DSeq-JEPA: Discriminative Sequential Joint-Embedding Predictive Architecture

Xiangteng He[*1,2], Shunsuke Sakai[*3], Shivam Chandhok[1,2], Sara Beery[4], Kun Yuan[5,6], Nicolas Padoy[5,6], Tatsuhito Hasegawa[3], and Leonid Sigal[1,2,7]

[1]University of British Columbia, Canada, [2]Vector Institute for AI, Canada, [3]University of Fukui, Japan [4]Massachusetts Institute of Technology, USA, [5]University of Strasbourg, France, [6]IHU Strasbourg, France, [7]Canada CIFAR AI Chair, Canada

Project Page: `https://dseqjepa-project.com`

Abstract. Recent advances in self-supervised visual representation learning have demonstrated the effectiveness of predictive latent-space objectives for learning transferable features. In particular, Image-based Joint-Embedding Predictive Architecture (I-JEPA) learns representations by predicting latent embeddings of masked target regions from visible context. However, it predicts target regions in parallel and all at once, lacking ability to order predictions meaningfully. Inspired by human visual perception, which attends selectively and progressively from primary to secondary cues, we propose DSeq-JEPA, a Discriminative Sequential Joint-Embedding Predictive Architecture that bridges latent predictive and autoregressive self-supervised learning. Specifically, DSeq-JEPA integrates a discriminatively ordered sequential process with JEPA-style learning objective. This is achieved by (i) identifying primary discriminative regions using an attention-derived saliency map that serves as a proxy for visual importance, and (ii) predicting subsequent regions in discriminative order, inducing a curriculum-like semantic progression from primary to secondary cues in pre-training. Extensive experiments across tasks – image classification (ImageNet), fine-grained visual categorization (iNaturalist21, CUB, Stanford Cars), detection/segmentation (MS-COCO, ADE20K), and low-level reasoning (CLEVR) – show that DSeq-JEPA consistently learns more discriminative and generalizable representations compared to I-JEPA variants.

Keywords: Discriminative sequential prediction · Joint-embedding predictive architecture · Self-supervised learning

1 Introduction

Self-supervised learning (SSL) has become a powerful paradigm for learning robust and transferable visual representations from unlabeled data. Early advances in contrastive learning [1–3] and masked image modeling [4–6] showed

* Equal contribution.

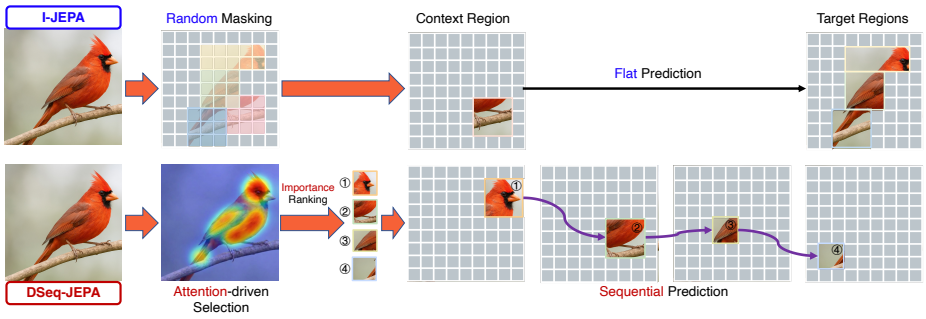

Fig. 1: Overview of the key differences between I-JEPA and DSeq-JEPA. I-JEPA predicts embeddings of randomly sampled target regions in a parallel and independent manner. In contrast, DSeq-JEPA ranks regions by attention and predicts region embeddings sequentially from high- to low-attention regions. Patches shown are only for illustration; the model predicts region embeddings rather than raw pixels.

that representations learned without manual annotation can rival supervised pretraining on many downstream tasks. More recently, Joint-Embedding Predictive Architectures (JEPAs) [7] have shifted the objective from contrasting instances or pixel reconstruction to latent-space prediction, offering a scalable and information-preserving formulation. Among these, I-JEPA [8] has emerged as a compelling framework for learning high-level visual representations through latent prediction of masked target regions. Despite its effectiveness, I-JEPA treats target regions largely uniformly and independently during pre-training.

In practice, visual information is rarely uniform: some regions carry primary semantic cues (*e.g.*, object-defining parts), while others provide secondary or contextual evidence. Moreover, visual perception is often *selective* and *sequential*, attending first to salient cues and then refining interpretation using additional context. This motivates a key research question:

> *Can predictive self-supervised learning benefit from an explicit notion of* ***where*** *and* ***in what order*** *to make predictions?*

In this work, we answer this question affirmatively and propose Discriminative Sequential Joint-Embedding Predictive Architecture (DSeq-JEPA), a new self-supervised pretraining framework that introduces a discriminative sequential inductive bias that is (loosely) motivated by human perception. This paradigm bridges two previously distinct lines of self-supervised learning: *predictive modeling* (JEPA-style) and *auto-regressive next region prediction* (iGPT-like [9]). Our key idea is to impose structure on target predictions by prioritizing semantically informative regions and predicting them in a progressive order. As illustrated in Figure 1, DSeq-JEPA first derives an attention-based saliency map to estimate visual importance and identify primary discriminative regions (*where to attend*), instead of I-JEPA's uniform sampling. It then performs sequential latent predictions over these regions from the most discriminative to the least (*in what*

order), rather than I-JEPA's flat, independent predictions. Ultimately, DSeq-JEPA yields a curriculum-like semantic progression from primary to secondary cue during pre-training.

