	73.5
SiamJEPA with RST (12th layer)	450	73.3
SiamJEPA with RST (I-JEPA protocol)	450‡	**74.2**

representation based solely on positional or local spatial information. To minimize the prediction objective under this perturbation, the model is encouraged to encode semantic information that is shared across patches, rather than relying primarily on their spatial locations. In this way, RST acts as a simple regularization mechanism that promotes the distribution of global semantic information across individual patch representations.

Building on this observation, we further propose an RST-based curriculum learning strategy. Specifically, we first train the model with RST to encourage the emergence of globally semantic patch representations. In the later stages of training, we remove the random shuffling and restore the original spatial correspondence between student and teacher tokens. This allows the model to subsequently refine spatially structured representations while retaining the semantic representations acquired during the initial RST stage.

4 Experiments

In this section, we conduct the ablation study of SiamJEPA and compared its performance with the masked autoencoder (MAE) (He et al., 2022), context autoencoder (Chen et al., 2024a), I-JEPA (Assran et al., 2023), and Dseq-JEPA (He et al., 2026). The main purpose of this experiment is not to achieve the state of the art performance in JEPA.

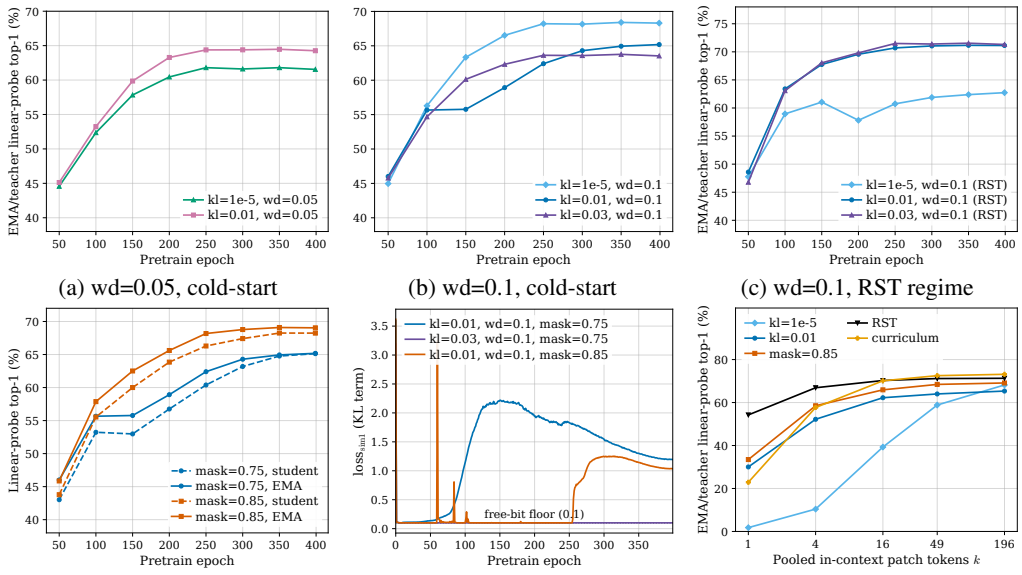
4.1 Setup and implementation details

All ablation studies are conducted under the same experimental setting (Table 2). We employ an EMA momentum schedule that increases from 0.99 during epochs 1–200, to 0.999 during epochs 201–300, and to 0.9999 during epochs 301–400.

All experiments are performed using NVIDIA V100, A100, or H100 GPUs. Based on the findings from these studies, we select the hyperparameters and train the final models for 400 pre-training epochs for the main comparisons.

Note that our implementation is built upon the official MAE codebase [1] rather than the official I-JEPA implementation. Consequently, although the training dynamics differ from those reported for I-JEPA, our implementation provides a common experimental framework for comparing reconstruction-based and JEPA-based methods.

[1] `https://github.com/facebookresearch/mae`



(a) wd=0.05, cold-start (b) wd=0.1, cold-start (c) wd=0.1, RST regime

(d) mask_ratio (kl=0.01, wd=0.1) (e) Training loss (loss$_{\text{sim1}}$) (f) Class information per token

Figure 2: Effect of the weight decay and mask ratio (cold-start unless noted). (a)–(d): EMA/teacher linear-probe top-1 vs. pretrain epoch; (e): the corresponding loss$_{\text{sim1}}$ (KL term) training dynamics for the configs in (b) and (d), bounded below by the free-bit floor of 0.1; (f): top-1 when only k random patch tokens of the clean image are mean-pooled after the full encoder (EMA/teacher, wd=0.1, $\lambda_{\text{KL}} = 0.01$ and mask_ratio=0.75 unless noted; RST at epoch 399; curriculum = init←RST ep200 at 450 epochs; Appendix B.2). (b) includes an intermediate free-bit scale ($\lambda_{\text{KL}} = 0.03$); (d) uses dashed=student, solid=EMA/teacher.

