DiJEPA: Self-Supervised Learning via Diffusion in Latent Space

Anonymous authors
Paper under double-blind review

Abstract

Joint-Embedding Predictive Architectures (JEPAs) learn semantic representations by predicting the latent embeddings of the masked target view from the visible context. Existing approaches use either direct regression in a continuous embedding space or token-softmax prediction over quantized targets. While effective, direct regression does not explicitly model a target distribution, whereas token-based prediction performs distribution matching in a discrete space; both rely on carefully designed distribution regularization to stabilize training and improve representation quality. We introduce DiJEPA, a JEPA-style framework that formulates visual representation learning as diffusion-based prediction over an online projection of the encoder's representation space. DiJEPA jointly trains a Feature Auto-Encoder (FAE) with the encoder, mapping high-dimensional representations into low-dimensional continuous latents whose distribution is optimized throughout representation learning. A diffusion predictor then reconstructs target-view latents from student-view representations through denoising, providing a distributional prediction objective in a continuous latent space. On ImageNet-1K with a ViT-L backbone, DiJEPA achieves 83.24% top-1 linear probing accuracy under a non-distilled setting, competitive with 83.3% from DINOv3. The jointly learned FAE latents can also be used for image generation, achieving 2.74 gFID without CFG in 80 epochs of latent diffusion training. Our results establish diffusion-based prediction over evolving FAE latents as a new predictive objective alongside direct regression and distribution matching over discrete tokens for self-supervised visual learning.

1 Introduction

Self-supervised learning (SSL) has become a central paradigm for representation learning (Balestriero et al., 2023; Grill et al., 2020; Bao et al., 2022; He et al., 2021; van den Oord et al., 2017), offering a scalable way to extract general-purpose knowledge from large-scale unlabeled data and transfer it across downstream tasks. A prominent recent direction in self-supervised learning is the family of Joint-Embedding Predictive Architectures (JEPAs) (LeCun, 2022; Assran et al., 2023b;a; Balestriero & LeCun, 2025; Li et al., 2026; Mur-Labadia et al., 2026), which learn representations by predicting the latent embedding of a masked or augmented view from the visible context, without explicitly reconstructing raw signals.

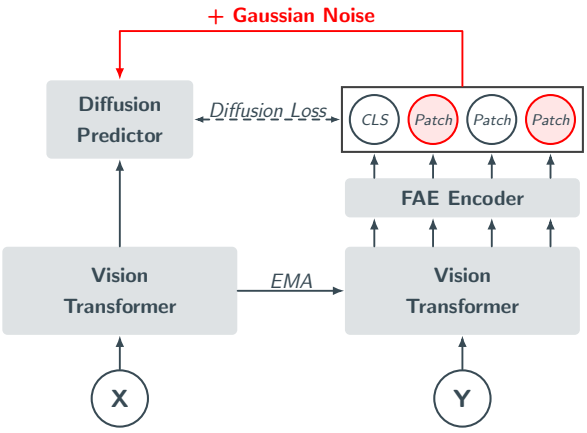

Figure 1: Diffusion Joint-Embedding Predictive Architecture.

However, a central challenge in JEPA-style learning is how to define the prediction target and training objective, as these choices determine both the collapse-prevention mechanism and the expressiveness of the learned space. Existing approaches have mostly developed along two lines. First, direct

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.

embedding prediction methods such as I-JEPA (Assran et al., 2023a) and Data2Vec (Baevski et al., 2022) train a predictor to match the representation of a masked view, typically with an L2 or smooth-L1 loss. Because continuous direct regression averages over modes, it cannot explicitly model the target distribution; when a masked region is ambiguous, the prediction collapses toward the mean of plausible representations. Methods in this family use either the stop-gradient operation and momentum teacher (Assran et al., 2023a; Baevski et al., 2022) or explicit regularizers such as VICReg (Bardes et al., 2021), SIGReg introduced by LeJEPA (Balestriero & LeCun, 2025), or RDMReg (Kuang et al., 2026) to prevent dimensional collapse. Second, quantized prediction methods such as DINO (Caron et al., 2021; Oquab et al., 2023; Sim'eoni et al., 2025) and iBOT (Zhou et al., 2021) map embeddings to tokens or cluster assignments and train different views to match the resulting softmax distributions. This formulation has proven highly effective, but its centering procedure requires fixed normalization steps such as Sinkhorn-Knopp balancing (Caron et al., 2020; Oquab et al., 2023) and KoLeo regularization (Oquab et al., 2023)

In this work, we introduce **DiJEPA** (**Di**ffusion **J**oint-**E**mbedding **P**redictive **A**rchitecture), which formulates JEPA-style learning as diffusion-based prediction over a dynamically learned, low-dimensional, and well-structured latent space. We instantiate DiJEPA with a Feature Auto-Encoder (FAE) (Gao et al., 2025), which maps high-dimensional representations into compact continuous latents. FAE adopts a single-attention-layer encoder and a representation reconstruction objective, encouraging the latent space to preserve as much information as possible from the original embeddings. Its KL regularization provides an implicit centering effect by removing redundant global information shared across patch embeddings. A denoising network then predicts the FAE latent of a target view conditioned on the context encoder's output. Because FAE is jointly trained with the backbone, the low-dimensional latent space serves as an evolving learning target whose distribution is optimized together with the representation model throughout training.

Relatedly, LeJEPA (Balestriero & LeCun, 2025) also encourages Gaussian structure in continuous self-supervised learning, but differs from DiJEPA in both its regularization objective and where the regularization is applied. LeJEPA applies SIGReg to image-level embeddings as a collapse-prevention mechanism, enabling training without an EMA teacher. In contrast, DiJEPA regularizes the learning target space across both image-level and patch-level embeddings, while leaving the backbone representations unconstrained. Our design is guided by the principle that patch representations should retain a balance between image-level information shared across patches and patch-specific information, avoiding treating individual patches as independent Gaussian samples or collapsing different patch representations toward a common image-level embedding. Accordingly, KL regularization in DiJEPAis used to shape a compact and structured continuous target space for diffusion-based prediction, making the predictive task more challenging and thereby encouraging higher-quality representations. This formulation also yields substantially stronger representation quality than LeJEPA, with DiJEPA achieving 83.24% ImageNet-1K linear-probing accuracy using a ViT-L backbone.

We apply DiJEPA in the DINOv3 training setting, where two types of augmented views are used: different crops of the same image for image-level CLS prediction, and masked image views for patch-level prediction. Our framework provides three forms of flexibility. First, FAE latents are continuous and low-dimensional, avoiding the information bottleneck of discrete tokens while remaining suitable for both understanding and generation. Second, diffusion-based prediction naturally models multimodal distributions over plausible targets, unlike point-estimate losses. Third, KL regularization in the FAE replaces fixed Sinkhorn-based centering with a learnable centering mechanism that is optimized jointly with the backbone, opening up a broader design space for JEPA-style learning targets. Building on this framework, we further explore several variants, including adaptive diffusion timestep shifting and using multiple CLS tokens in the image-level objective.

We empirically study DiJEPA on ImageNet-1K (Deng et al., 2009) using a ViT-L backbone (Dosovitskiy et al., 2021). Following standard self-supervised evaluation practice, we evaluate the frozen backbone with linear probing under a non-distilled setting. DiJEPA achieves 83.24% top-1 accuracy, matching 83.3% from DINOv3 under a comparable non-distilled protocol. The jointly learned FAE latents can also be used for image generation, achieving 2.74 gFID without CFG in 80 epochs of latent diffusion training. These results demonstrate that diffusion-based prediction in a learned latent space is a viable alternative to both projection-based regression and token-softmax prediction for JEPA-style representation learning.

