Where Do Affordances Crystallize? Video vs. Image Self-Supervised Encoders for Embodied Perception

Abrar Zahin Raihan Aurchi Chowdhury
Bangladesh University of Engineering and Technology
Dhaka, Bangladesh

Abstract

Self-supervised video models are strong backbones for robotic perception. We probe 25 layer positions of architecturally matched ViT-L encoders (V-JEPA 2, V-JEPA 2.1, DINOv2) on the UMD Part Affordance Dataset and show that video *SSL models crystallize affordance signal at layer 19—four layers earlier than DINOv2 (image SSL, layer 23), suggesting a consistent effect of temporal predictive objectives on representation depth. Both V-JEPA variants share the same peak depth, yet V-JEPA 2.1's intermediate-layer supervision yields markedly stronger early-layer representations (layer 2 mAP:* 92.2% *vs.* 85.0% *for V-JEPA 2). Head ablation at each peak layer further reveals a sparse set of attention heads that concentrate the affordance signal; keeping only the top-3 heads achieves a* $5.3\times$ *attention compute reduction with* $\leq 0.1\%$ *mAP loss.*

1. Introduction

Robotic manipulation requires understanding object affordances—the action possibilities an object offers [7]. Self-supervised video models, especially the JEPA family [2], transfer strongly to robotic tasks without task-specific fine-tuning. Video SSL models like V-JEPA learn implicit world models through temporal prediction [4], making them natural candidates for embodied perception. V-JEPA 2.1 [6] extends V-JEPA 2 with a Dense Predictive Loss applied hierarchically at intermediate encoder layers, yielding a 20-point grasping improvement—yet its evaluation considers only final-layer features, leaving internal representational structure unexamined. DINOv2 [8] provides our static-image SSL baseline. Linear probes [1] and attention head pruning [3, 5] are established interpretability tools; to our knowledge, we apply them to affordance features in frozen video encoders for the first time.

We make three contributions: (1) a layer-wise probe study (layers 0–24) showing video SSL models crystallize affordance signal four layers earlier (layer 19) than DINOv2 (layer 23), with V-JEPA 2.1's intermediate supervision producing stronger early-layer representations despite identical peak depth; (2) an attention head ablation at each model's peak layer revealing sparse "affordance heads"; and (3) a pruning analysis showing $\approx 5.3\times$ attention compute reduction at $\leq 0.1\%$ mAP loss.

2. Method

Layer-wise probing. For a frozen ViT-L encoder ($L{=}24$ blocks, $D{=}1024$, $H{=}16$ heads), we extract representations at 25 positions ($\ell{=}0$: patch embedding; $\ell{=}1\ldots24$: block outputs). DINOv2 uses the CLS token; JEPA models use mean-pooled patch tokens. A one-vs-rest logistic regression ($C{=}1$, L-BFGS) is trained per layer to predict each of the 7 binary affordance labels; mAP is reported on the held-out test set.

Attention head ablation. At peak layer ℓ^*, we ablate head h by zeroing its $d_h{=}64$ output columns before the output projection [5]; importance is $\Delta_h = \text{mAP}_{\text{full}} - \text{mAP}_{-h}$. We sweep top-$k$ head subsets to quantify the efficiency–accuracy tradeoff; concentration is measured as each head's share of the total positive Δ_h mass.

Experimental setup. UMD Part Affordance Dataset [7] (7 affordance classes): stratified 2,000-image subsample (80/10/10 split, $224{\times}224$, ImageNet normalization). Static images are replicated $T{=}8$ times as pseudo-video for JEPA models. A random ViT-L establishes the chance-level floor. All encoders are **frozen**; only probe weights are trained.