4.2 Comparison to other SSL methods

As shown in Table 1, SiamJEPA achieves performance comparable to that of MAE and CAE while requiring substantially fewer training epochs. In particular, SiamJEPA outperforms MAE using less than one-quarter of the training epochs and achieves performance comparable to CAE, demonstrating significantly improved training efficiency.

Compared with I-JEPA (Assran et al., 2023) and DSeq-JEPA (He et al., 2026), SiamJEPA without curriculum learning achieves lower final linear-probing performance. With the RST-based curriculum, however, SiamJEPA largely closes this gap. Specifically, training with RST for the first 200 epochs and then continuing with the standard non-shuffled SiamJEPA objective for another 250 epochs achieves 73.3% EMA/teacher linear-probing accuracy after 450 epochs in total. (Table 3). This result is comparable to the 72.9% reported by I-JEPA and the 73.5% reported by DSeq-JEPA, both after 600 training epochs, while requiring substantially fewer training epochs. Notably, unlike I-JEPA and DSeq-JEPA, SiamJEPA relies on a much simpler masking strategy, suggesting that the semantic-to-spatial curriculum can provide an effective alternative to more sophisticated masking or prediction schemes. Since Table 1's 12th-layer mean-pooling protocol differs from I-JEPA's own evaluation, we additionally re-evaluate the same checkpoint using I-JEPA's official linear-probing protocol (avg-pooled/4-layer-concat features, best of a 12-head grid): SiamJEPA reaches 74.2% top-1, exceeding both I-JEPA (72.9%) and DSeq-JEPA (73.5%) even under a matched evaluation protocol, again with substantially fewer training epochs (450 vs. 600).

4.3 Effect of Siamese student encoder

To investigate the role of the Siamese student encoders, we conduct an ablation study on the KL regularization weight λ_{KL}, weight decay, and masking ratio. Figure 2 summarizes the results without RST (mask ratio $= 0.75$; panels (a)–(b)). With the default weight decay of 0.05, increasing λ_{KL} from 10^{-5} to 0.01 consistently improves linear-probing accuracy. Interestingly, this trend reverses with stronger weight decay ($\text{wd} = 0.1$): at epoch 399, $\lambda_{\text{KL}} = 10^{-5}$ outperforms $\lambda_{\text{KL}} = 0.01$. Panel (b) shows that linear-probing accuracy continues to improve under weak regularization, reaching

68.3% by epoch 399, even as panel (e) shows its KL loss steadily increasing, suggesting that stronger student–student agreement is not always beneficial.

In contrast, strong KL regularization substantially improves performance when combined with RST (panel (c)), suggesting that Siamese consistency becomes particularly important under RST. Panel (d) further shows that increasing the mask ratio from 0.75 to 0.85 avoids the same saturation behavior, while the KL loss remains close to the free-bit floor, consistent with a smaller discrepancy between the two student views.

Overall, the benefit of Siamese consistency depends strongly on the training regime. Strong KL regularization is particularly effective with RST, but can be unnecessary or even detrimental when sufficient consistency is already maintained. These results suggest that the Siamese objective serves as a complementary consistency constraint rather than simply maximizing agreement between the student encoders.

4.4 Random shuffle teacher experiments

As shown in Figure 2(a)–(b), SiamJEPA's linear-probing performance without curriculum learning falls short of state-of-the-art methods, possibly because final-layer patch representations retain too much spatial and local information at the expense of semantics. Under RST (panel (e)), despite removing the teacher's spatial correspondence entirely, linear-probing performance is comparable to or better than the non-shuffled counterpart, suggesting that exact spatial correspondence is not necessary and that RST instead encourages more semantic patch representations.

We verify this semantic bias directly with a linear probe that predicts a frozen final-layer patch token's original grid position out of 196 classes (Appendix C, Table 3): SiamJEPA without RST retains almost perfect positional information (94.3%, chance 0.51%), whereas RST loses most of it (18.7%, still $37\times$ above chance), consistent with a shift toward DINO/BYOL-style global-content encoding. Figure 3 tracks this along the RST-based curriculum: positional decodability (b) stays low throughout RST training and recovers substantially once the objective switches back, at either switching point (epoch 200 or 399), while the corresponding ImageNet linear-probe accuracy (a) continues to improve. The same bias appears in test-time robustness (Figure 4): with only 10% of patch tokens retained, RST retains 52.0% top-1 versus 41.3% without RST, showing RST relies less on specific spatial locations, though it is not completely position-agnostic.