2 Related Work

Self-Supervised Learning from Evolving Deep Representations. Early self-supervised methods such as SimCLR (Chen et al., 2020), MoCo (He et al., 2020), and wav2vec (Schneider et al., 2019) learn representations through contrastive objectives. A growing family of methods instead learns by predicting representations that evolve during training. BYOL (Grill et al., 2020) directly regresses continuous representations produced by a momentum teacher, while SimSiam (Chen & He, 2021) avoids the use of a momentum encoder by matching global representations across augmented views. Data2Vec (Baevski et al., 2022) predicts contextualized latent representations from masked inputs, while Data2Vec 2.0 (Baevski et al., 2023) improves representation quality and training efficiency. I-JEPA (Assran et al., 2023a) further improves representation quality by predicting spatial target blocks from a large visible context.

Instead of directly regressing continuous representations, SwAV (Caron et al., 2020) and DINO (Caron et al., 2021) learn by matching distributions over discrete prototypes across augmented views. iBOT (Zhou et al., 2021) introduces an online tokenizer jointly trained with evolving embedding. DINOv2 (Oquab et al., 2023) and DINOv3 (Sim'eoni et al., 2025) combine image-level and patch-level prediction objectives and further scale this recipe to large-scale training, obtaining strong performance on both semantic and pixel-level tasks.

Another line of work uses explicit distribution constraints to prevent collapse. SwAV (Caron et al., 2020) uses Sinkhorn–Knopp balancing for its cluster assignments, while VICReg (Bardes et al., 2021) and RDMReg (Kuang et al., 2026) use explicit distribution regularization, with LeJEPA (Balestriero & LeCun, 2025) introducing SIGReg as a Gaussian regularization objective to prevent collapse.

There are also several related variants with different target constructions. wav2vec 2.0 (Baevski et al., 2020) applies a contrastive objective over quantized latent targets, while BEiT (Bao et al., 2022) predicts discrete tokens produced by a pretrained tokenizer. HuBERT (Hsu et al., 2021) obtains discrete deep targets through offline clustering and refreshes them between training rounds. W2V-BERT (Chung et al., 2021) and NEPA (Xu et al., 2025) construct online targets from earlier convolutional or patch-embedding representations rather than deep teacher representations.

Despite these variations, existing methods for learning from evolving deep representations primarily rely on projection-based regression or matching distributions over discrete targets. DiJEPA instead formulates learning as diffusion-based prediction over evolving continuous deep representations.

Generative objectives for representation learning. A separate line of work studies diffusion or score-matching objectives for representation learning with targets that do not evolve together with the learned deep representation. DRL (Abstreiter et al., 2021) studies how manual control of the level of details encoded in the representation affects downstream classification task performance. Mukhopadhyay et al., 2023 shows diffusion model naturally encodes more information for image classification than GANs. DDAE (Xiang et al., 2023) verifies end-to-end diffusion pre-training can also learn strongly linear-separable features on small datasets including CIFAR10 and Tiny-ImageNet. DiffMAE (Wei et al., 2023) formulates masked autoencoding as a conditional diffusion process, reconstructing masked image regions through iterative denoising process and also achieves strong linear probing performance with diffusion-based reconstruction loss. l-DAE (Chen et al., 2024) first maps images into a low-dimensional latent space using either a separately trained patch-wise VAE or a fixed PCA projection, and then applies diffusion-based denoising in this non-evolving latent space for self-supervised representation learning. D-JEPA (Chen et al., 2025) combines a standard JEPA prediction loss for representation learning with a separate per-token diffusion loss over frozen MAR latents for autoregressive generation. In contrast, DiJEPA introduces a diffusion-based prediction objective over online FAE latents that evolve jointly with the underlying deep representations throughout training.

Representation Alignment for Generation Recent work has shown that aligning generative representations with pretrained visual features can lead to faster convergence and improved quality. REPA (Yu et al., 2024) aligns the noisy hidden states of diffusion transformers with clean embeddings from a pretrained ViT, leading to faster convergence and improved fidelity. Subsequent methods further explore this direction: REG (Wu et al., 2025) encourages semantic consistency between latent tokens and image features during training, while REPA-E (Leng et al., 2025) jointly optimizes the tokenizer and diffusion model with a related alignment objective. Different from direct alignment methods, VA-VAE (Yao et al., 2025) observes that high-dimensional latent spaces can improve

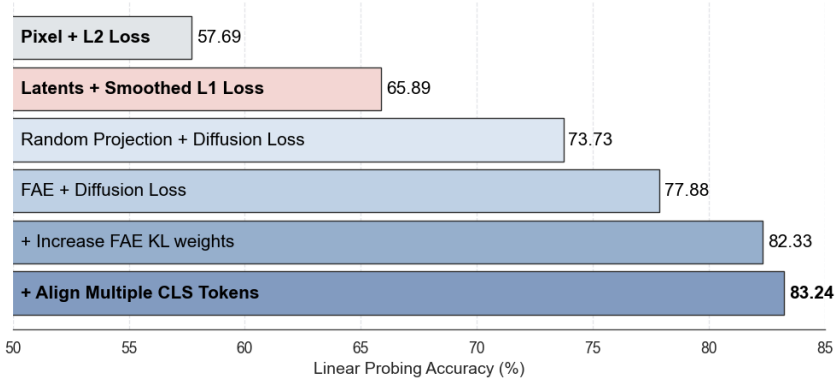

Figure 2: Experiments on ImageNet-1K. The first four rows are pilot experiments, while the remaining rows evaluate design choices in our final configuration. See Appendix B for settings.

reconstruction but make generative training more difficult, motivating latent spaces that are better aligned with pretrained visual representations. RAE (Zheng et al., 2025) directly uses pretrained encoder features as the diffusion latent space by introducing a wider DDT head to process high-dimensional representations. FAE (Gao et al., 2025) instead compresses pretrained visual features into a compact latent space with a lightweight feature auto-encoder, preserving their semantic structure while maintaining a generation-friendly latent dimensionality. Unlike these approaches, DiJEPA jointly learns the compact FAE latent space with the visual encoder during self-supervised pre-training, avoiding a separate stage for constructing or aligning a generation-oriented latent space.

To our knowledge, DiJEPA is among the first methods to apply diffusion-based prediction directly to deep online latent representations for self-supervised visual learning.

3 Method

3.1 Diffusion in an Evolving Latent Space

Before arriving at the final design, we conduct pilot experiments to test the viability of several candidate designs and compare them with simple alternatives. These experiments are performed under the same preliminary training setting, and the hyperparameters are not intended to represent optimal results. Rather, they are used to identify which design choices are important for constructing effective JEPA-style learning targets.