3. Results

Layer-wise probing. Figure 1 shows mAP across all 25 layer positions. Both V-JEPA 2 and V-JEPA 2.1 peak at layer 19 (mAP$=99.7\%$ and 98.3%), four layers before DINOv2 (layer 23, mAP$=99.9\%$), attributable to their *video* predictive objective. V-JEPA 2.1's slightly lower peak probe mAP is consistent with its intermediate supervision producing representations optimized for fine-tuning rather than

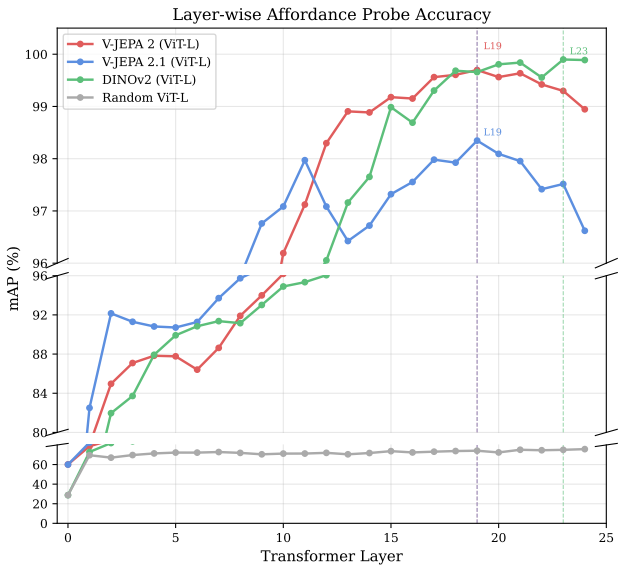


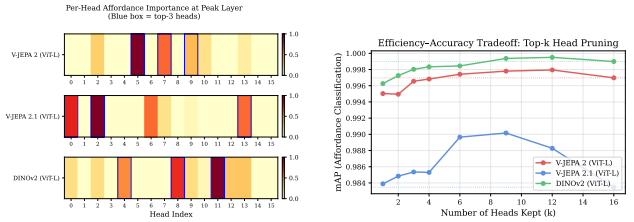
Figure 1. Layer-wise mAP across 25 positions. Dashed lines mark peak layers. Both V-JEPA models peak at layer 19; DINOv2 peaks at layer 23.

Table 1. Layer-wise linear probe at each model's peak layer. Values are Average Precision (%). Bold = best per column.

Model	Peak Layer	*grasp*	*cut*	*scoop*	*wrap-grasp*	*poke*	*support*	*contain*	mAP
V-JEPA 2 (ViT-L)	19	99.6	**99.8**	98.5	100.0	**100.0**	**100.0**	**100.0**	99.7
V-JEPA 2.1 (ViT-L)	19	99.3	99.1	95.3	99.9	98.4	96.4	**100.0**	98.3
DINOv2 (ViT-L)	23	**99.8**	99.5	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**99.9**
Random ViT-L	24	97.1	92.5	65.9	87.6	44.9	61.5	80.8	75.8

linear separability of frozen features. Its intermediate supervision nonetheless yields markedly stronger early-layer representations: layer 2 mAP is 92.2% vs. 85.0% for V-JEPA 2 (+7.2 pp), converging to the same peak. The random baseline peaks at 75.8% overall, well below all trained encoders; its inflated AP on frequent classes (*e.g.*, *grasp*: 97.1%) reflects label prevalence rather than learned structure. Table 1 summarizes peak-layer results.

Head ablation & pruning efficiency. Figures 2a and 2b show per-head importance and the efficiency–accuracy tradeoff at each model's peak layer. The top-3 heads account for 76%, 74%, and 63% of total positive affordance importance for V-JEPA 2, V-JEPA 2.1, and DINOv2, confirming sparse signal concentration. For V-JEPA 2.1, all $k \geq 1$ subsets meet or exceed full-model mAP, consistent with intermediate supervision creating partially redundant head roles. Keeping $k=3$ heads ($\approx 5.3\times$ theoretical attention FLOP reduction) costs only $\leq 0.1\%$ mAP for V-JEPA 2 and DINOv2; V-JEPA 2.1 gains $+0.19$ pp (Table 2).