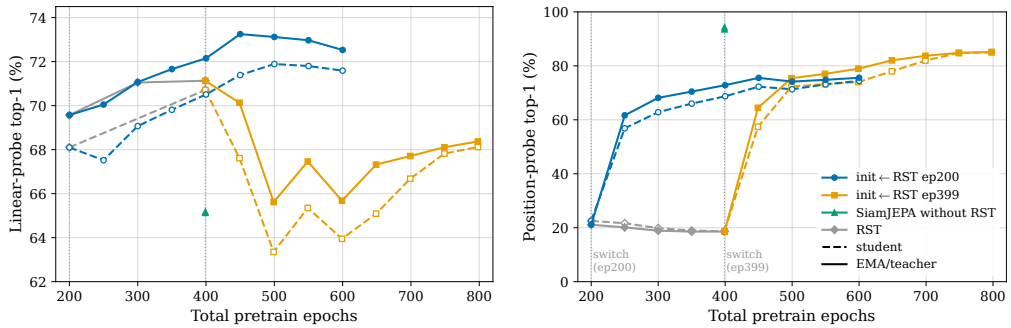
Under 50% image occlusion, the curriculum model (switched at epoch 200) is most robust (49.5%), ahead of RST-only (43.5%) and SiamJEPA without RST (34.7%), indicating that the curriculum combines RST's robustness with the spatial information recovered afterward.

This bias is even visible at the single-token level (Figure 2(f)): a single random patch token already yields 54.2% top-1 under RST, versus 30.0% without RST, while a weak KL term ($\lambda_{\mathrm{KL}} = 10^{-5}$) makes a single token almost uninformative (1.8%, despite 68.2% once averaged over all 196 tokens) – indicating that RST and strong Siamese consistency independently push class information down into individual tokens. Further results are in Appendix C.

4.5 Random shuffle curriculum learning

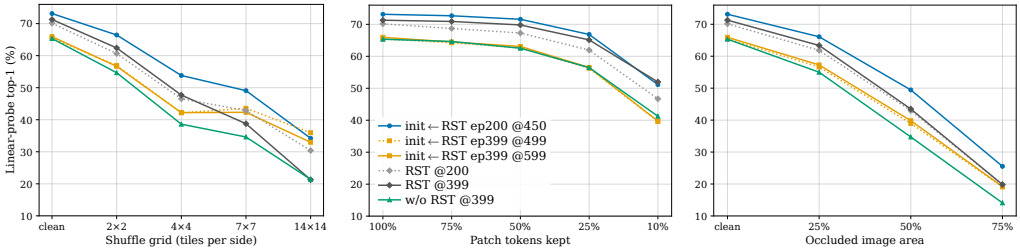
One of our key findings is that RST rapidly improves linear probing performance during the early stages of training. However, its improvement tends to saturate in later epochs. In our experiments, using a larger weight decay alleviates this saturation and improves performance under longer training, but an additional mechanism appears necessary to further improve SiamJEPA. In contrast, the non-shuffled counterpart preserves spatial correspondence between teacher and student tokens, which may be beneficial for further refining the learned representations.

Figure 3 (a) reports the results of this curriculum learning strategy. As can be seen, the subsequent non-shuffled training continues to improve the representation rather than exhibiting the saturation observed with RST alone. These results suggest that semantic representation learning induced by RST and spatially aligned representation learning induced by the standard JEPA objective play complementary roles at different stages of training. As Figure 3 (b) above shows, positional information is lost under RST and recovered after switching back, but matching curricula on total training budget alone does



(a) ImageNet linear probe (b) Position decodability

Figure 3: Position decodability and ImageNet linear-probe top-1 of the same checkpoints along the RST-based curriculum. Dotted vertical lines mark the switch from RST to the SiamJEPA-without-RST objective (epoch 200 or 399). Dashed = student, solid = EMA/teacher. The position probe uses 2,000/500 images and a single seed (roughly ± 1 point).



(a) Patch shuffle (b) Token drop (random) (c) Occlusion (mean-colour block)

Figure 4: Test-time robustness of the trained linear probes (EMA/teacher encoders, 10,000 ImageNet val images, no retraining; every random choice is shared across encoders). Patch shuffle permutes tiles of a $g \times g$ grid (14×14 is a pure patch-token permutation); token drop feeds only a random fraction of patch tokens to the encoder; occlusion fills a square block of the given image area at a random location with the mean colour.

not equalize final ImageNet accuracy, indicating that *when* RST is switched off matters beyond how long it runs.

