Self-Supervised Learning for Robot Grasping

Jed Guzelkabaagac
Technical University of Munich
Munich, Germany
jed.guzelkabaagac@tum.de

Boris Petrović
Technical University of Munich
Munich, Germany
boris.petrovic@tum.de

***Abstract*—Robotic grasping of novel objects remains difficult: large 3D shape variability and the cost of collecting labeled grasps hinder generalization. Self-supervised learning (SSL) on unlabeled point clouds offers transferable representations for downstream grasp prediction. While data2vec demonstrates modality-agnostic distillation in vision, speech, and language [1], the impact of 3D SSL encoders—such as Point–JEPA [2] and Point2Vec [3]—on robotic grasping remains underexplored. Our objective is to test whether SSL pretraining improves grasp-joint prediction, particularly in low-label regimes. We compare representative approaches (masked prediction, self-distillation, and augmentation strategies) within a unified pipeline and evaluate label efficiency under fixed object-level splits.**

I. Related Work

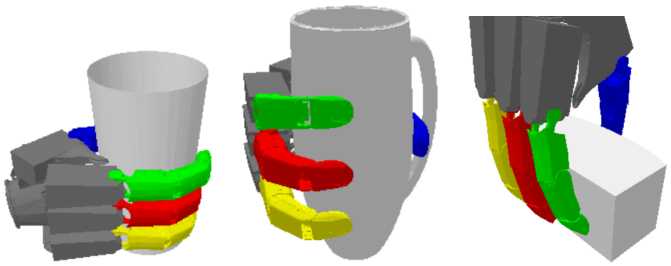
Self-supervised learning (SSL) for 3D data has largely progressed along three directions.

(i) Reconstruction-based methods learn by masking inputs and reconstructing them in the input space. On point clouds and LiDAR, this includes point/voxel masked autoencoding; e.g., Voxel-MAE reconstructs masked voxels for sparse automotive LiDAR and improves downstream tasks with fewer labels [4]–[7].

(ii) Teacher–student (latent-target) objectives predict contextualized *latent* targets rather than raw inputs. data2vec provides a modality-agnostic recipe that predicts latent representations of the full input from a masked view, while Point2Vec adapts this idea to point clouds and addresses positional-leakage pitfalls unique to 3D, reporting strong transfer on ModelNet40/ScanObjectNN [1], [3].

(iii) Joint-embedding predictive architectures (JEPA) predict target *representations* for spatially contiguous blocks from a single context block, learning semantic features without heavy view augmentations [8], [9]. Point–JEPA brings this design to point clouds with a simple sequencer that orders patch centers so contiguous context/target blocks can be sampled despite unordered points, avoiding input-space reconstruction or extra modalities while remaining competitive on standard 3D benchmarks [2], [9].

Despite this progress, the impact of predictive, context-aware pretraining on *grasp–joint prediction* remains under-explored [10]–[12]. In this study, we adopt a Point–JEPA backbone for object embeddings and pair it with an inference-aware, multi-hypothesis loss for joint angles, comparing against training from scratch and analyzing label-efficiency gains under fixed object-level splits.



(a) (b) (c)

Fig. 1: Example grasp predictions from our method. Given an object point cloud and wrist pose, the model outputs a stable joint configuration for the hand.

II. Technical Outline

We convert meshes to point clouds and tokenize them into local patches, yielding input-space patch embeddings suitable for transformer processing [13].

We utilize a pretrained JEPA backbone trained on ShapeNet point clouds via Point–JEPA [2]. A block sampler selects *spatially contiguous* target windows, and a simple sequencer orders patch centers so these windows are contiguous in the 1-D token sequence, following the JEPA masking/targeting scheme [8]. The *context encoder* sees the input with targets masked; the *target encoder* processes the full input to produce latent targets. A lightweight predictor maps context features to the target space, and training minimizes a regression loss between predicted and stop-gradient target features on the masked blocks. The target encoder is an EMA copy of the context encoder, so no negatives or input-space reconstruction are required. After pretraining, we discard predictor and target encoders and retain only the context encoder to produce contextualized patch features; these are pooled with attention into a global object embedding (Fig. 2). For visual clarity, Fig. 2 shows single masked patches, whereas pretraining masks *blocks*.