The proposed design is motivated by two complementary observations. First, *where* the model predicts matters: focusing on discriminative regions can concentrate learning on informative content instead of treating all targets equally. Second, *in what order* the model predicts matters: sequential prediction encourages progressive representation building, where earlier high-value cues support subsequent predictions. Importantly, these components are synergistic: discriminative region selection provides an informative ordering prior, and sequential latent prediction converts this prior into a structured self-supervised learning signal. From this perspective, DSeq-JEPA can be viewed as breaking the permutation symmetry over target regions implicit in independent region prediction and replacing it with a semantically meaningful prediction order.

We validate the effectiveness of DSeq-JEPA through extensive experiments across diverse downstream tasks and evaluation settings. Compared with I-JEPA variants, DSeq-JEPA consistently improves transfer performance on image classification (ImageNet), fine-grained visual categorization (iNaturalist21, CUB-200-2011, Stanford Cars), dense prediction tasks including detection and segmentation (MS-COCO, ADE20K), and low-level reasoning benchmarks (CLEVR). Beyond quantitative gains, qualitative analyses show that DSeq-JEPA attends to and predicts more discriminative regions, while ablations demonstrate that both discriminative region selection and sequential latent prediction are necessary and that their combination yields the strongest improvements.

Contributions: Our contributions can be summarized as follows:

- We introduce **DSeq-JEPA**, a JEPA-based self-supervised pre-training framework that augments latent predictive learning with **discriminative region prioritization** and **sequential next-region embedding prediction**.
- We propose an attention-derived saliency-based mechanism to identify informative regions and construct a discriminative **prediction order**, enabling a curriculum-like progression from primary to secondary cues.
- Through extensive experiments and ablations, we show that DSeq-JEPA consistently improves over strong baselines across classification, fine-grained recognition, dense prediction, and low-level reasoning benchmarks.

By explicitly modeling both **where** and **in what order** to predict, DSeq-JEPA advances joint-embedding predictive learning toward a more selective and progressive form of self-supervised visual pre-training.

2 Related Work

Self-supervised learning (SSL) has become a central paradigm for learning transferable visual representations from unlabeled data. Existing SSL methods can be broadly grouped into three families:

Joint-Embedding Architectures are designed to bring the embeddings closer together for compatible inputs, and push the embeddings further apart for incompatible ones. This includes: (i) contrastive learning approaches, *e.g.*, SimCLR [1], MoCo [2, 3, 10], and SwAV [11], which form the foundation of this paradigm, optimizing instance-level discrimination via large batches or memory banks; and (ii) non-contrastive approaches, *e.g.*, BYOL [12], SimSiam [13], which further show that strong representations can emerge without explicit negatives via asymmetric prediction and momentum-based designs that avoid collapse.

Generative Architectures aim to reconstruct a target signal $\mathbf{y}$ (*e.g.*, pixels, or patches) from a related input signal $\mathbf{x}$ by employing a decoder network. Masked image modeling serves as a prominent technique in this category, including MAE [4], BEiT [5], SimMIM [6], and iGPT [9], where models learn to fill in masked regions of an image. Although highly effective for representation learning, their objectives are driven by content reconstruction, which can allocate substantial modeling capacity to low-level appearance recovery rather than directly encouraging high-level semantic abstraction. Notably, iGPT [9] adopts an autoregressive formulation that predicts image tokens sequentially. Relatedly, RandSAC [14] explores self-supervision through random segments with autoregressive coding, suggesting that sequential prediction can serve as a useful inductive bias for visual representation learning. However, these approaches do not explicitly model which regions are more informative and reconstruct pixels.

Joint-Embedding Predictive Architectures formulate SSL as prediction in latent embedding space, *i.e.*, predicting the embedding of a target signal from a compatible context signal through a predictor, rather than reconstructing pixels. By operating in the latent space, JEPAs encourage abstraction and provide a scalable predictive learning framework. I-JEPA [8], the image-based instantiation, predicts latent embeddings of masked target regions from visible context using a lightweight predictor, avoiding both pixel-level reconstruction and contrastive pairing. DMT-JEPA [15] changes the context/target embeddings by aggregating features of semantically similar neighboring patches, mainly changing what latent target is predicted. C-JEPA [16] further augments the JEPA formulation with contrastive objectives to improve training stability and invariance alignment, while LeJEPA [17] aims to provide a more principled and heuristic-free formulation.

Despite their strong abstraction capabilities, existing JEPA-based image pretraining methods treat target regions largely uniformly and predict them independently, without an explicit mechanism to determine which regions are most discriminative or in what order predictions should proceed. As shown in Figure 2, DSeq-JEPA aims to improve JEPA-based representation learning by introducing two complementary inductive biases: (i) *discriminative region prioritization* (*where to predict*), which biases prediction toward informative target regions instead of uniform target sampling, and (ii) *sequential latent prediction* (*in what order to predict*), which replaces flat unordered prediction with a structured sequential predictive process. Compared with generative autoregressive models

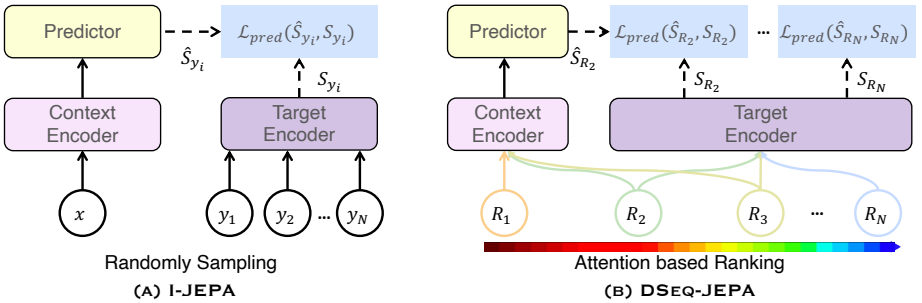

Fig. 2: (A) I-JEPA predicts embeddings of target regions $(y_1, \ldots, y_N)$ from a context region x, using a predictor network. In contrast, **(B) DSeq-JEPA** uses an attention map to identify and select primary discriminative regions (*where*), and then performs sequential latent prediction from the most discriminative to the least (*in what order*), through a predictor to predict the embeddings of the next discriminative regions $\{R_2, \ldots, R_N\}$ based on the sequence of identified regions.

such as iGPT [9], DSeq-JEPA operates in the latent space rather than reconstructing image tokens, and uses a discriminative region ordering rather than a fixed token order.