(a) Head importance Δ_h. Blue: top-3.

(b) mAP vs. heads kept k.

Figure 2. Head ablation at each model's peak layer (left: per-head importance; right: pruning efficiency, dotted = full-model baseline).

Table 2. Head pruning efficiency at the peak layer. mAP (%) when keeping only top-k affordance heads. ΔmAP (pp): positive = gain over full model. Approx. speedup $\approx H/k$ where $H = 16$ heads.

Model	k heads	mAP	ΔmAP (pp)	Approx. Speedup
V-JEPA 2 (ViT-L)	1	99.5	−0.19	16.0×
V-JEPA 2 (ViT-L)	2	99.5	−0.20	8.0×
V-JEPA 2 (ViT-L)	3	99.7	−0.04	5.3×
V-JEPA 2 (ViT-L)	4	99.7	−0.01	4.0×
V-JEPA 2 (ViT-L)	6	99.7	+0.04	2.7×
V-JEPA 2 (ViT-L)	9	99.8	+0.08	1.8×
V-JEPA 2 (ViT-L)	12	99.8	+0.10	1.3×
V-JEPA 2 (ViT-L)	16	99.7	0.00	1.0×
V-JEPA 2.1 (ViT-L)	1	98.4	+0.04	16.0×
V-JEPA 2.1 (ViT-L)	2	98.5	+0.14	8.0×
V-JEPA 2.1 (ViT-L)	3	98.5	+0.19	5.3×
V-JEPA 2.1 (ViT-L)	4	98.5	+0.18	4.0×
V-JEPA 2.1 (ViT-L)	6	99.0	+0.62	2.7×
V-JEPA 2.1 (ViT-L)	9	99.0	+0.67	1.8×
V-JEPA 2.1 (ViT-L)	12	98.8	+0.48	1.3×
V-JEPA 2.1 (ViT-L)	16	98.3	0.00	1.0×
DINOv2 (ViT-L)	1	99.6	−0.27	16.0×
DINOv2 (ViT-L)	2	99.7	−0.17	8.0×
DINOv2 (ViT-L)	3	99.8	−0.10	5.3×
DINOv2 (ViT-L)	4	99.8	−0.07	4.0×
DINOv2 (ViT-L)	6	99.8	−0.05	2.7×
DINOv2 (ViT-L)	9	99.9	+0.04	1.8×
DINOv2 (ViT-L)	12	100.0	+0.05	1.3×
DINOv2 (ViT-L)	16	99.9	0.00	1.0×

4. Conclusion

Both V-JEPA models crystallize affordance signal four layers earlier (layer 19) than DINOv2 (layer 23), attributable to their video-domain predictive objectives. Within the JEPA family, V-JEPA 2.1's intermediate supervision further improves early-layer representation quality (+7.2 pp mAP at layer 2) without shifting peak depth. Pruning to the top-3 heads at each model's peak layer achieves a $\approx 5.3\times$ theoretical FLOP reduction in peak-layer attention with $\leq 0.1\%$ mAP loss, offering a practical path to efficient affordance-aware robot perception.

References

[1] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016. 1

[2] Adrien Bardes, Quentin Garrido, Jean Ponce, Michael Rabbat, Yann LeCun, and Mahmoud Assran. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025. 1

[3] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. 1

[4] Yann LeCun. A path towards autonomous machine intelligence. *OpenReview*, 2022. 1

[5] Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 1

[6] Lorenzo Mur-Labadia, Adrien Bardes, et al. V-JEPA 2.1: Unlocking dense features in video self-supervised learning. *arXiv preprint arXiv:2603.14482*, 2026. 1

[7] Austin Myers, Ching L. Teo, Cornelia Fermüller, and Yiannis Aloimonos. Affordance detection of tool parts from geometric features. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 1374–1381, 2015. 1

[8] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. 1