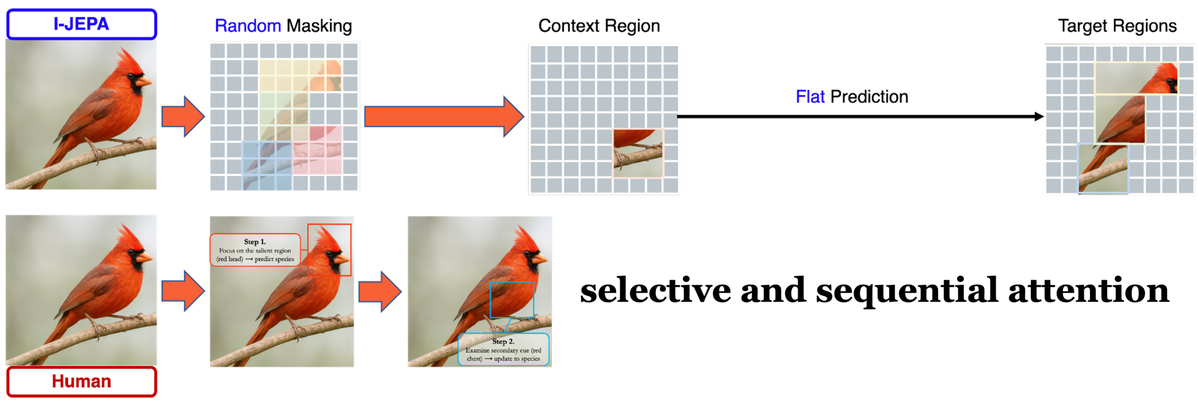
DSeq-JEPA: Discriminative Sequential Joint-Embedding Predictive Architecture

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] Leonid Sigal[1,2,7]

[1]University of British Columbia, [2]Vector Institute for AI, [3]University of Fukui, [4]Massachusetts Institute of Technology, [5]University of Strasbourg, [6]IHU Strasbourg, [7]Canada CIFAR AI Chair,
* Equal contribution xiangteng.he@ubc.ca sshunsuke0102@gmail.com

Why Rethink JEPA Target Prediction?



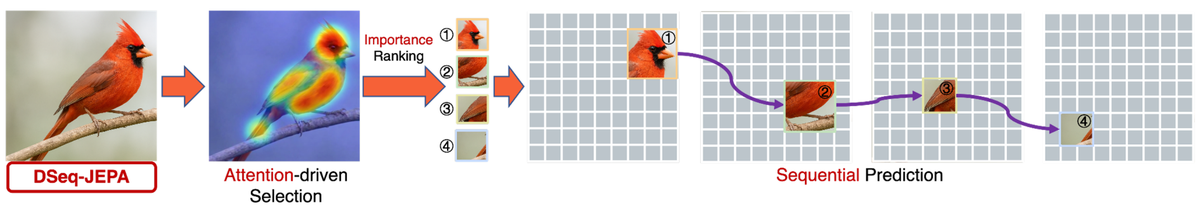
I-JEPA learns visual representations by predicting the latent embeddings of masked image regions from visible context.

However, its target prediction has two important properties:

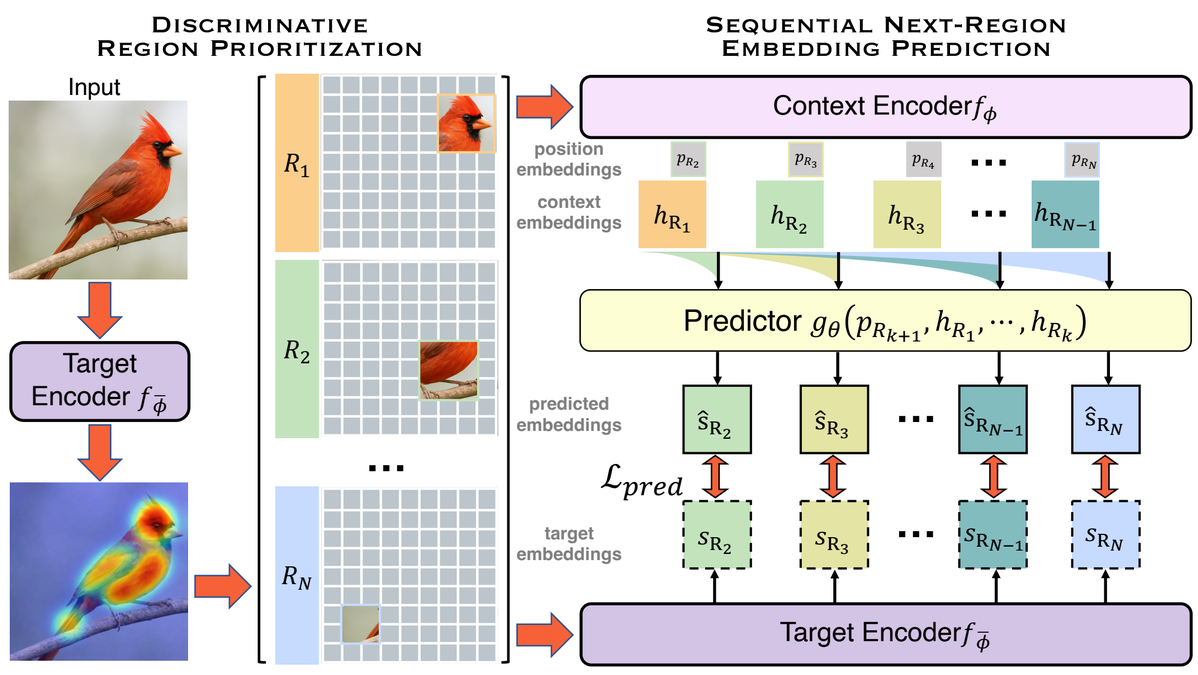
- **Uniform:** target regions are sampled without explicitly considering their visual importance.
- **Independent:** multiple target regions are predicted in parallel, without modeling dependencies among them.