3 Methodology

3.1 Overall Framework

Our overall framework is illustrated in Figure 2. Our goal is to enhance self-supervised learning by introducing a preferential strategy for predictive region selection and an explicit prediction order. To this end, we propose a framework that explicitly models both *where* to predict and *in what order* predictions should proceed. As shown in Figure 3, the framework comprises two key components: *(1) Discriminative Region Prioritization*, which produces an attention-derived saliency map to identify informative regions (*where* to predict) and to derive importance scores for ranking the selected regions (*in what order* to predict); and *(2) Sequential Next-Region Embedding Prediction*, which performs latent prediction sequentially from the most discriminative region to the least.

3.2 Discriminative Region Prioritization

Given an input image, we first generate an attention-derived saliency map $\mathbf{A}$ using the target encoder, which provides spatial guidance for locating informative visual content. Specifically, the saliency map is computed from the similarity between the auxiliary class token and patch embeddings at a selected transformer block, highlighting patches that contribute strongly to global image semantics. Formally, given an image $\mathbf{x} \in \mathbb{R}^{C\times H\times W}$, we feed it into the target encoder $f_{\bar{\phi}}$, obtaining feature embeddings $\mathbf{s} = f_{\bar{\phi}}(\mathbf{x}) \in \mathbb{R}^{D\times h\times w}$, where D is the embedding

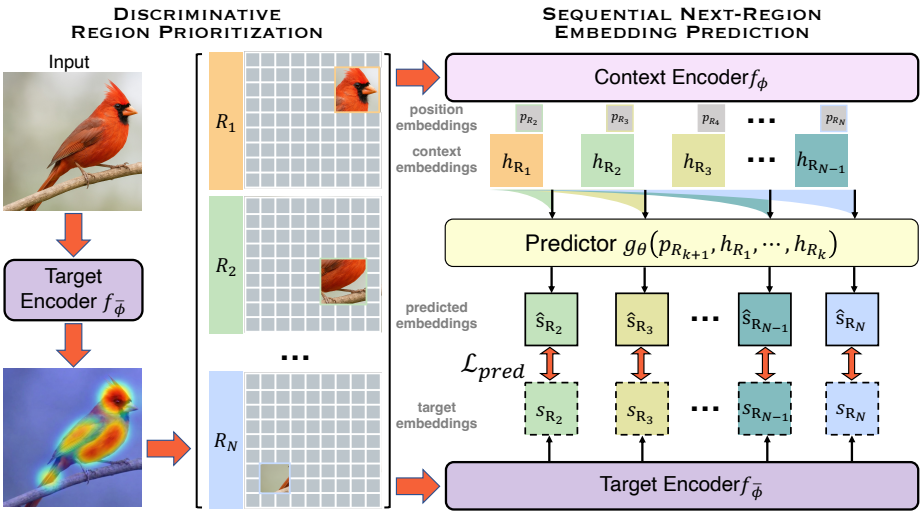

Fig. 3: Overview of DSeq-JEPA. Given an input image, the target encoder $f_{\bar{\phi}}$ produces an attention-derived saliency map used to select and rank the top N discriminative regions $\mathcal{R} = \{R_k\}_{k=1}^{N}$ (*Discriminative Region Prioritization*). The context encoder f_ϕ then encodes these regions, and the predictor g_θ performs *Sequential Next-Region Embedding Prediction*: at step k, it predicts the embedding of R_{k+1} from the context features of previous regions and the positional token $\mathbf{p}_{R_{k+1}}$, and aligns it with the corresponding target-encoder embedding using latent prediction loss $\mathcal{L}_{\text{pred}}$.

dimension, and h, w denote the number of vertical and horizontal patches, respectively. We compute a similarity map $\mathbf{A} \in \mathbb{R}^{h \times w}$ between the embeddings of the class token and all patches at the l-th transformer block. While we use this simple similarity-based attention for region selection, the formulation is general and can accommodate other attention mechanisms that provide semantically meaningful guidance (See Table 7). We then normalize $\mathbf{A}$ to $\tilde{\mathbf{A}}$ using Eq. (1),

$$\tilde{\mathbf{A}} = \frac{\mathbf{A} - \min(\mathbf{A})}{\max(\mathbf{A}) - \min(\mathbf{A})}, \tag{1}$$

and apply Otsu's method [18], which adaptively selects a data-driven threshold per image, to obtain a binary mask,

$$\mathbf{M}(i,j) = \mathbb{1}\left[\tilde{\mathbf{A}}(i,j) \geq \tau\right]. \tag{2}$$

We apply connected-component labeling to $\mathbf{M}$ under a relaxed connectivity criterion (8-neighborhood) to obtain the candidate set $\{R_k\}$. After removing small components ($|R_k| < 0.15hw$), each remaining region is assigned a discriminative score ρ_k defined as the average normalized saliency response within the region (Eq. (3)),

$$\rho_k = \frac{1}{|R_k|} \sum_{(i,j) \in R_k} \tilde{\mathbf{A}}(i,j). \tag{3}$$