Previous work such as l-DAE (Chen et al., 2024) explored applying diffusion to VAE latents derived from pixels for self-supervised pre-training. However, applying diffusion to evolving representation latents is more challenging than applying it to fixed latent spaces. Existing methods often avoid this difficulty by carefully stabilizing the target distribution. For example, DINOv2 uses a carefully designed teacher softmax temperature schedule, while HuBERT (Hsu et al., 2021) trains the tokenizer offline and updates it only between training rounds. This issue is particularly important for diffusion-based prediction. In generative modeling (Ma et al., 2024; Rombach et al., 2021), diffusion objectives are known to be optimization-sensitive and often require smaller learning rates, less weight decay, and stable training targets. In BYOL-style representation learning, however, the target embeddings evolve rapidly during training, making diffusion-based prediction more prone to instability.

To evaluate the viability of diffusion-based latent prediction, we first apply the diffusion loss with a fixed-form projector that provides stable targets. We use random projection to reduce the target dimensionality. This design choice is motivated by prior SSL work showing that fixed-form projectors can perform comparably to optimized projectors in both image (Richemond et al., 2023) and audio (Chiu et al., 2022) settings. It is worth noting that the self-supervised setting differs from generation models such as RAE (Zheng et al., 2025). RAE uses a wider-channel DDT head, allowing the model to summarize high-rank global information through the iterative generation process. In contrast, in self-supervised learning, the diffusion objective cannot effectively take high-rank latents as targets without increasing the predictor dimension. This is supported by the latent-dimension ablation in DAE (Chen et al., 2024) and our ablation in Table 1c. Thus, a dimensionality-reducing projection is necessary, even if the projection is random.

Algorithm 1 DiJEPA Training Schema

```
# student_backbone, teacher_backbone are backbone ViTs
# fae_encoder, fae_decoder are part of FAE

def DiJEPA(view_x, view_y):
    embeddings_x = student_backbone(view_x) # [B, S, D]
    embeddings_y = teacher_backbone(view_y) # [B, S, D]

    latents_y = fae_encoder(embeddings_y) # [B, S, D*]
    reconstructed_y = fae_decoder(latents_y) # [B, S, D]
    loss_fae = l2_loss(embeddings_y - reconstructed_y)
    loss_fae += kl_weights * kl_divergence(latents_y, normal_distribution)

    t = uniform((B, S),)
    noised_latents = diffusion_scheduler(latents_y, t)
    predicted_latents = diffusion_decoder(embeddings_x, noised_latents)

    images_level_objective = diffusion_loss(
        latents_y[:, CLS] - predicted_latents[:, CLS]
    )
    patch_level_objective = diffusion_loss(
        latents_y[:, MASKED_PATCHES] - predicted_latents[:, MASKED_PATCHES]
    )
    loss_vit = images_level_objective + patch_level_objective
    return loss_fae + loss_vit
```

We compare this design against pixel-space reconstruction with an ℓ_2 loss and latent-space reconstruction with a smooth-ℓ_1 loss. After verifying that diffusion-based prediction is viable with a fixed random projection, we move to an online FAE target space. Fixed random projections provide stability and are meaningful at the beginning of training, but their embedding distribution may become suboptimal as representation learning progresses. In contrast, online FAE targets are less stable and require more careful design, but they evolve with the backbone, allowing the target space to adapt as the representation improves.

These pilot experiments validate the main design choices of DiJEPA. First, representation-space prediction outperforms pixel-space reconstruction, confirming that higher-level targets are more suitable for JEPA-style learning than raw-pixel reconstruction. Second, diffusion-based prediction successfully converges and outperforms smooth-ℓ_1 regression even with a fixed random projection, showing that the gain also comes from the diffusion learning objective rather than solely from the learned FAE target space. Consistent with this, DINO (Caron et al., 2021) reports that replacing cross-entropy over prototype assignments with MSE reduces ViT-S linear-probing accuracy from 76.1% to 62.4%. Third, using online FAE targets further improves over frozen random projections, indicating that the target distribution benefits from evolving together with the backbone rather than remaining fixed throughout training. We further analyze the characteristics of the learned target space in Appendix I.1. The FAE produces compact latents with nearly full effective dimensionality and substantially stronger Gaussian structure for CLS and individual patch tokens.

3.2 Learning Objective

In training, we follow the DINOv3 data processing and crop augmentation pipeline. We use two types of augmented views: different crops from the same image for the image-level objective, and masked image views for the patch-level objective.

Learning target. For both objectives, prediction is performed through the same diffusion mechanism in the FAE latent space. Given a teacher embedding h^* from the EMA network and the corresponding student embedding h, we first encode h^* with the FAE encoder to obtain a compact latent x^*. We then add Gaussian noise to obtain x_t. The diffusion predictor then uses the student embedding h together with the noised teacher latent x_t to predict the clean teacher latent x^*.

Patch-level objective. For patch-level learning, we randomly mask a subset of input patches for the student while keeping the teacher view unmasked. Here, h^* is the teacher patch embedding corresponding to a masked student patch, and h is the corresponding student patch embedding. All patch latents are noised and decoded independently.

Image-level objective. For image-level learning, h^* and h are the teacher and student CLS embeddings from different crops of the same image.

Diffusion loss. A key difference between our model and prior discrete-token prediction methods or contrastive learning methods is that we use a diffusion-style reconstruction objective in the FAE latent space. In the decoder, we predict the clean teacher FAE latent based on noised ground truth and student model output. We use cosine diffusion timestep schedule.

For each i-th FAE token x_i^* from EMA teacher, and a given timestep t, and student model output h_i, the L_2 loss is defined by

$$\mathcal{L}_{i,t,\epsilon} = \left\| x_i^* - \mathrm{Dec}\big(h_i,\ \sqrt{\overline{\alpha}_t}\, x_i^* + \sqrt{1-\overline{\alpha}_t}\, \epsilon\big) \right\|^2$$

where the notation $\overline{\alpha}_t$ follows the original DDPM (Ho et al., 2020) paper representing how much ground truth we keep at timestep t, ϵ represents a noise sampled from normal distribution which is independently sampled for each FAE token and timestep.

Because the scale of $\mathcal{L}_{i,t,\epsilon}$ from each sample are largely determined by the noise timestep, position of the masked unit, scale of raw input, in order to balance the sample weights and make the loss meaningful for learning representations, similar to E3TTS (Gao et al., 2023) and EDM2 (Karras et al., 2024), we predict the mean square error(or in some sense, variance) for each masked unit and use it to convert the loss into log scale:

$$\mathcal{L} = \sum_i \sum_{t\sim[0,1],\ \epsilon\sim\mathcal{N}(0,I)} \left(\frac{\mathcal{L}_{i,t,\epsilon}}{\sigma_{i,t}^2} + \log \sigma_{i,t}^2 \right)$$

To normalize the variance, similar to GiVit (Tschannen et al., 2024), we use softplus function to restrict the $\sigma_{i,t}^2 \geq \sigma_{min}^2$, and in experiment, we set $\sigma_{min}^2 = 1e-10$.

$$\log \sigma_{i,t}^2 = \mathrm{softplus}\big(\mathrm{logvar} - \log \sigma_{\min}^2\big) + \log \sigma_{\min}^2]$$

In this paper, to further improve stability when predicting a moving target, we disentangle the loss function into two part: KL term, L2 loss term, where $[\cdot]_{\mathrm{sg}}$ denotes the stop-gradient operator.

$$\mathcal{L}_{\mathrm{KL},\, i,t} = \sum_{\epsilon\sim\mathcal{N}(0,I)} \left(\frac{[\mathcal{L}_{i,t,\epsilon}]_{\mathrm{sg}}}{\sigma_{x_i,t}^2} + \log \sigma_{i,t}^2 \right)$$

$$\mathcal{L}_{\mathrm{L2},\, i,t} = \sum_{\epsilon\sim\mathcal{N}(0,I)} \frac{\mathcal{L}_{i,t,\epsilon}}{[\sigma_{i,t}^2]_{\mathrm{sg}}}$$

To balance contributions across different timesteps and improve training stability, we shrink the stop-gradient denominator by applying a power transform to the L2 term:

$$\mathcal{L}_{\mathrm{L2},\, i,t}^{\gamma} = \sum_{\epsilon\sim\mathcal{N}(0,I)} \frac{\mathcal{L}_{i,t,\epsilon}}{\big([\sigma_{i,t}^2]_{\mathrm{sg}}\big)^{\gamma}}, \quad 0 < \gamma < 1.$$

In our experiments, we set $\gamma = 0.95$ for CLS latents prediction and $\gamma = 0.25$ for masked patch latents prediction.