For grasp prediction, the pooled object embedding is concatenated with the 7D wrist pose and passed to a lightweight head that emits K joint-angle hypotheses along with logits (and optionally scores). Training uses a winner-takes-all / min-over-K objective to preserve multi-modality (see Sec. III-D); at inference, we select the top–logit hypothesis [14].

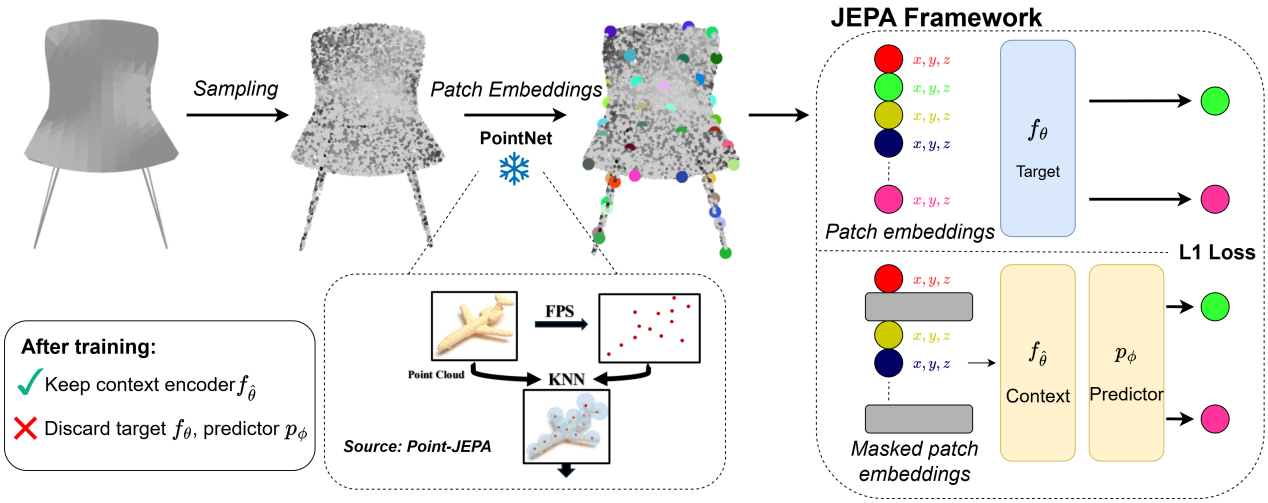


Fig. 2: **Point–JEPA pretraining [2].** Object mesh is converted to a point cloud, tokenized into patch embeddings, and passed to a self-distilled JEPA to obtain contextualized patch features.

III. Methodology

A. Dataset & split protocol

Data and filtering. We use the DLR–Hand II dataset built atop ShapeNet object categories, comprising approximately 10 to 500 unique shapes per category. Each shape provides around 240 grasp samples, each consisting of a 7D wrist pose, a 12D joint configuration, and a grasp quality score. After discarding all grasps with score < 1.5, our dataset contains on the order of 1,500 meshes and 370,000 total grasps.

Object-level, category-stratified splits. To prevent overfitting and pose leakage from multiple grasps of the same mesh, we split at the object level. Sample-level splits proved overly easy and invalidated the value of SSL-based encoding, and so were not used.

Fixed evaluation (shared across label budgets). A single evaluation suite is fixed once and reused across all budgets: (i) **val**: 10% of objects, (ii) **test_object**: another 10% of objects, and (iii) **test_category**: all objects from two held–out synsets.

The *test_category* evaluation (held-out synsets) was pre-specified to assess cross-category transfer. Due to earlier than anticipated GPU deallocation, we could not run it. We keep the protocol and release split JSONs and evaluation scripts; all quantitative claims in this report pertain to *val*.

Low–data training subsets. For $p \in \{1, 10, 25, 100\}\%$, we sample the train set from the remaining objects with proportional *category stratification* (at least one object per synset).

Two split packs and reproducibility. We construct two independent train split packs (A/B) with different manifest seeds; the evaluation fixtures (val, test_object, test_category) are identical across packs. Unless otherwise noted, table entries report the mean $\pm$ standard deviation computed *across both split packs and, where available, across initialization seeds* (seeds 0 and 1). The exact split JSONs and the split-generation script are included with the project materials.