Algorithm 1: Discriminative Region Prioritization

Input: Saliency map $\mathbf{A}$, current epoch t, total pre-training epochs T
Output: Discriminative regions $\{R_k\}_{k=1}^{N}$

Compute $\lambda \leftarrow \min\big(1, \max(0,\ t/T)\big)$;
Normalize $\mathbf{A}$ to $\tilde{\mathbf{A}}$ via Equation (1);
Obtain binary mask $\mathbf{M}$, connected regions $\{R_k\}$;
Compute discriminative score ρ_k for each region via Equation (3);
Select top-N regions $\mathcal{R} = \{R_k\}_{k=1}^{N}$;
for $k = 1$ **to** N **do**
    Sample $b_k \sim \text{Bernoulli}(\lambda)$;
    if $b_k = 1$ **then**
        $R_k := R_k$;
    else
        /* Randomly sample region with center(u, v) and size(w, h) */
        $R_k :=$ R(u,v,w,h);
return $\{R_k\}_{k=1}^{N}$

Unlike prior works [7, 16], we employ irregular, non-overlapping masks for the extracted regions. We rank regions by ρ_k, select the top $N-1$ regions as primary discriminative regions, and define the final region R_N as the remaining image area to ensure complete image coverage:

$$\mathcal{R} = \{R_k\}_{k=1}^{N}, \qquad R_N = \Omega \setminus \bigcup_{k=1}^{N-1} R_k, \tag{4}$$

where Ω denotes the full image region. This ensures complete coverage of the image without gaps, yielding more stable and discriminative region-based representation learning.

To improve stability when saliency maps are noisy early in pre-training, we adopt a probabilistic curriculum. Specifically, the probability λ of applying discriminative region selection is linearly increased from 0 to 1 over pre-training epochs, so that each sample uses our discriminative region selection with probability λ and random region sampling with probability $1 - \lambda$. This progressive transition stabilizes optimization and gradually shifts the model from diffuse attention to structured saliency (see Algorithm 1).

3.3 Sequential Next-Region Embedding Prediction

Given the ranked discriminative regions, DSeq-JEPA predicts the next-region embeddings sequentially from the most discriminative region to the least. Let $\mathcal{R} = \{R_k\}_{k=1}^{N}$ denote the selected regions sorted in descending order of discriminativeness. At each step $k \in \{1, \ldots, N-1\}$, the predictor g_θ estimates the latent embedding of the $(k+1)$-th region, $\hat{\mathbf{s}}_{R_{k+1}}$, conditioned on the representations of the preceding discriminative regions from the context encoder, $\mathbf{h}_{R_1}, \ldots, \mathbf{h}_{R_k}$,

together with the positional token $\mathbf{p}_{R_{k+1}}$ of the next target region:

$$\hat{\mathbf{s}}_{R_{k+1}} = g_\theta(\mathbf{p}_{R_{k+1}}, \mathbf{h}_{R_1}, \ldots, \mathbf{h}_{R_k}). \tag{5}$$

Compared with I-JEPA's flat, independent target prediction, this formulation introduces ordered dependencies across regions and yields a structured and ordered sequential latent prediction process.

3.4 Training Objective

To enforce the alignment between the predicted representation $\hat{\mathbf{s}}_{R_{k+1}}$ and its ground-truth counterpart $\mathbf{s}_{R_{k+1}}$, we adopt the smoothed ℓ_1 (Huber) loss. Formally, the sequential latent prediction objective is defined as

$$\mathcal{L}_{\text{pred}} = \frac{1}{N-1} \sum_{k=1}^{N-1} \sum_{j=1}^{D} \psi\Big(\big(\hat{\mathbf{s}}_{R_{k+1}} - \mathbf{s}_{R_{k+1}}\big)_j \Big), \tag{6}$$

$$\psi(x) = \begin{cases} \frac{1}{2}x^2, & |x| < \delta, \\ \delta\left(|x| - \frac{1}{2}\delta\right), & |x| \geq \delta, \end{cases} \qquad \delta = 1. \tag{7}$$

Here, D denotes the embedding dimension, and $(\cdot)_j$ indexes the j-th feature dimension. This objective encourages stable and discriminative alignment between predicted and target embeddings, making sequential latent prediction smooth while remaining robust to outliers.

4 Experiments

4.1 Evaluation setup

To comprehensively evaluate the effectiveness of DSeq-JEPA in learning discriminative and transferable visual representations, we pre-train all self-supervised models on ImageNet-1K [19] using the standard protocol from I-JEPA [8] and C-JEPA [16]. We evaluate on five downstream benchmark groups: (1) image classification (ImageNet-1K [19]); (2) fine-grained visual categorization (iNaturalist21 [20], CUB-200-2011 [21], and Stanford Cars [22]); (3) object detection and instance segmentation (MS-COCO [23]); (4) semantic segmentation (ADE20K [24]); and (5) low-level reasoning (CLEVR [25]). We follow I-JEPA/C-JEPA pre-training and evaluation protocols whenever applicable for fair comparison.

4.2 Main Results

Image classification and fine-grained visual categorization. Table 1 reports results for linear probing and full fine-tuning on ImageNet-1K, together with fine-grained visual categorization (iNat21, CUB, and Cars). Linear probing on ImageNet-1K is a standard SSL evaluation protocol and directly reflects

Table 1: Linear probing and full fine-tuning evaluations. We report *Top-1 accuracy* on image classification (ImageNet) and fine-grained visual categorization (iNat21, CUB, Cars). Average performance across all datasets is reported in the last column.