Diffusion Predictor After obtaining the FAE-encoded teacher latents, we feed them into an FFN-only diffusion predictor; its architecture is illustrated in Appendix D. The first FFN takes the noised latent as input and uses the timestep embedding as its conditioning signal. The second FFN takes the encoder output as input and uses the representation of the noised latent, produced by the first FFN, as its conditioning signal. This design keeps the gradient path from the loss to the encoder output short, while preserving a straightforward residual skip connection.

3.3 FAE

We train FAE jointly with the ViT backbone. The student FAE compresses teacher ViT outputs into low-dimensional latents and reconstructs the original outputs. Meanwhile, the student ViT is trained to predict the FAE latents produced from the corresponding teacher representations.

A key modification from the original FAE is the use of stronger KL regularization. As shown in Figure 2 and Ablation 1a, unlike the FAE used only for generation, the FAE used for representation pretraining benefits from substantially stronger KL regularization. In particular, stronger distribution

control encourages latents from different images to be more clearly separated in the target space during early training, providing a more cluster-like learning target initially, while gradually transitioning toward a more continuous representation space as training progresses.

We also find that stabilizing the latent ℓ_2 norm is important when using FAE latents as diffusion learning targets. As the backbone evolves during training, large fluctuations in the latent norm can destabilize the diffusion objective and lead to rank collapse. To address this, we decouple the KL divergence into two terms: a global term that controls only the overall ℓ_2 norm of the latent distribution, and a local term that retains the remaining KL regularization:

$$\mathcal{L}_{\mathrm{KL}}^{\mathrm{global}} = \frac{1}{2}\left(\mathbb{E}[\|z\|_2^2] - \log\big(\mathbb{E}[\|z\|_2^2]\big) - 1\right), \quad \text{where } \mathbb{E}[\|z\|_2^2] = \mathbb{E}[\mu^2 + \sigma^2]$$

$$\mathcal{L}_{\mathrm{KL}}^{\mathrm{local}} = \mathcal{L}_{\mathrm{KL}} - \mathcal{L}_{\mathrm{KL}}^{\mathrm{global}} = \frac{1}{2}\,\mathbb{E}\big[\log\big(\mathbb{E}[\|z\|_2^2]\big) - \log\sigma^2\big]$$

We reweight the two terms independently, assigning a larger weight to $\mathcal{L}_{\mathrm{KL}}^{\mathrm{global}}$ for tighter control of the latent norm:

$$\mathcal{L}_{\mathrm{KL}}^{\mathrm{reweighted}} = \lambda^{\mathrm{global}}\,\mathcal{L}_{\mathrm{KL}}^{\mathrm{global}} + \lambda^{\mathrm{local}}\,\mathcal{L}_{\mathrm{KL}}^{\mathrm{local}}, \quad \text{when } \lambda^{\mathrm{global}} = \lambda^{\mathrm{local}} = 1,\ \mathcal{L}_{\mathrm{KL}}^{\mathrm{reweighted}} = \mathcal{L}_{\mathrm{KL}}$$

3.4 Loss Design Exploration

By removing the limitation of quantization and leveraging a flexible learnable centering mechanism, DiJEPA opens up a broader design space for JEPA-style learning targets. Building on this flexibility, we explore several loss-design variants, including multiple CLS latents and an adaptive diffusion timestep schedule to balance their contributions. These explorations demonstrate that DiJEPA can support richer JEPA-style targets while maintaining stable training.

Multiple CLS latents in the loss. DINOv2 (Darcet et al., 2023) introduces multiple register tokens into the ViT backbone, but does not directly include them in the loss. In contrast, DiJEPA uses a more general design that maps multiple CLS tokens into separate FAE latents and includes them directly in the learning objective, with the goal of encouraging different latents to capture complementary aspects of the image. As shown in Figure 2 and Table 1, increasing the number of CLS latents provides gains similar to those observed with register tokens in DINOv2. Our kNN analysis further suggests that different CLS latents focus on different aspects of the representation.

Adaptive diffusion schedule. Different crop pairs from the same image may share different degrees of similarity across different CLS latents. Some crop pairs may align strongly on one latent, while other crop pairs may align more strongly on another. Forcing all latent pairs to align with the same diffusion time schedule is therefore suboptimal. To address this, we introduce adaptive diffusion timestep shift. The timestep shift are predicted from frozen CLS embedding pairs and used to adjust noise sampling in the diffusion process. Within each example, the shifts are regularized by a KL divergence term and by controlling the average α_t. Our ablation shows that adaptive diffusion timestep shift stabilizes multi-CLS training and improves performance.

4 Experiments

4.1 Main Experiments

We implement our method in the DINOv3 codebase (Sim'eoni et al., 2025) and follow its non-distilled ImageNet-1K training and data-processing setup. Specifically, we pre-train a ViT-L backbone for 125,000 training steps; detailed hyperparameters are provided in the appendix.

We evaluate the learned representations using two standard classification protocols: k-NN and linear probing. For linear probing, the pre-trained backbone is frozen throughout training, and only a linear classifier is trained for 12,500 steps with batch size 1024. During evaluation, images are resized to 256 pixels on the shorter side and center-cropped to 224×224. No data augmentation is applied, and we report top-1 accuracy on the ImageNet-1K validation set. For k-NN evaluation, we sweep over different numbers of nearest neighbors and perform distance-weighted voting.

As shown in Table 1, DiJEPA achieves 83.24% linear probing top-1 accuracy, competitive with 83.3% from DINOv3 under the comparable non-distilled setting. These results establish diffusion-

Method	Arch.	Data	kNN	Linear
Weakly supervised				
SimCLR v2	RN152 (2×)	INet-1k	–	79.1
CLIP	ViT-L/14	WIT-400M	79.8	84.3
SWAG	ViT-H/14	IG3.6B	82.6	85.7
OpenCLIP	ViT-G/14	LAION-2B	83.2	86.2
EVA-CLIP	ViT-g/14	–	**83.5**	86.4
PEcore	ViT-G/14	–	–	**89.3**
SigLIP 2	ViT-g/16	–	–	89.1
AIMv2	ViT-3B/14	–	–	87.9
Self-supervised				
MAE	ViT-L/16	INet-1k	–	76.0
	ViT-H/14	INet-1k	49.4	76.6
SwAV	ViT-S/16	INet-22k	66.3	73.5
BYOL	RN200w2	INet-22k	73.9	79.6
CAE	ViT-L/16	INet-1k	–	78.1
DINO	ViT-S/8	INet-1k	78.6	79.2
SEERv2	RG10B	IG2B	–	79.8
MSN	ViT-L/7	INet-1k	79.2	80.7
EsViT	Swin-B/W=14	INet-1k	79.4	81.3
Mugs	ViT-L/16	INet-1k	80.2	82.1
data2vec	ViT-L/16	INet-1k	–	77.3
I-JEPA	ViT-L/16	INet-1k	–	77.5
LeJEPA	ViT-L/16	INet-1k	–	77.1†
iBOT	ViT-L/16	INet-22k	72.9	82.3
DINOv3	ViT-L/16	INet-1k	**82.2**	**83.3**
DiJEPA	ViT-L/16	INet-1k	–	82.4
DiJEPA(4 CLS tokens)	ViT-L/16	INet-1k	80.3	83.2