Interestingly, while the original SiamJEPA without curriculum learning underperforms I-JEPA and Dseq-JEPA with ViT-B, the proposed curriculum enables SiamJEPA to outperform I-JEPA in linear probing. Although our results are based on a single run and may therefore be affected by experimental randomness, they suggest that SiamJEPA can match or potentially outperform strong JEPA baselines while using a substantially simpler masking strategy.

5 Conclusion

In this paper, we investigated the role of Siamese student encoders within the Joint Embedding Predictive Architecture (JEPA) framework. To this end, we proposed SiamJEPA, a JEPA-based representation learning framework that incorporates Siamese student encoders together with an exponential moving average (EMA) teacher network. SiamJEPA processes two independently masked views using Siamese student encoders and learns to predict latent representations of masked target regions. Through extensive ablation studies on ImageNet linear probing with a ViT-Base backbone, we systematically investigated the effects of key design choices, including KL regularization, masking strategies, weight decay, and the free-bit threshold. Our results demonstrate that incorporating Siamese student encoders consistently improves representation quality over a corresponding JEPA-like baseline. Moreover, we find that the Siamese architecture serves as an effective regularizer, leading to more discriminative latent representations and faster convergence during the early stages of training. Overall, our findings provide new insights into the role of Siamese student encoders in JEPA and offer practical guidance for designing and optimizing future JEPA-style representation learning methods.

AI USE STATEMENT

We used ChatGPT and Claude Code to assist with implementation and debugging, experiments including hyperparameter tuning, literature searches, manuscript drafting and revision, and figure preparation. The authors take full responsibility for the final manuscript, code, references, experimental results, and reported findings.

REFERENCES

Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Mike Rabbat, and Nicolas Ballas. Masked siamese networks for label-efficient learning. In *ECCV*, 2022.

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*, 2023.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

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

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *ICLR*, 2022.

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

Suzanna Becker and Geoffrey E Hinton. Self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, 355(6356):161–163, 1992.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020a.

Xiaokang Chen, Mingyu Ding, Xiaodi Wang, Ying Xin, Shentong Mo, Yunhao Wang, Shumin Han, Ping Luo, Gang Zeng, and Jingdong Wang. Context autoencoder for self-supervised representation learning. *International Journal of Computer Vision*, 132(1):208–223, 2024a.

Xinlei Chen and Kaiming He. Exploring simple Siamese representation learning. In *CVPR*, 2021.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020b.

Yusi Chen, Huanqiu Zhang, Mia Cameron, and Terrrence Sejnowski. Predictive sequence learning in the hippocampal formation. *Neuron*, 112:2645–2658, 2024b.

Ninad Daithankar, Alexi Gladstone, Yann LeCun, and Heng Ji. You don't need strong assumptions: Visual representation learning via temporal differences. *arXiv preprint arXiv:2606.15956*, 2026.

Emily Denton and Rob Fergus. Stochastic video generation with a learned prior. In *ICML*, 2018.

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 *ICLR*, 2021.

Alexandre Eymaël, Renaud Vandeghen, Anthony Cioppa, Silvio Giancola, Bernard Ghanem, and Marc Van Droogenbroeck. Efficient image pre-training with Siamese cropped masked autoencoders. In *ECCV*, 2024.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent-a new approach to self-supervised learning. In *NeurIPS*, 2020.

Agrim Gupta, Jiajun Wu, Jia Deng, and Fei-Fei Li. Siamese masked autoencoders. *NeurIPS*, 2023.

Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering Atari with discrete world models. In *ICLR*, 2021.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022.

Xiangteng He, Shunsuke Sakai, Shivam Chandhok, Sara Beery, Kun Yuan, Nicolas Padoy, Tatsuhito Hasegawa, and Leonid Sigal. DSeq-JEPA: Discriminative sequential Joint-Embedding predictive architecture. *ECCV*, 2026.

Satoki Ishikawa, Makoto Yamada, Han Bao, and Yuki Takezawa. PhiNets: Brain-inspired non-contrastive learning based on temporal prediction hypothesis. In *ICLR*, 2025.

Huiwon Jang, Dongyoung Kim, Junsu Kim, Jinwoo Shin, Pieter Abbeel, and Younggyo Seo. Visual representation learning with stochastic frame prediction. In *ICMLR*, 2024.