Method	Arch.	Epochs	Image Classification		FGVC			Overall Avg.
			Linear	Fine-tune	iNat21	CUB	Cars	
Methods without view data augmentations								
MAE [4]	ViT-B/16	1600	68.0	83.6	33.5	59.8	60.2	61.0
I-JEPA [8]	ViT-B/16	600	72.4	83.5	35.9	65.3	65.9	64.6
DSeq-JEPA	ViT-B/16	600	**73.5** +1.1	**84.0** +0.5	**36.4** +0.5	**66.2** +0.9	**67.3** +1.4	**65.5** +0.9
C-JEPA [16]	ViT-B/16	600	73.5	84.2	36.2	64.9	66.1	65.0
DSeq-C-JEPA	ViT-B/16	600	**73.8** +0.3	**84.3** +0.1	**36.6** +0.4	**66.5** +1.6	**67.4** +1.3	**65.7** +0.7
NEPA [26]	ViT-B/14	1600	-	83.8	-	-	-	-
MAE [4]	ViT-L/16	1600	76.0	85.9	37.1	63.4	61.9	-
I-JEPA [8]	ViT-L/16	600	77.1	86.6	38.8	66.9	68.1	67.5
DSeq-JEPA	ViT-L/16	600	**77.9** +0.8	**86.8** +0.2	**39.5** +0.7	**68.1** +1.2	**68.9** +0.8	**68.2** +0.8
C-JEPA [16]	ViT-L/16	600	78.0	86.9	39.0	65.8	67.7	67.5
DSeq-C-JEPA	ViT-L/16	600	**78.4** +0.4	**87.2** +0.3	**39.7** +0.7	**68.3** +2.5	**68.8** +1.1	**68.5** +1.0
NEPA [26]	ViT-L/14	800	-	85.3	-	-	-	-
MAE [4]	ViT-H/14	1600	77.2	-	-	-	-	-
I-JEPA [8]	ViT-H/16_{448}	300	81.1	87.1	38.9	67.2	67.7	68.4
DSeq-JEPA	ViT-H/16_{448}	300	**82.4** +1.3	**87.8** +0.7	**39.3** +0.4	**68.9** +1.7	**70.1** +2.4	**69.7** +1.5
Methods with view data augmentations								
LeJEPA [17]	ViT-H/14	100	79.0	-	-	-	-	-
DINO [27]	ViT-B/16	1600	78.2	82.8	30.2	33.6	27.0	50.4
iBOT [28]	ViT-B/16	1600	79.5	84.0	38.9	67.7	68.1	67.6

representation quality. Across *ViT-B/16, ViT-L/16, and ViT-H/16* backbones, DSeq-JEPA consistently improves over I-JEPA in linear probing (+1.1, +0.8, and +1.3 Top-1 accuracy, respectively) and also improves full fine-tuning (+0.5, +0.2, and +0.7, respectively). To further assess stability, we additionally repeat DSeq-JEPA under the ImageNet ViT-B/16 setting with three random seeds, obtaining 73.8±0.4 Top-1 accuracy, suggesting that the result is stable across runs. The gains extend to fine-grained benchmarks, where DSeq-JEPA improves over I-JEPA on all reported settings (e.g., ViT-B/16: +0.5 on iNat21, +0.9 on CUB, and +1.4 on Cars; ViT-L/16: +0.7, +1.2, and +0.8, respectively), yielding a consistent improvement in the FGVC average. Moreover, at larger scale, DSeq-JEPA with ViT-H/16 substantially outperforms LeJEPA [17] by +3.4 points in ImageNet linear probing. These trends support our hypothesis that discriminative region prioritization and sequential next-region embedding prediction improve both global semantic separation and fine-grained cue discrimination.

Compared with C-JEPA, which augments I-JEPA with contrastive regularization, DSeq-JEPA achieves comparable ImageNet performance under both linear probing and full fine-tuning, while consistently outperforming C-JEPA on fine-grained benchmarks (iNat21/CUB/Cars), indicating that discriminative sequential prediction learns more part-aware and discriminative representations. When adding the same contrastive regularization, DSeq-C-JEPA surpasses C-JEPA across settings, suggesting that contrastive alignment and discriminative sequential prediction are complementary.

Table 2: Performance on detection and segmentation tasks. We report AP^{box}, AP^{mask} on MS-COCO, and $mIoU$ on ADE20K.

Method	Arch.	Epochs	AP^{box}	AP^{mask}	mIoU
Methods without view data augmentations					
MAE [4]	ViT-B/16	1600	50.3	44.9	48.1
I-JEPA [8]	ViT-B/16	600	49.9	44.5	47.6
DSeq-JEPA	ViT-B/16	600	**50.5** +0.6	**45.0** +0.5	**48.1** +0.5
C-JEPA [16]	ViT-B/16	600	50.7	45.3	48.7
DSeq-C-JEPA	ViT-B/16	600	**50.9** +0.2	**45.7** +0.4	**48.9** +0.2
NEPA [26]	ViT-B/14	1600	-	-	48.3
Methods with view data augmentations					
DINO [27]	ViT-B/16	1600	50.1	43.4	46.8
iBOT [28]	ViT-B/16	1600	51.2	44.2	50.0

Relative to representative SSL baselines (MAE, DINO and iBOT), DSeq-JEPA remains competitive under a simpler JEPA-style single-view pre-training recipe, compared to reconstruction- and multi-view-based baselines. On ViT-B/16, DINO and iBOT (with multi-view augmentations) achieve higher ImageNet linear-probing accuracy, whereas DSeq-JEPA remains competitive and yields strong fine-tuning performance despite relying on a simpler single-view recipe. At larger scale, DSeq-JEPA with ViT-H attains the highest ImageNet accuracy under both linear probing and full fine-tuning. In addition, DSeq-JEPA is training-efficient: it uses a *single-view* recipe (vs. DINO/iBOT's *multi-view augmentation*) and runs for *600/300* epochs, whereas MAE/DINO/iBOT require *1600* epochs. This efficiency advantage of JEPA-style single-view pre-training is consistent with observations from I-JEPA, where a large ViT-H/14 I-JEPA model requires less compute than a smaller iBOT ViT-S/16 model [7].