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

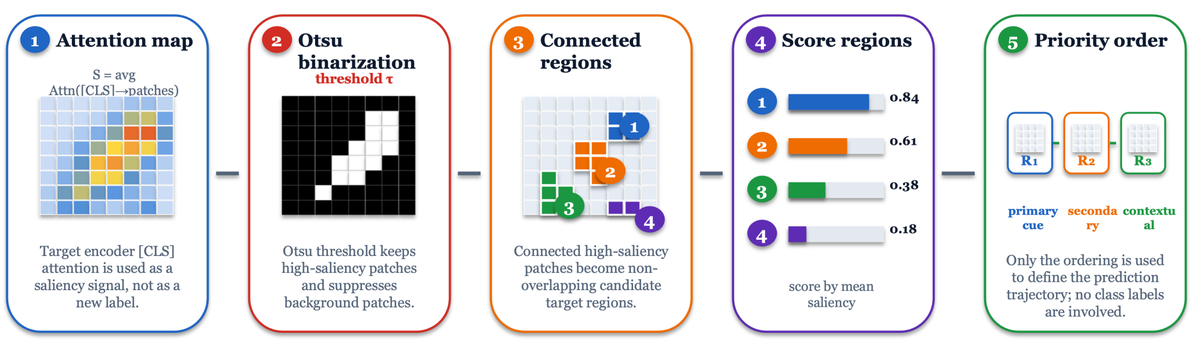
DSeq-JEPA: Discriminative Sequential Prediction



DSeq-JEPA turns JEPA target prediction from an unordered set prediction into a discriminatively ordered sequence prediction.



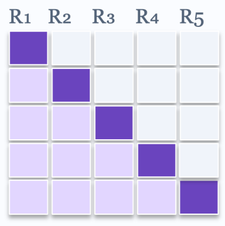
① Where to Predict: Discriminative Region Prioritization



We convert the target encoder's attention into a compact set of **informative, irregular, non-overlapping target regions**.

② In What Order: Sequential Latent Prediction

Autoregressive prediction in latent space



causal target schedule

The predictor sees a progressively richer target history instead of four independent target slots.

Instead of predicting all targets independently, DSeq-JEPA constructs a **target trajectory**: predict the next region embedding conditioned on context and earlier regions.

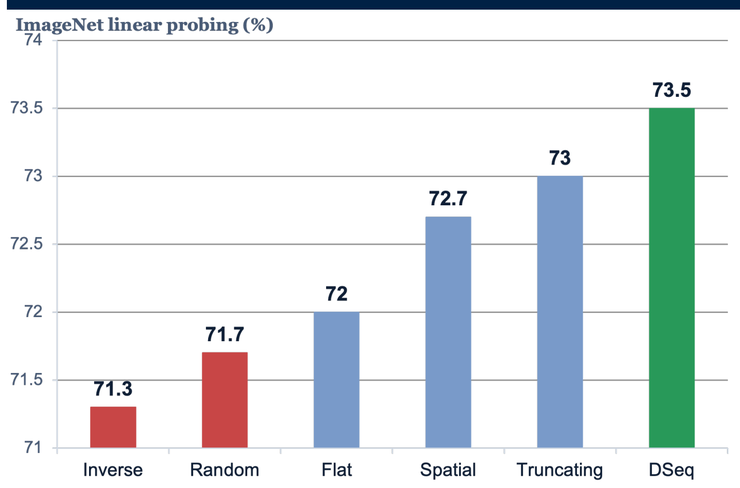
Ablation: Selection and Sequence Are Synergistic

Discriminative selection and sequential prediction are synergistic.