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

James L McClelland, Bruce L McNaughton, and Randall C O'Reilly. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, 102(3):419, 1995.

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, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Pratyaksh Rao, Wancong Zhang, Randall Balestriero, Yann LeCun, and Giuseppe Loianno. Skyjepa: Learning long-horizon world models for zero-shot sim-to-real control of quadrotors. *arXiv preprint arXiv:2606.23444*, 2026.

Jürgen Schmidhuber and Daniel Prelinger. Discovering predictable classifications. *Neural Computation*, 5(4):625–635, 1993.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NIPS*, 2017.

Haiyu Wu, Randall Balestriero, and Morgan Levine. VISReg: Variance-Invariance-Sketching regularization for JEPA training. *arXiv*, 2026. doi: 10.48550/arXiv.2606.02572. URL `https://arxiv.org/pdf/2606.02572.pdf`.

Makoto Yamada, Kian Ming A Chai, Ayoub Rhim, Satoki Ishikawa, Mohammad Sabokrou, and Yao-Hung Hubert Tsai. Brain-inspired stochastic joint embedding representation learning. *arXiv preprint arXiv:2505.11129*, 2025.

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

Richard Zemel and Geoffrey E Hinton. Discovering viewpoint-invariant relationships that characterize objects. *NIPS*, 1990.

Table 2: Pretraining and linear probing configurations.

Setting	MAE	SiamJEPA
Backbone	ViT-Base	ViT-Base
Decoder depth	8	1
Pretrain Dataset	ImageNet-1K	ImageNet-1K
Pretrain Epochs	400	400
Input Resolution	224^2	224^2
Mask Ratio	0.75	{0.75,0.85}
Optimizer	AdamW	AdamW
Base Learning Rate	1.5e-4	1.5e-4
Weight Decay	0.05	{0.05,0.1}
Effective Batch Size	4096	8192
EMA	No	{0.99, 0.999, 0.9999}
Linear Probe Epochs	90	90
Linear Probe Resolution	224^2	224^2

A Extended related work

Joint embedding architecture: One of the most popular SSL approaches is contrastive learning. SimCLR (Chen et al., 2020a) learns representations by pulling together augmented views of the same image (positive pairs) while pushing apart different images (negative pairs), but relies on large negative-sample counts and thus benefits from very large batch sizes. Momentum Contrast (MoCo) (Chen et al., 2020b) alleviates this with a momentum encoder and a fixed-size negative-sample queue.

Subsequent work showed negative samples are not strictly necessary: BYOL (Grill et al., 2020) and SimSiam (Chen & He, 2021) achieve strong performance via stop-gradient operations and exponential moving averages instead. Another family relies on explicit regularization, e.g., Barlow Twins (Zbontar et al., 2021) and Variance-Invariance-Covariance Regularization (VICReg) (Bardes et al., 2022), which encourage informative, non-collapsed representations through variance and covariance constraints. Masked Siamese Networks (MSN) (Assran et al., 2022) extend BYOL to masked modeling, using a student and an EMA teacher encoder that align CLS-token embeddings, and outperform BYOL particularly in the 1% ImageNet-1K label-efficient setting.

PhiNet (Ishikawa et al., 2025), a brain-inspired SSL framework drawing on the biological circuitry of the hippocampus and neocortex, employs Siamese encoders with temporal prediction and EMA mechanisms, and is more robust to collapse than SimSiam with favorable continual-learning properties. Unlike other methods, which use a single student encoder, PhiNet uses Siamese student encoders.

Masked modeling with pixel reconstruction: Most of the aforementioned methods were originally built on CNNs such as ResNet; Transformer-based SSL has since become dominant. A seminal example is Masked Autoencoders (MAE) (He et al., 2022), which reconstruct masked image patches from visible tokens. Another line combines masked modeling with Siamese architectures: SiamMAE (Gupta et al., 2023) predicts future video frames from the current frame and a masked future observation via two encoders and a decoder, and CropMAE (Eymaël et al., 2024) and RSP (Jang et al., 2024) follow similarly, showing strong performance on downstream tasks such as pose estimation and segmentation. However, these methods remain fundamentally reconstruction-based, aiming to recover image content.