KL Weight	25K	50K	75K	100K	125K
0.002	67.8	70.8	73.5	76.7	79.0
0.05	73.4	73.9	76.7	80.1	80.7
0.075	71.4	75.2	77.6	77.6	81.7
0.1	69.6	74.8	78.1	81.2	**82.4**
0.2	62.6	74.5	78.5	81.0	81.8
0.1–0.01	64.5	75.5	78.7	81.7	82.3

(a) Different KL weights.

Model	# CLS	Adaptive Schedule	Probing
DINOv3	1		83.03 %
DINOv3	4		83.14 %
DiJEPA	1	No	82.37 %
DiJEPA	4	No	82.56 %
DiJEPA	4	Yes	83.24 %

(b) Multiple CLS and adaptive schedule.

FAE Dim	Layers	SwiGLU	Probing
64	4	–	82.37%
128	4	–	80.16%
64	6	–	81.91%
64	4	Yes	Failed

(c) Diffusion Predictor Structure.

Table 1: **Main results and ablation studies. Left: Main ImageNet-1K representation results,** with baseline numbers collected from iBOT (Zhou et al., 2021) and DINOv2 (Oquab et al., 2023); the DINOv3 result is taken from the official DINOv3 repository, while the LeJEPA result is taken from the LeJEPA paper (Balestriero & LeCun, 2025) and is reported using online linear probing. **Right: Ablation study results.** The DINOv3 baselines used in ablation and analysis are from our own reproductions under the same ImageNet-1K training setup as DiJEPA.

based prediction in the FAE latent space as a strong alternative to projection-based regression and token-softmax prediction for JEPA-style representation learning.

4.2 Generation Experiments Result

For all models evaluated in our implementation, representations are extracted directly from 256×256 images, corresponding to a 16×16 latent grid. To isolate the effect of jointly learning the representation and FAE latent space, we additionally train the same FAE on a frozen DINOv3 ViT-L backbone with four storage tokens, which achieves 83.14% ImageNet-1K linear-probing accuracy. DINOv3 + FAE and DiJEPA use the same generation setup. For VA-VAE, we report the 64-epoch result from the original paper and the 80-epoch reproduction from RAE.

As shown in Table 2, DiJEPA achieves a gFID of 2.74 after 80 epochs of latent diffusion training, compared with 3.91 for DINOv3 ViT-L + FAE, while improving the Inception Score from 159.92 to 216.61. These results suggest that jointly optimizing the visual representation and compact FAE target space produces latents better suited for downstream generative modeling than fitting an FAE post hoc to a pretrained representation. Although the final gFID remains slightly behind the original FAE and RAE results, those methods use larger ViT-G backbones pretrained on substantially larger datasets, or models distilled from such backbones.

4.3 Dense Prediction

We further evaluate DiJEPA on dense prediction tasks, including semantic segmentation on ADE20K and monocular depth estimation on NYUv2, following the DINOv3 experimental settings for both

Base Model	Alignment Method	Epochs	IS ↑	gFID ↓	sFID ↓	Precision ↑	Recall ↑
DINOv2 ViT-B Distilled	REPA	800	–	5.90	–	–	–
DINOv2 ViT-L Distilled	VA-VAE	64	130.2	5.14	4.22	0.76	**0.62**
DINOv2 ViT-L Distilled	VA-VAE	80	–	4.29	–	–	–
DINOv2 ViT-B Distilled	RAE	80	214.80	2.16	–	0.82	0.59
DINOv2 ViT-G	FAE	80	207.60	**2.08**	**4.20**	0.82	0.59
DINOv3 ViT-L	FAE	80	159.92	3.91	4.47	**0.84**	0.54
DiJEPA ViT-L	–	80	**216.61**	2.74	5.31	0.83	0.54

Table 2: ImageNet-1K 256×256 generation results. Epochs denote latent diffusion training epochs.

tasks. We use our reproduced DINOv3 ViT-L checkpoint with four storage tokens as the baseline. As shown in Table 5, DiJEPA achieves 43.54 mIoU on ADE20K, compared with 44.39 for DINOv3. On NYUv2, DiJEPA obtains an RMSE of 0.474, compared with 0.444 for DINOv3 (Table 6). The relatively small gap on these dense prediction benchmarks is consistent with the ImageNet representation results, indicating that DiJEPA preserves transferable spatial information in addition to image-level semantic features.

4.4 Ablation Studies

We conduct ablation studies to analyze the key design choices in DiJEPA, including the strength of KL regularization discussed in Section 3.3, the use of multiple CLS latents with adaptive timestep shift discussed in Section 3.4, and the diffusion predictor architecture discussed in Section 3.2. The results are summarized in Table 1.

KL regularization strength. Unlike the original FAE setting for generation, representation learning benefits from stronger KL regularization. We evaluate six KL-weight settings from 2×10^{-3} to 0.2; Table 1a reports linear-probing accuracy throughout training. Smaller KL weights converge faster early but yield worse final accuracy, with 0.1 performing best among fixed weights. We stop at 0.2, as stronger regularization increases sensitivity to backbone fluctuations and can cause FAE latents to collapse toward random representations. We additionally evaluate a linearly decayed schedule from 0.1 to 0.01, which achieves representation quality comparable to the fixed 0.1 setting while producing smaller final latent variance, making the resulting latents more suitable for generation.

Diffusion Predictor structure. Table 1c shows that 64-dimensional latents with 4 decoder layers give the best probing accuracy. Increasing the latent dimension to 128 substantially hurts performance, suggesting that larger latents are harder to diffuse. Increasing decoder depth from 4 to 6 layers also slightly hurts performance. Replacing the standard MLP with SwiGLU fails to converge. We hypothesize that SwiGLU destabilizes gradient dynamics under backbone fluctuations, causing gradient-norm growth. Overall, a lightweight decoder with compact latents performs best.

Multiple CLS latents and adaptive timestep shift. Table 1b shows that increasing DINOv3 from 1 to 4 CLS tokens gives only a small gain. This gain is smaller than in prior register-token experiments Darcet et al. (2023), possibly due to their larger-scale ImageNet-21K training. For DiJEPA, 4 CLS latents give a similar gain without adaptive timestep shift, but a substantially larger gain when combined with it. This suggests that adaptive timestep assignment helps different CLS latents specialize and better exploit the added capacity.

Overall, these ablations support our design choices in DiJEPA: stronger KL regularization, a compact diffusion predictor, and adaptive timestep shift for multiple CLS latents.