	Flat	Sequential
Uniform regions	**Baseline** ImageNet 72.4, iNat21 35.9	**Seq. only** ImageNet 72.3, iNat21 34.9
Discrim. regions	**Regions only** ImageNet 72.0, iNat21 35.7	**DSeq-JEPA** ImageNet 73.5, iNat21 36.4

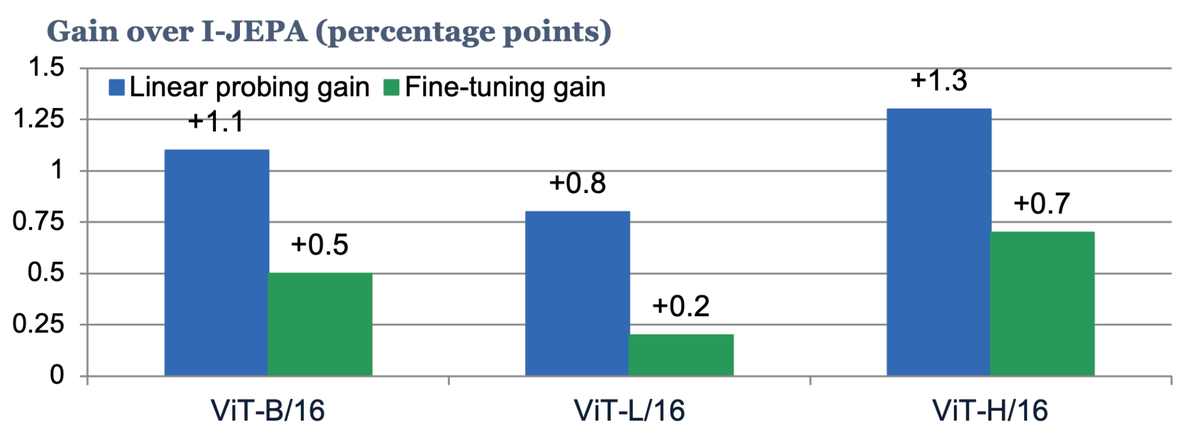
- **Sequential only:** an arbitrary sequence can introduce noisy supervision.
- **Selection only:** informative regions are still treated as independent targets.
- **DSeq-JEPA:** selection defines a meaningful prediction path, while sequential prediction learns dependencies along this path.

Ablation: The Prediction Order Matters

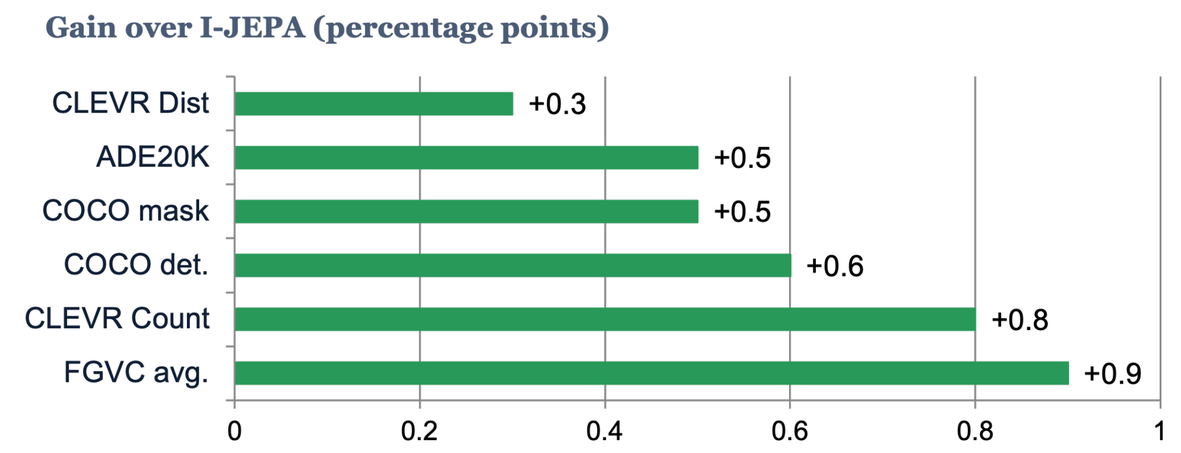


1. **A poor sequence can even hurt the representation** — **Random** and **inverse** ordering fall below **flat** prediction
2. **A weak visual prior gives a partial gain** — **Spatial** ordering improves over flat prediction
3. **Discriminative order provides the strongest curriculum** — **DSeq** reaches 73.5

Stronger Representations Across Model Scales



Improvements Transfer Beyond ImageNet



Take-Home Message

1. **Where matters.** Predicting informative visual regions provides a meaningful target space.
2. **Order matters.** A semantically grounded sequence creates a structured latent prediction trajectory.
3. **They matter together.** Discriminative region selection defines the path; sequential prediction learns along it.

DSeq-JEPA = JEPA + Discriminative Target Selection + Sequential Latent Prediction

Resources

Project Page
dseqjepa-project.com

Paper
DSeq-JEPA

Code & Models
GitHub Repository