B Ablation study

B.1 Effect of weight decay

Next, we investigate the effect of weight decay by varying its coefficient, as shown in Figure 2. Consistent with the observations reported for PhiNet (Ishikawa et al., 2025), SiamJEPA is relatively robust to the choice of weight decay. At $\lambda_{\mathrm{KL}} = 0.01$, larger weight decay is particularly beneficial during longer training: with the default wd=0.05, linear probing performance saturates by epoch 300 (63.56%→63.78%→63.59% at epochs 300/350/399), whereas wd=0.1 continues improving through epoch 399 (63.19%→64.77%→65.13%), eventually overtaking wd=0.05. This benefit is

not restricted to strong KL regularization: at $\lambda_{\mathrm{KL}} = 10^{-5}$, wd=0.1 is, if anything, even more advantageous, reaching 67.53% at epoch 399 (fully converged) versus wd=0.05's final 60.84%. This is notably different from an earlier, pre-fix observation of training instability for $\lambda_{\mathrm{KL}} = 10^{-5}$ under wd=0.1 (Appendix Section C); after the `forward_encoder` mask-shuffle fix, we no longer observe this failure mode, and $\lambda_{\mathrm{KL}} = 10^{-5}$/wd=0.1 is in fact the strongest cold-start cell in the grid so far.

These results suggest that stronger weight decay improves generalization during extended pre-training for SiamJEPA, and that this benefit holds across KL regularization strengths rather than being specific to strongly KL-regularized configurations.

B.2 Per-token semantic information

To measure how much class information individual patch tokens carry, we pass the clean image through the frozen encoder and mean-pool only k randomly chosen patch tokens before applying the trained linear probe (Figure 2(f); EMA/teacher encoders, 10,000 ImageNet val images, the same random tokens for all encoders). Unlike token dropping, the attention context of all tokens is kept intact. Because the probe was trained on 196-token means, small k is out of distribution for the probe, so the numbers are best read as a ranking across encoders. RST stands out: a single patch token already yields 54.2% top-1 accuracy (76% of its clean accuracy), compared with 30.0% for SiamJEPA without RST, directly supporting the view that RST spreads global semantic information over individual patch tokens. A higher mask ratio (0.85) barely changes the per-token information (33.5%), so its robustness to occlusion has a different origin. The curriculum model reads less class information from a single token (22.9%), presumably because positional information is re-introduced into each token in the second stage, yet it matches RST with 16 tokens (70.1%). With a weak KL term ($\lambda_{\mathrm{KL}} = 10^{-5}$), single tokens are almost uninformative to the probe (1.8%) although the full average reaches 68.2%; this encoder is also the most dependent on the position embedding, which suggests large token-specific positional components that cancel only in the average.

C Random shuffle teacher ablation study

The ablations reported in this section were all conducted under the Random Shuffle Teacher (RST) training regime introduced in Section 3.4, in which the teacher's patch tokens are randomly permuted before the prediction objective is computed. These sweeps additionally use mean pooling from an intermediate (10th) transformer layer as their evaluation protocol.

C.1 Position decodability and semantic bias of RST

We test whether shuffling the teacher's patch order washes out positional information by training a linear probe on frozen final-layer patch tokens (pre-pool, no `[CLS]`) to predict the token's original grid position out of 196 classes. Table 3 compares checkpoints trained under identical hyperparameters ($\mathrm{KL} = 0.01$, $\mathrm{wd} = 0.1$), differing only in whether `forward_encoder` shuffled patch order for the teacher branch (RST) or preserved it (SiamJEPA without RST), together with two curriculum models initialized from RST at epoch 200 and 399. SiamJEPA without RST retains almost perfect positional information (94.3% top-1, chance 0.51%), while RST loses most of it (18.7% top-1) though it remains roughly $37\times$ above chance, indicating the representation is not fully scrambled. This supports the hypothesis that shuffling the teacher's patch order during pretraining biases the learned representation toward a more DINO/BYOL-style global-content encoding at the expense of spatial position.

Figure 5 visualizes the additional configurations of Table 4 across the full pretrain trajectory (student and EMA/teacher), rather than only at their final checkpoints.

C.2 Effect of Siamese student encoder for RST

To investigate the role of the Siamese student encoders for RST, we conduct an ablation study on the KL regularization weight λ_{KL}, the weight decay parameter, and the free-bit threshold used in the KL divergence computation. The KL term encourages the outputs of the two Siamese student encoders to

Table 3: Position-decodability of frozen final-layer patch tokens. A linear probe predicts which of the 196 spatial grid positions a frozen patch token was extracted from (chance $1/196 \approx 0.51\%$). All encoders use KL $= 0.01$, wd $= 0.1$, mask_ratio$= 0.75$, and differ only in whether `forward_encoder` shuffled the teacher's patch order during pretraining, and whether the run was subsequently continued under the SiamJEPA-without-RST (non-shuffled) objective. The last column repeats the ImageNet linear-probe top-1 of the very same checkpoint, so that positional and semantic quality can be read side by side.