5 Conclusion

We presented DiJEPA, a JEPA-style framework that formulates self-supervised learning as diffusion-based prediction over jointly learned FAE latents. On ImageNet-1K, DiJEPA achieves competitive non-distilled ViT-L representation quality, while the learned FAE latents can be used for image generation, achieving 2.74 gFID without CFG. These results establish diffusion-based prediction over continuous representation latents as a promising objective for both understanding and generation.

AI USE STATEMENT

In this work, we used generative AI tools for polishing writing. We have not used generative AI tools for other tasks with required disclosure, and the rest of the required disclosure tasks are not applicable to this work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the model architecture, training objectives, and experimental setup in the main text and appendix. The appendix further provides training hyperparameters, implementation details, and evaluation protocols to support reproducibility. We are also seeking approval to release the source code in the future.

REFERENCES

Korbinian Abstreiter, Sarthak Mittal, Stefan Bauer, Bernhard Schölkopf, and Arash Mehrjou. Diffusion-based representation learning. *arXiv preprint arXiv:2105.14257*, 2021.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15619–15629, 2023a.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael G. Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629, 2023b. URL https://api.semanticscholar.org/CorpusID:255999752.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: a framework for self-supervised learning of speech representations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. data2vec: A general framework for self-supervised learning in speech, vision and language, 2022.

Alexei Baevski, Arun Babu, Wei-Ning Hsu, and Michael Auli. Efficient self-supervised learning with contextualized target representations for vision, speech and language, 2023.

Randall Balestriero and Yann LeCun. Lejepa: Provable and scalable self-supervised learning without the heuristics. *ArXiv*, abs/2511.08544, 2025. URL https://api.semanticscholar.org/CorpusID:282922448.

Randall Balestriero, Mark Ibrahim, Vlad Sobal, Ari Morcos, Shashank Shekhar, Tom Goldstein, Florian Bordes, Adrien Bardes, Gregoire Mialon, Yuandong Tian, et al. A cookbook of self-supervised learning. *arXiv preprint arXiv:2304.12210*, 2023.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers, 2022.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *ArXiv*, abs/2105.04906, 2021. URL https://api.semanticscholar.org/CorpusID:234357520.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *ArXiv*, abs/2006.09882, 2020. URL https://api.semanticscholar.org/CorpusID:219721240.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9630–9640, 2021.

Dengsheng Chen, Jie Hu, Xiaoming Wei, and Enhua Wu. Denoising with a joint-embedding predictive architecture, 2025. URL https://arxiv.org/abs/2410.03755.

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

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15750–15758, 2021.

Xinlei Chen, Zhuang Liu, Saining Xie, and Kaiming He. Deconstructing denoising diffusion models for self-supervised learning, 2024.

Chung-Cheng Chiu, James Qin, Yu Zhang, Jiahui Yu, and Yonghui Wu. Self-supervised learning with random-projection quantizer for speech recognition, 2022.

Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. w2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 244–250, 2021. doi: 10.1109/ASRU51503.2021.9688253.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. *ArXiv*, abs/2309.16588, 2023. URL https://api.semanticscholar.org/CorpusID:263134283.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

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, 2021. URL https://arxiv.org/abs/2010.11929.

Yuan Gao, Nobuyuki Morioka, Yu Zhang, and Nanxin Chen. E3 tts: Easy end-to-end diffusion-based text to speech, 2023.

Yuan Gao, Chen Chen, Tianrong Chen, and Jiatao Gu. One layer is enough: Adapting pretrained visual encoders for image generation. *ArXiv*, abs/2512.07829, 2025. URL https://api.semanticscholar.org/CorpusID:283694483.

Quentin Garrido, Randall Balestriero, Laurent Najman, and Yann Lecun. Rankme: Assessing the downstream performance of pretrained self-supervised representations by their rank, 2023. URL https://arxiv.org/abs/2210.02885.

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

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 29:3451–3460, oct 2021. ISSN 2329-9290. doi: 10.1109/TASLP.2021.3122291. URL https://doi.org/10.1109/TASLP.2021.3122291.

Tero Karras, Miika Aittala, Jaakko Lehtinen, Janne Hellsten, Timo Aila, and Samuli Laine. Analyzing and improving the training dynamics of diffusion models, 2024.

Yilun Kuang, Yash Dagade, Tim G. J. Rudner, Randall Balestriero, and Yann LeCun. Rectified lpjepa: Joint-embedding predictive architectures with sparse and maximum-entropy representations. *ArXiv*, abs/2602.01456, 2026. URL https://api.semanticscholar.org/CorpusID:285269150.

Yann LeCun. A path towards autonomous machine intelligence, June 2022. URL https://openreview.net/pdf?id=BZ5a1r-kVsf. Version 0.9.2.

Xingjian Leng, Jaskirat Singh, Yunzhong Hou, Zhenchang Xing, Saining Xie, and Liang Zheng. Repa-e: Unlocking vae for end-to-end tuning of latent diffusion transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18262–18272, 2025.

Xianhang Li, Chen Huang, Chun-Liang Li, Eran Malach, Joshua M. Susskind, Vimal Thilak, and Etai Littwin. Rethinking JEPA: Compute-efficient video self-supervised learning with frozen teachers. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=3cB9243E9i.

Nanye Ma, Mark Goldstein, Michael S. Albergo, Nicholas Matthew Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision*, 2024. URL https://api.semanticscholar.org/CorpusID:267027717.

Soumik Mukhopadhyay, Matthew Gwilliam, Vatsal Agarwal, Namitha Padmanabhan, Archana Swaminathan, Srinidhi Hegde, Tianyi Zhou, and Abhinav Shrivastava. Diffusion models beat gans on image classification, 2023.

Lorenzo Mur-Labadia, Matthew Muckley, Amir Bar, Mido Assran, Koustuv Sinha, Mike Rabbat, Yann LeCun, Nicolas Ballas, and Adrien Bardes. V-jepa 2.1: Unlocking dense features in video self-supervised learning. *arXiv preprint arXiv:2603.14482*, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russ Howes, Po-Yao (Bernie) Huang, Shang-Wen Li, Ishan Misra, Michael G. Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *ArXiv*, abs/2304.07193, 2023. URL https://api.semanticscholar.org/CorpusID:258170077.

Pierre Harvey Richemond, Allison Tam, Yunhao Tang, Florian Strub, Bilal Piot, and Felix Hill. The edge of orthogonality: A simple view of what makes BYOL tick. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 29063–29081. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/richemond23a.html.

Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10674–10685, 2021. URL https://api.semanticscholar.org/CorpusID:245335280.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. pp. 3465–3469, 09 2019. doi: 10.21437/Interspeech.2019-1873.

Oriane Sim'eoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michael Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Herv'e J'egou, Patrick Labatut, and Piotr Bojanowski. Dinov3. 2025. URL https://api.semanticscholar.org/CorpusID:280649654.

Michael Tschannen, Cian Eastwood, and Fabian Mentzer. Givt: Generative infinite-vocabulary transformers, 2024.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 6309–6318, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Chen Wei, Karttikeya Mangalam, Po-Yao (Bernie) Huang, Yanghao Li, Haoqi Fan, Hu Xu, Huiyu Wang, Cihang Xie, Alan Loddon Yuille, and Christoph Feichtenhofer. Diffusion models as masked autoencoders. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16238–16248, 2023. URL https://api.semanticscholar.org/CorpusID:257985028.