B. Preprocessing & object representation

Meshes are converted to point clouds and tokenized into local groups to create patch embeddings. We use 1024 points per object and a tokenizer with 64 groups (group size 32, radius 0.05). These patch embeddings feed the self–supervised encoder (Fig. 2); a global object embedding is then produced by *attention pooling*, which we found to be more stable and accurate than mean/max pooling.

C. SSL backbone (Point–JEPA)

We use the Point–JEPA *context encoder* as the backbone (predictor/target discarded after pretraining). Tokenized point-cloud patches with positional encodings are encoded and attention-pooled into a single object embedding. We fine-tune end-to-end with a *two-tier learning rate* (smaller on the pretrained backbone, larger on the new grasp head); concrete settings and schedules appear in Sec. IV.

D. Joint estimation head & loss

We predict K candidate joint configurations $\{\hat{\mathbf{j}}_k\}_{k=1}^{K}$ for each object embedding and wrist pose, together with a logit vector $\boldsymbol{\ell} \in \mathbb{R}^K$ used to rank hypotheses at test time. Grasping is inherently *multi-modal*: several distinct joint settings can succeed for the same pose, and single-output MSE collapses these modes to their mean, often yielding an interpolated (and infeasible) grasp (see Fig. 4). To preserve distinct modes, we use a winner-takes-all (WTA) / min-over-K objective, a standard approach for ambiguous prediction tasks.

$$k^* = \arg\min_k \|\hat{\mathbf{j}}_k - \mathbf{j}\|^2 \tag{1}$$

$$L = \|\hat{\mathbf{j}}_{k^*} - \mathbf{j}\|^2 + \alpha\, \mathrm{CE}(\boldsymbol{\ell},\, k^*) \tag{2}$$

The first term penalizes only the closest hypothesis, avoiding mean-collapse; the cross-entropy term teaches the logits to act as a proxy selector for the winning mode. We then use the

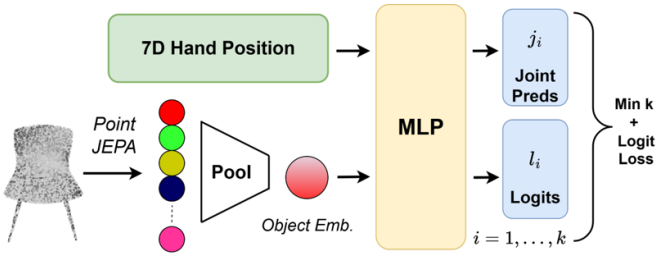


Fig. 3: **Grasp head architecture.** Contextualized object features from Point–JEPA are pooled into a global embedding and concatenated with the 7D wrist pose. An MLP predicts K joint configurations and corresponding logits.

same selector at inference (pick the top–logit hypothesis), so evaluation does not rely on an oracle "best-of-K". This mirrors multi-hypothesis practice in trajectory forecasting, where sets of candidates are scored and one is chosen at test time.

In preliminary trials we also considered Mixture Density Networks (MDNs) to model multi-modality probabilistically, but found them brittle in this high-dimensional joint space (training instability and covariance issues) and less aligned with our top-1 selection at inference; the WTA formulation remained simpler and more stable.

All experiments use $K{=}5$ to balance coverage and compute (brief pilots with $K \in \{3, 4\}$ were inconclusive; a full sweep is left to future work). Beyond the primary top–logit RMSE, we report two diagnostics that probe the multi-hypothesis mechanism without relying on an oracle at evaluation time: (i) a *selection gap* (top–logit RMSE minus best-of-K RMSE) to assess selector fidelity, and (ii) *Coverage*@15°, the fraction of samples for which at least one hypothesis lies within 15° of the ground truth.

E. Evaluation protocol & metrics

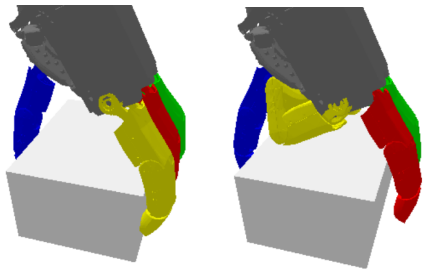
At inference, we choose the head with the largest logit ("top–logit"). The main metric is **validation top–logit RMSE** on joint angles. We additionally compute *coverage*@$\{10°, 15°, 20°\}$ (a grasp counts as covered if any head is within the angular threshold) and a diversity proxy (mean pairwise dispersion across heads) to quantify multimodality (Appendix).