Encoder	Branch	Epoch	Pos. top-1 (%)	Pos. top-5 (%)	IN-1k top-1 (%)
SiamJEPA without RST	student	399	**94.3**	**99.9**	65.13
	EMA/teacher	399	93.5	99.9	65.18
RST	student	399	18.7	57.6	70.71
	EMA/teacher	399	18.6	57.2	71.13
init←RST ep200	student	450	72.3	99.0	71.39
	EMA/teacher	450	75.5	99.3	**73.25**
init←RST ep399	student	599	74.0	99.1	63.95
	EMA/teacher	599	78.9	99.3	65.68

Table 4: Additional ways of leveraging the pre-fix RST (Random Shuffle Teacher) checkpoint, extending Table 3 with configurations not otherwise tabulated there (SiamJEPA without RST, RST, and init←RST ep200 @450 already appear in Table 3). init← = fresh optimizer/epoch-0 restart (`-init_checkpoint`); resume← = optimizer/epoch-preserving continuation (`-resume`).

Configuration	Epoch	Student top-1 (%)	EMA/teacher top-1 (%)
resume←RST ep200	399	71.13	71.86
init←RST ep200	599	71.59	72.53
init←RST ep399	799	68.12	68.37

become similar, while the free-bit threshold prevents the two distributions from becoming completely identical by imposing a minimum KL value.

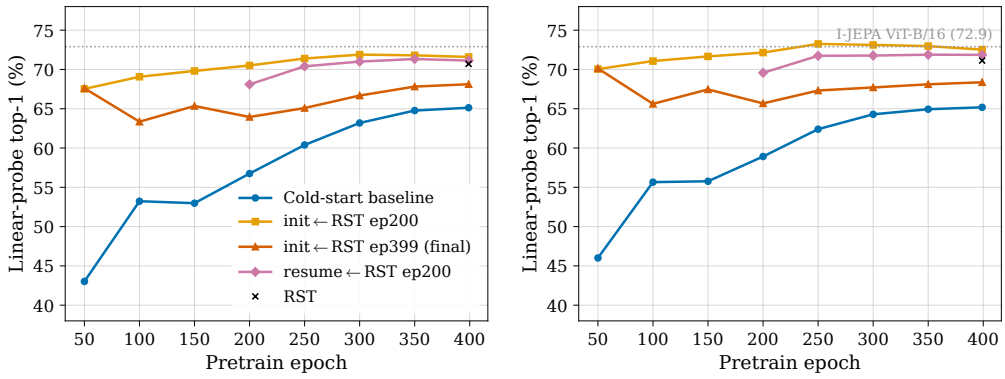
Figure 6 shows the learning curves for small and large values of λ_{KL}. As expected, a small regularization weight results in a large KL divergence (greater than 40), whereas a large regularization weight keeps the KL divergence close to the free-bit threshold of 0.1. Despite this substantial difference in the KL term, the overall learning behavior of the total loss is similar in both cases, although training with the larger regularization weight converges more slowly. These observations suggest that the primary learning signal for acquiring useful representations comes from the masked prediction objective (i.e., the Sim-2 loss), whereas the KL term mainly serves as a regularizer. In particular, the KL regularization provides a beneficial inductive bias during the early stages of training, but overly strong regularization may unnecessarily constrain optimization and slow convergence.

C.3 Predictor depth

The effect of predictor depth remains an open question. Owing to computational constraints and the large hyperparameter space, we have not yet been able to determine whether the observations described below are intrinsic properties of SiamJEPA (RST) or are specific to our experimental setting.

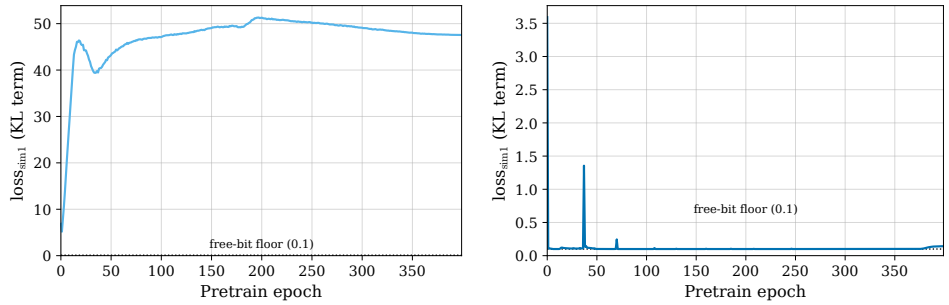
Empirically, we found that SiamJEPA (RST) achieves better linear probing performance when using a shallow Transformer predictor with only one or two layers. Throughout this paper, we therefore employ a single-layer predictor, which is substantially shallower than those commonly used in JEPA-based methods. If this observation generalizes beyond our current setting, it would suggest that SiamJEPA (RST) can achieve competitive performance with a significantly smaller predictor, leading to a more parameter-efficient architecture.