Ge Wu, Shen Zhang, Ruijing Shi, Shanghua Gao, Zhenyuan Chen, Lei Wang, Zhaowei Chen, Hongcheng Gao, Yao Tang, Jian Yang, Mingqiang Cheng, and Xiang Li. Representation entanglement for generation:training diffusion transformers is much easier than you think. *ArXiv*, abs/2507.01467, 2025. URL https://api.semanticscholar.org/CorpusID:280091678.

Weilai Xiang, Hongyu Yang, Di Huang, and Yunhong Wang. Denoising diffusion autoencoders are unified self-supervised learners, 2023.

Sihan Xu, Ziqiao Ma, Wenhao Chai, Xuweiyi Chen, Weiyang Jin, Joyce Chai, Saining Xie, and Stella X. Yu. Next-embedding prediction makes strong vision learners, 2025. URL https://arxiv.org/abs/2512.16922.

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models, 2025. URL https://arxiv.org/abs/2501.01423.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. *arXiv preprint arXiv:2410.06940*, 2024.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. *arXiv preprint arXiv:2510.11690*, 2025.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Loddon Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *ArXiv*, abs/2111.07832, 2021. URL https://api.semanticscholar.org/CorpusID:244117494.

A LIMITATIONS

Our experiments focus on ImageNet-1K with a ViT-L backbone under a non-distilled self-supervised training setting. This provides a controlled setting for comparing DiJEPA with strong JEPA-style baselines, but it does not fully establish how the method scales to larger datasets, larger backbones, or longer pre-training schedules. Extending the evaluation to broader datasets and architectures remains important future work.

DiJEPA introduces additional trainable components beyond standard direct regression or token-softmax objectives, including the FAE projection and diffusion predictor. These components provide a flexible continuous latent-space objective, but also increase computational cost and implementation complexity. Our current hyperparameters are tuned for ImageNet-1K with a ViT-L backbone, and may require adaptation when scaling to different datasets, architectures, or training budgets. Compared with standard JEPA-style objectives, DiJEPA introduces an additional trainable component FAE to construct the latent prediction space. Although this module is lightweight, it still adds some training cost and introduces extra hyperparemeters.

Finally, our generation experiments are intended to test whether the learned latents retain the generation-friendly properties of FAE, rather than to provide a fully optimized comparison with generation-specialized systems. The current setup uses a smaller backbone than the original FAE and RAE settings, and the shorter latent sequence length may further limit generation performance. As a result, the generation results should be interpreted as evidence of compatibility between representation learning and generation, rather than as a definitive benchmark of generative performance.

B PILOT EXPERIMENT SETTINGS

The pilot experiments in Figure 2 follow the DINOv3 training and data-processing setup unless otherwise specified. For improved training stability, we reduce the weight decay to 0.04 in these experiments. The pixel-reconstruction baseline is trained for 250K steps without the image-level objective, whereas all other settings are trained for 125K steps with the image-level objective enabled.

We retain the DINO-style masking strategy throughout these experiments: masking is applied to only 50% of images, and the masking ratio for each masked image is sampled uniformly from 10% to 50%. This masking configuration is inherited from DINOv3 and is not specifically optimized for masked-prediction objectives. Consequently, the pixel-reconstruction and latent-reconstruction pilot baselines may be suboptimal under this setting. For reference, MAE (He et al., 2021) shows that both masking ratio and training duration can substantially affect masked-prediction performance.

C FAE Encoder Structure

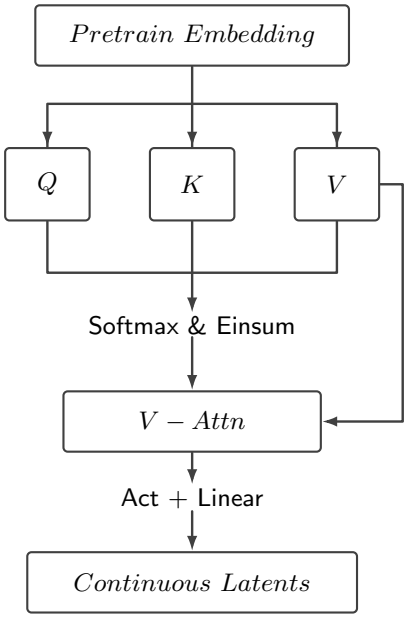

Figure 3: FAE Encoder with better performance when initialization

D Diffusion Predictor Architecture

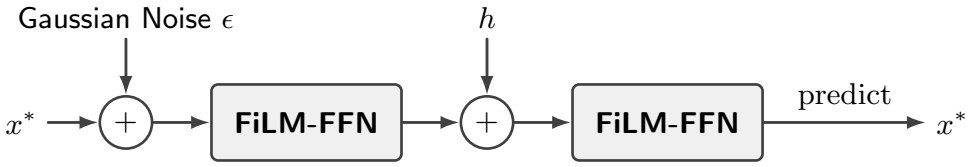

Figure 4: Architecture of the diffusion predictor. The feed-forward network incorporates conditional embeddings through FiLM layers.

E HYPERPARAMETERS

Training	
Backbone	ViT-L/16
Training Steps	125000
Total batch size	2048
Optimizer	AdamW
Precision	bf16
GPU	B200
Crops	2 large + 8 small
Global crop scale	$[0.32, 1.0]$
Local crop scale	$[0.05, 0.55]$
Masking prob.	0.5
Masking ratio	$[0.1, 0.5]$

Diffusion	
Schedule	cosine
Timestep shift	0.125
δ	0.01
Head layers	4
Norm layer	LayerNorm

Backbone	
Architecture	ViT-L
Patch size	16
Drop path	0.3
LayerScale	10^{-5}
FFN layer	4
MLP ratio	4.0
Storage tokens	4
Position embedding	RoPE

FAE	
FAE dimension	64
Depth	6
MLP ratio	16
Number of heads	64
Position embedding	RoPE

Optimization and Schedules	
LR schedule	10^{-3}, cosine
FAE LR schedule	2×10^{-4}, cosine
Layer-wise decay	0.9
AdamW β_1, β_2	0.9, 0.999
Gradient clip	3.0
LR multipliers	patch emb. 0.2 FAE Enc. 0.1 Diff. Pred. 0.25
Weight decay	backbone $0.1\rightarrow0.2\rightarrow0.4$ FAE enc. $0.0\rightarrow0.1\rightarrow0.15$
EMA decay	0.008
KL weight	$0.1\rightarrow0.01$
Adaptive weight	$0.25\rightarrow0.1$
Adaptive decay	112500 steps

Table 3: Hyperparameters used for ImageNet-1K pre-training.

F TRAINING COST AND MODEL SIZE

Compared with training the DINOv3 backbone alone, DiJEPA jointly trains the ViT backbone and FAE, together with the diffusion predictor, resulting in approximately 30% higher per-step time for ViT-L and 37% for ViT-G. This joint training simultaneously learns the visual representation and the compact FAE latent space, avoiding the need for a separate post-hoc FAE training stage when the representation is subsequently used for generative modeling. In our generation experiments, for example, fitting the same FAE to a pretrained DINOv3 ViT-L requires a separate 400K-step training stage. The additional per-step cost of DiJEPA therefore reflects the joint optimization of both representation learning and latent-space construction within a single pre-training process.