Additionally, we compare with NEPA [26], a recent sequential embedding-prediction baseline. Although the training configurations are not strictly matched, both DSeq-JEPA and DSeq-C-JEPA achieve higher ImageNet fine-tuning accuracy on both ViT-B (84.0/84.3 vs. 83.8) and ViT-L (86.8/87.2 vs. 85.3), suggesting that semantically grounded next-region prediction is more effective than generic sequential embedding prediction.

Detection and Segmentation. To further evaluate transferability to dense prediction, we evaluate object detection and instance segmentation on MS-COCO [23] and semantic segmentation on ADE20K [24]. As shown in Table 2, DSeq-JEPA consistently improves over I-JEPA across all reported metrics (+0.6 AP^{box}, +0.5 AP^{mask}, and +0.5 mIoU), while DSeq-C-JEPA achieves the best overall performance among JEPA variants, surpassing both I-JEPA and C-JEPA. DSeq-C-JEPA also exceeds NEPA on ADE20K mIoU (48.9 vs. 48.3), suggesting that discriminative ordering benefits not only sequential prediction but also spatial generalization. Despite using a lighter single-view 600-epoch

Table 3: Performance on low-level tasks. We report *Clevr/Count* (object counting) and *Clevr/Dist* (distance estimation) with a linear probe on frozen ViT-L/16 features, measured by *Top-1 accuracy*.

Metric	I-JEPA	DSeq-JEPA	C-JEPA	DSeq-C-JEPA
Clevr/Count	85.6	**86.4** +0.8	86.8	**87.1** +0.3
Clevr/Dist	71.2	**71.5** +0.3	71.6	**71.9** +0.3

Table 4: Effects of region generation and prediction strategy (ViT-B/16).

Region Generation		Prediction Strategy		ImageNet	iNat21
Uniform sampling	*Discriminative selection*	*Flat*	*Sequential*		
✓		✓		72.4	35.9
✓			✓	72.3	34.9
	✓	✓		72.0	35.7
	✓		✓	**73.5**	**36.4**

recipe, DSeq-JEPA/DSeq-C-JEPA outperform MAE and DINO across all reported dense-prediction metrics, and DSeq-C-JEPA achieves higher AP^{mask} than iBOT. Overall, these results indicate that discriminative sequential prediction promotes more spatially structured and semantically coherent representations that transfer effectively beyond classification to pixel-level prediction.

Low-level Reasoning. As shown in Table 3, on low-level reasoning tasks, DSeq-JEPA achieves 86.4 on Clevr/Count and 71.5 on Clevr/Dist [25] with ViT-L/16, improving over I-JEPA by +0.8 and +0.3, respectively. With contrastive regularization, DSeq-C-JEPA further improves to 87.1/71.9, surpassing C-JEPA on both tasks. The fact that improvements persist on these low-level metrics indicates that DSeq-JEPA does not merely enhance semantic categorization, but also strengthens geometric and structural priors in the learned representation.

4.3 Ablation Study

Do discriminative selection and sequential prediction act synergistically? We ablate the two key components of DSeq-JEPA—*discriminative region prioritization* and *sequential next-region embedding prediction*—by varying region generation (uniform vs. discriminative) and prediction strategy (flat vs. sequential). Table 4 reports ImageNet and iNat21 linear-probing accuracy with a ViT-B/16 backbone. The baseline configuration (uniform sampling + flat prediction) corresponds to I-JEPA, which treats all regions equally and predicts them in parallel. Enabling either component in isolation degrades performance: sequential prediction without an informative order (uniform + sequential) breaks permutation symmetry arbitrarily and introduces noisy supervision, while discriminative selection without sequential conditioning (selective + flat) collapses

Table 5: Effects of different prediction orders (ViT-B/16).

Order scheme	Flat	Random	Spatial	Inverse	Truncating	**DSeq (ours)**
ImageNet	72.0	71.7	72.7	71.3	73.0	**73.5**

inter-region dependencies into independent targets. In contrast, combining both components (selective + sequential) yields the best results. These findings show that the modules are not independently effective but *synergistic*: discriminative region selection supplies a semantically meaningful trajectory (where), and sequential prediction exploits this trajectory to capture region-to-region dependencies (in what order).

How important is the prediction order? We next isolate the effect of prediction order while keeping the same set of discriminative regions. Table 5 compares several ordering schemes on ImageNet with ViT-B/16: *Flat* (I-JEPA-style, no autoregression), *Random* (random region order), *Spatial* (regions sorted by center coordinates in row-major order), *Truncating* (predict only up to Top-3 regions), and our *DSeq* order (DSeq-JEPA). We observe that random ordering harms performance, indicating that autoregression without a meaningful trajectory forces the model to learn from mixed-difficulty signals. A simple spatial order brings a measurable improvement (+0.7 over flat), suggesting that purely geometric structure provides limited benefit. The inverse order only reaches 71.3, showing that the direction of the prediction trajectory matters. Our discriminative sequential order achieves the highest accuracy (+1.5 over flat), showing that encoding a semantically grounded prediction trajectory over regions is crucial for reaping the benefits of sequential prediction.