We also observed that the final linear probing performance depends on which encoder layer is used for evaluation. This suggests that the layer containing the most transferable representation may vary with



(a) Student (b) EMA / teacher

Figure 5: Ways of leveraging the pre-fix RST (Random Shuffle Teacher) checkpoint after the encoder-mask-shuffle fix, compared against SiamJEPA without RST trained from scratch (cold-start) and the uncorrected RST run itself. Dotted line: I-JEPA ViT-B/16 reference (72.9%).



(a) $\lambda_{\mathrm{KL}} = 0.00001$. (b) $\lambda_{\mathrm{KL}} = 0.01$.

Figure 6: RST-regime training-$\mathrm{loss}_{\mathrm{sim1}}$ (KL term) dynamics for $\lambda_{\mathrm{KL}} = 0.00001$ and $\lambda_{\mathrm{KL}} = 0.01$ (both wd=0.1, mask_ratio=0.75). With a small regularization coefficient, the KL divergence between the representations produced by the two Siamese student encoders remains large throughout training (rising above 40). In contrast, with a larger regularization coefficient, the KL divergence stays close to the free-bit threshold (0.1 in our experiments) for essentially the whole 399 epochs, aside from brief transient spikes.

the predictor depth, consistent with the observations in the previous section. It is possible that this behavior changes for larger models or more challenging datasets. However, thoroughly investigating this phenomenon would require substantially more computational resources than were available in this work. We therefore leave a systematic study of predictor depth and representation dynamics for future work.

C.4 Test-time robustness

Figure 4 evaluates the trained linear probes under test-time perturbations without any retraining. RST is the most robust to dropping patch tokens: with only 10% of the tokens kept, it retains 52.0% top-1 accuracy (73% of its clean accuracy), compared with 41.3% (63%) for SiamJEPA without RST, consistent with semantic information being spread across patch tokens. However, RST is not position-agnostic. Setting the patch position embedding to zero reduces its accuracy from 71.3% to 16.4%, a larger drop than for SiamJEPA without RST ($65.3\% \rightarrow 21.9\%$), and a 14×14 patch shuffle, which only reassigns position embeddings to patch contents, is as harmful for RST as for SiamJEPA without RST (21.3% vs. 21.4%). Thus, RST removes positional information from the final-layer representation, while its intermediate computation still relies on the position embedding, increasingly so with longer RST training (34.6% at epoch 200 vs. 16.4% at epoch 399 without the position embedding). Under occlusion with a mean-colour block, which keeps all tokens and is

closest to a real occluder, the curriculum model switched at epoch 200 is clearly the most robust (49.5% at 50% occluded area, vs. 43.5% for RST and 34.7% for SiamJEPA without RST).

D Discussion & Future work

Several important directions remain for future work. First, our study primarily focused on relatively small-scale architectures to better understand the role of Siamese student encoders. Whether the observed benefits extend to larger-scale models, such as ViT-Large and ViT-Huge, remains an important open question.

Second, we found that the final performance is highly sensitive to implementation details and hyperparameter choices. Although our experimental setup consistently outperformed masked autoencoder baselines, there is likely substantial room for further improvement through architectural refinements and more systematic hyperparameter optimization.

Third, this work focused exclusively on image-based pretraining. Previous studies have shown that Siamese student encoders are particularly effective for video representation learning (Gupta et al., 2023; Jang et al., 2024; Yamada et al., 2025). Moreover, as JEPA-based approaches have rapidly gained attention, several recent works have begun incorporating Siamese student encoder designs into video predictive learning frameworks (Daithankar et al., 2026; Rao et al., 2026). These developments suggest that a deeper understanding of Siamese student encoders may have broader implications beyond image representation learning, particularly for video understanding and world modeling.

SiamJEPA is inspired by the brain-inspired representation learning framework PhiNets (Ishikawa et al., 2025; Yamada et al., 2025). Exploring additional neuroscience-inspired mechanisms may further improve representation quality while providing new insights into the relationship between biological and artificial learning systems.