Configuration	Backbone	Head	FAE	Step Time (s)
DINOv3 ViT-L (bs=64)	303.15M	$23.60\text{M}\times2$	–	0.28494
DINOv3 ViT-L FAE (bs=256)	303.15M	–	76.71M	0.2124
DiJEPA ViT-L (bs=64)	303.15M	$93.61\text{M}\times2$	76.71M	0.36908
DINOv3 ViT-G (bs=24)	1,134.65M	$58.64\text{M}\times2$	–	0.46206
DiJEPA ViT-G (bs=24)	1,134.65M	$209.67\text{M}\times2$	228.36M	0.63444

Table 4: **Training cost and model size.** Parameter breakdown and per-step training time of DINOv3 and DiJEPA.

G ADE20K Experiment Results

Method	mIoU ↑	Acc ↑	aAcc ↑	Dice ↑	F-score ↑	Precision ↑	Recall ↑
DINOv3 ViT-L	44.39	55.83	79.98	58.57	58.57	64.65	55.83
DiJEPA ViT-L	43.54	54.70	79.53	57.51	57.51	64.31	54.70

Table 5: **ADE20K semantic segmentation results.** We follow the DINOv3 experimental setting. The DINOv3 baseline is our reproduced ViT-L checkpoint with four storage tokens.

H NYU Depth Experiment Results

Method	A1 ↑	Abs Rel ↓	RMSE ↓
DINOv3 ViT-L	0.84695	0.13542	0.44409
DiJEPA ViT-L	0.82075	0.14571	0.47395

Table 6: **NYUv2 monocular depth estimation results.** We follow the DINOv3 experimental setting. The DINOv3 baseline is our reproduced ViT-L checkpoint with four storage tokens.

I Embedding Analysis

We first study the Gaussianity and effective dimensionality of the backbone representations and FAE latents using LeJEPA-style SIGReg analysis and RankMe. We then follow the embedding analysis in DINOv3 to provide complementary visualizations of the learned representations.

I.1 Gaussianity and Effective Dimensionality

We further analyze the distributional properties of the learned representations using RankMe (Garrido et al., 2023), SIGReg, and the sliced ECF discrepancy. For CLS, Patch Mean, FAE CLS, and FAE Patch Mean, the mini-batch SIGReg statistic follows LeJEPA and is computed with a global batch size of 1,024. For Patch and FAE Patch, it is computed over the 196 tokens within each image and then averaged across images. All metrics are evaluated over 1,250 ImageNet-1K training batches.

The sliced ECF discrepancy is the sliced Epps–Pulley statistic normalized by the sample count, providing a sample-count-normalized measure of deviation from a standard Gaussian distribution. Lower values therefore indicate stronger Gaussianity.

As shown in Table 7, the 64-dimensional FAE CLS and patch latents achieve RankMe values of 62.538 and 63.797, respectively, indicating nearly full effective dimensionality. They also exhibit substantially lower SIGReg and sliced ECF discrepancies than the corresponding backbone features. The effect is strongest for FAE CLS and individual patch latents, while pooled patch means remain less Gaussian.

Metric	Model	CLS	Patch Mean	Patch	FAE CLS	FAE Patch Mean	FAE Patch
RankMe ↑	DINOv3	749.567	751.106	895.574	–	–	–
RankMe ↑	DiJEPA	593.473	630.841	758.410	62.538	55.721	63.797
Mini-batch SIGReg ↓	DINOv3	376.082	210.861	38.655	–	–	–
Mini-batch SIGReg ↓	DiJEPA	269.979	508.907	72.234	44.028	383.555	2.842
Sliced ECF discrepancy ↓	DINOv3	0.36692	0.20553	–	–	–	–
Sliced ECF discrepancy ↓	DiJEPA	0.26285	0.49668	–	0.04202	0.37452	–

Table 7: **Gaussianity and effective dimensionality of backbone representations and FAE latents.** RankMe measures effective dimensionality, while SIGReg and sliced ECF discrepancy measure deviation from a standard Gaussian distribution.

I.2 Dense Sparse Matching

We conduct the Dense Sparse Matching analysis both on ViT embeddings and on FAE.

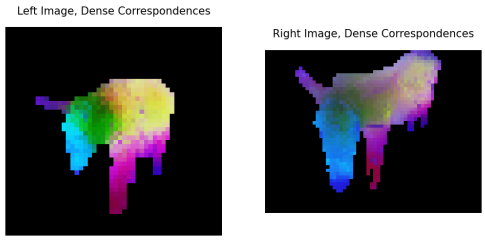

Figure 5: Dense Correspondences on ViT embeddings.



Figure 6: Sparse Correspondences on ViT embeddings.

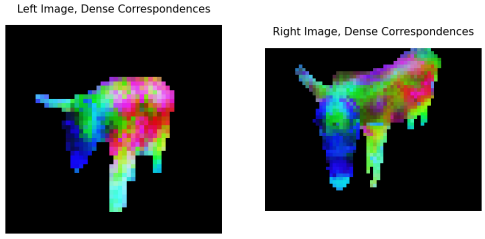

Figure 7: Dense Correspondences on FAE latents.



Figure 8: Sparse Correspondences on FAE latents.

Dense-Sparse Matching experiments indicate that FAE provides a stronger centering effect while largely preserving the original similarity structure. This centering behavior consequently influences the final DiJEPA embeddings.

I.3 FOREGROUND SEGMENTATION

We conduct the Foreground Segmentation analysis both on ViT embeddings.



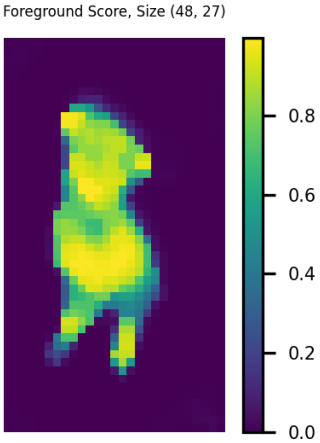

Figure 9: Foreground Score of DINOv3's dog.

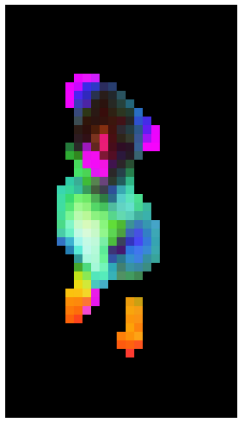

Figure 10: PCA of DINOv3's dog.



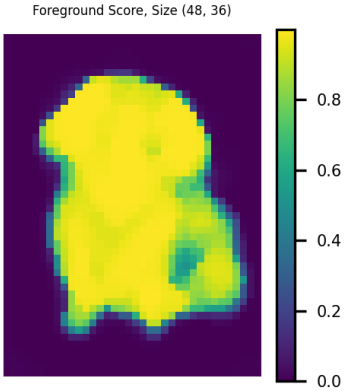

Figure 11: Foreground Score of our dog.

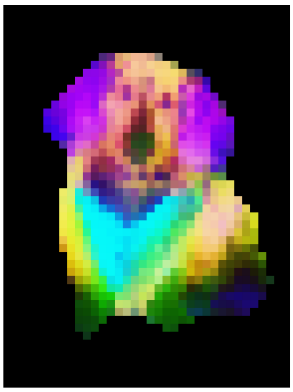

Figure 12: PCA of Our dog.

The Foreground Score experiments indicate that DiJEPA embeddings encode foreground-aware information that can be effectively transferred to downstream tasks.