Furthermore, the step-truncation diagnostic shows that predicting only up to the Top-3 regions (*Truncating*) yields 73.0 vs. 73.5 with the full chain (Top-5), indicating that later, lower-saliency regions still provide complementary supervision beyond the most salient cues. Together, these ablations support our design choice of using a discriminative, full-length prediction order rather than generic or truncated autoregression. Viewed from a learning-dynamics perspective, the discriminative order induces an implicit easy-to-hard curriculum: early steps focus on stable, highly informative regions, and later steps progressively integrate more context-dependent cues. To quantify this effect, we analyze the per-step prediction loss along the next-region prediction sequence using DSeq-JEPA ViT-B checkpoint at epoch 450. We sample 10,000 ImageNet images and compute the average loss for each Top-k region prediction. As shown in Figure 4, early steps (Top-2 and Top-3) have lower prediction error, whereas later steps (Top-4 and Top-5) are clearly harder with noticeably higher loss. This trend suggests that the model first learns to predict stable, highly informative regions and then progressively integrates weaker or more context-dependent cues, and the sequential predictor exploits this ordered trajectory to shape the learned representations.

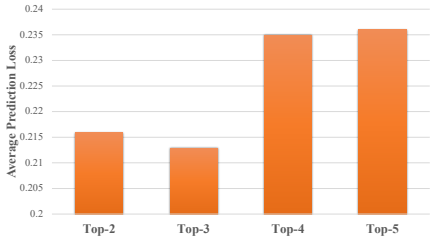

Fig. 4: Per-step prediction difficulty (ViT-B/16 checkpoint at epoch 450).

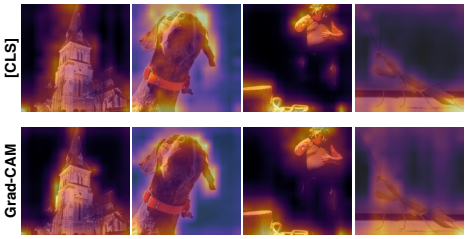

Fig. 5: `[CLS]`-based saliency vs. Grad-CAM.

Table 6: Effect of `[CLS]` token.

Method	I-JEPA	+`[CLS]`
ImageNet	72.4	72.4

Table 7: Robustness to attention proxies.

Method	Grad-CAM-based	`[CLS]`-based
ImageNet	73.4	73.5

Effect of the `[CLS]` Token. To keep the architecture comparable to I-JEPA, we introduce an auxiliary `[CLS]` token that is used only as an anchor to compute saliency maps, without adding extra layers or parameters. As shown in Table 6, adding this token does not improve I-JEPA's performance, indicating that it provides no representational gain by itself. Thus, the improvements of DSeq-JEPA stem from the proposed discriminative region prioritization and sequential next-region prediction, rather than from any architectural modification.

Robustness to Attention Proxies. Table 7 assesses whether DSeq-JEPA depends on a specific saliency estimator by replacing the `[CLS]`–based proxy with a label-free Grad-CAM–style proxy [29]. This variant attains 73.4 ImageNet linear-probing accuracy (vs. `[CLS]`–based: 73.5), indicating that our gains are robust to the choice of attention proxy. Figure 5 shows that our `[CLS]`-based saliency closely overlaps with the Grad-CAM proxy, achieving a Top-20 patch IoU of 0.41 while being simpler and gradient-free. Together with the prediction-order ablations in Table 5, this suggests that performance is driven not by a particular saliency mechanism, but by exploiting a semantically meaningful prediction trajectory: arbitrary orderings can hurt performance, whereas orderings that better correlate with discriminative content consistently yield gains, and the proxy simply provides a practical way to approximate such an ordering prior.

Sensitivity to the Number of Regions N. We fix the region number to $N = 5$, following the I-JEPA setting. To assess sensitivity to this choice, we vary $N \in \{3, 5, 7\}$ for DSeq-JEPA with ViT-B/16 and pre-train on ImageNet under the same protocol. The resulting ImageNet linear-probing accuracies are 72.9, 73.5, and 73.4, respectively. The overall performance fluctuates only mildly, indicating that DSeq-JEPA is not sensitive to the exact number of discriminative regions.

Table 8: Pre-training overhead and inference efficiency.

Method	Pre-training			Inference	
	Time(h)	Mem(GB)	Total GFLOPs	#Params(M)	GFLOPs/image
I-JEPA	24.2	31.5	96.4	86.6	17.7
C-JEPA	24.5	31.5	98.0	86.6	17.7
DSeq-JEPA	26.5	38.2	111.6	86.6	17.8
DSeq-C-JEPA	26.8	38.2	117.2	86.6	17.8

Using too few regions ($N = 3$) slightly under-utilizes contextual information, while increasing to $N = 7$ brings no further benefit.

4.4 Pre-training Overhead and Inference Efficiency

Table 8 reports pre-training and inference costs on ImageNet with ViT-B/16 under the same hardware setup. Notably, compared with I-JEPA/C-JEPA, DSeq-JEPA's inference complexity remains essentially unchanged (86.6M parameters; 17.7 vs. 17.8 GFLOPs/image), since discriminative sequential prediction is only used during pre-training. The modest added computation introduced by DSeq-JEPA is confined to pre-training: it adds about +2.0 hours wall-clock time (24.2/24.5 $\rightarrow$ 26.5h), increases total training compute from 96.4/98.0 to 111.6 GFLOPs, and raises peak memory usage from 31.5 to 38.2 GB due to sequential region tokens and causal masking. DSeq-C-JEPA follows the same pattern.

4.5 Visualization Analysis

Qualitative Comparison with I-JEPA. As shown in Figure 6, I-JEPA yields relatively diffuse attention maps and rectangular target masks that often cover multiple, partially overlapping regions. In contrast, DSeq-JEPA produces compact, semantically meaningful regions ordered by discriminativeness, progressively focusing on key object parts (e.g., the bird's beak and forehead) and yielding clearer region correspondences and more interpretable predictions.

Emergence of Semantic Structure during Pre-training. Figure 7 visualizes the evolution of patch-level clustering as pre-training proceeds. Early in training, clusters are fragmented and noisy; over time, they self-organize into coherent, object-aligned groups, even in the absence of labels. This illustrates how DSeq-JEPA's representations self-organize during training: patches with similar semantics gradually move closer in the learned embedding space. This progression indicates that DSeq-JEPA gradually induces fine-grained semantic structure via its attention-guided, sequential prediction mechanism.

5 Conclusion

We introduced DSeq-JEPA, a JEPA-based SSL framework that augments latent prediction with *discriminative region prioritization* and *sequential next-region*

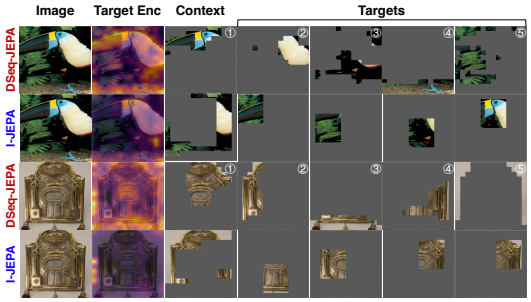

Fig. 6: Qualitative visualization of learned attention and selected context/target regions using ViT-B/16 model. For DSeq-JEPA, the numbered regions (①–⑤) correspond to discriminative regions ordered by their estimated importance.

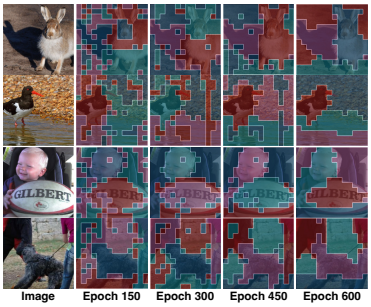

Fig. 7: Evolution of patch-level clustering during pre-training. From left to right: input image and 4-cluster results from ViT-B/16 checkpoints at 150, 300, 450, and 600 epochs.

embedding prediction. By explicitly modeling both *where* to attend and *in what order* to predict, DSeq-JEPA progressively activates primary cues and then integrates secondary contexts, yielding more structured and fine-grained representations. Comprehensive evaluations across multiple benchmarks demonstrate consistent gains over JEPAs, confirming DSeq-JEPA's effectiveness and generality. In future work, we plan to extend these ideas to vision–language pre-training, enabling more selective visual grounding and structured cross-modal alignment.

Acknowledgements

This work was funded, in part, by the UBC Data Science Institute, UBC AI & Health Network, Mitacs, Vector Institute for AI, Canada CIFAR AI Chair, NSERC Canada Research Chair (CRC), NSERC Discovery Grant, NSF CAREER Award No. 2441060 and NSF (2330423) & NSERC (585136) Global Center on AI and Biodiversity Change. Resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, the Digital Research Alliance of Canada, companies[1] sponsoring the Vector Institute. Additional hardware support was provided by John R. Evans Leaders Fund CFI grant and Digital Research Alliance of Canada under the Resource Allocation Competition award. This work also benefited from JSPS KAKENHI Grant-in-Aid for Scientific Research (B) and (C) under Grant 26K02879 and 23K11164, and from French state funds managed by the National Research Agency via the IHU Strasbourg (ANR-10-IAHU-0002) and the ENACT AI Cluster (ANR-23-IACL-0004).

[1] https://vectorinstitute.ai/#partners

References

1. Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020.
2. 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*, pages 9729–9738, 2020.
3. Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.
4. Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.
5. Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.
6. Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9653–9663, 2022.
7. Yann LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022.
8. 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*, pages 15619–15629, 2023.
9. Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International conference on machine learning*, pages 1691–1703. PMLR, 2020.
10. Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9640–9649, 2021.
11. Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.
12. Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.
13. Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021.
14. Tianyu Hua, Yonglong Tian, Sucheng Ren, Michalis Raptis, Hang Zhao, and Leonid Sigal. Self-supervision through random segments with autoregressive coding (randsac). In *International Conference on Learning Representations*, 2022.
15. Shentong Mo and Sukmin Yun. Dmt-jepa: Discriminative masked targets for joint-embedding predictive architecture. *arXiv preprint arXiv:2405.17995*, 2024.

16. Shentong Mo and Shengbang Tong. Connecting joint-embedding predictive architecture with contrastive self-supervised learning. *Advances in neural information processing systems*, 37:2348–2377, 2024.
17. Randall Balestriero and Yann LeCun. Lejepa: Provable and scalable self-supervised learning without the heuristics. *arXiv preprint arXiv:2511.08544*, 2025.
18. Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-9(1):62–66, 1979.
19. Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of IEEE/CVF conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
20. Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018.
21. Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.
22. Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 554–561, 2013.
23. Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European conference on computer vision*, pages 740–755. Springer, 2014.
24. Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of IEEE/CVF conference on computer vision and pattern recognition*, pages 633–641, 2017.
25. Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of IEEE/CVF conference on computer vision and pattern recognition*, pages 2901–2910, 2017.
26. 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. *arXiv preprint arXiv:2512.16922*, 2025.
27. Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.
28. Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. Image bert pre-training with online tokenizer. In *International Conference on Learning Representations*, 2022.
29. Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.