IV. Results

Unless stated otherwise, all metrics are reported on the fixed *val* split with object-level stratification, averaged over split packs and seeds where available. The pre-registered *test_category* evaluation could not be executed due to a provider-side compute outage (see Sec. III); we therefore refrain from cross-category claims.

A. Main outcomes

Table I summarizes top–logit RMSE (radians; lower is better). Three patterns emerge:



(a) (b)

Fig. 4: **Min-over-K illustrated.** For a fixed object and wrist pose, two hypotheses $(\hat{\mathbf{j}}_1, \hat{\mathbf{j}}_2)$ yield stable, collision-free grasps (a–b), while their mean is infeasible. Training (Eq. 2) regresses only the hypothesis nearest to $\mathbf{j}$ by L2 in joint-angle space and trains logits via cross-entropy; inference uses the top–logit.

TABLE I: **Validation top–logit RMSE** (radians; lower is better). Mean ± SD across seeds and split packs.

Train split	Scratch	JEPA	Δ (rel.)
1%	0.363 ± 0.002	**0.335 ± 0.003**	+7.7%
10%	0.335 ± 0.003	**0.303 ± 0.009**	+9.6%
25%	0.332 ± 0.002	**0.246 ± 0.012**	+25.9%
100%	0.235 ± 0.002	**0.234 ± 0.008**	+0.4%

- **Low labels (1–10%):** JEPA pretraining yields consistent gains (≈8–10% relative error reduction), indicating improved label efficiency in scarce-data regimes.
- **Moderate labels (25%):** The largest improvement appears at 25% (≈26% relative), a "sweet spot" where pretrained context helps most before fully supervised training saturates.
- **Full labels (100%):** Performance is effectively at parity; differences are within seed variance (JEPA ≈ scratch), suggesting that with enough supervision, training from random initialization can catch up.

B. Inference-aware selection and coverage

To assess whether the logits reliably choose the correct hypothesis, we define the *selection gap*

$$\Delta_{\text{sel}} = \text{RMSE}_{\text{top–logit}} - \text{RMSE}_{\text{best-of-}K},$$

with smaller values indicating better alignment between the learned selector and the oracle winner. At 1% and 10% budgets, JEPA reduces the gap relative to scratch (e.g., 1%: 0.142 vs 0.165 rad; 10%: 0.157 vs 0.176 rad), demonstrating stronger inference-aware selection (see Appendix, Fig. 9).

As a complementary multi-hypothesis metric, we report Coverage@15°: a grasp is covered if *any* head lies within 15° of the ground-truth joint angles. Coverage increases with label budget and is consistently higher with JEPA in the low-label settings (e.g., at 10%: 0.955 vs 0.938; at 1%: 0.866 vs 0.861), indicating that pretraining improves the probability that at least one hypothesis is close to the target (Appendix, Fig. 10).

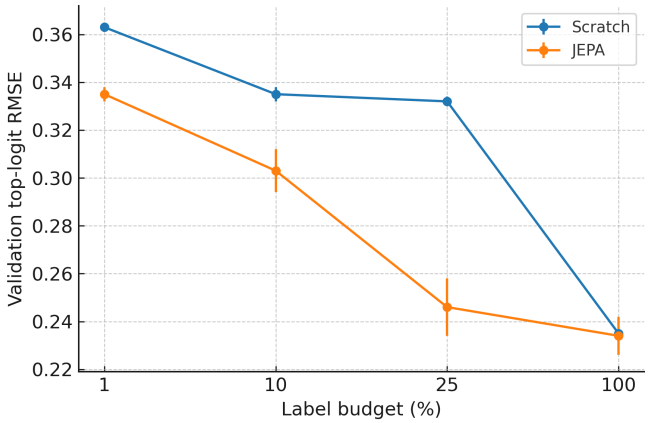


Fig. 5: **Label efficiency.** Validation top–logit RMSE versus label budget (1%, 10%, 25%, 100%). JEPA pretraining improves performance in low-label regimes, with the largest gap at 25%; at 100% both methods converge. Points show mean over seeds/split packs; error bars denote ± 1 SD.

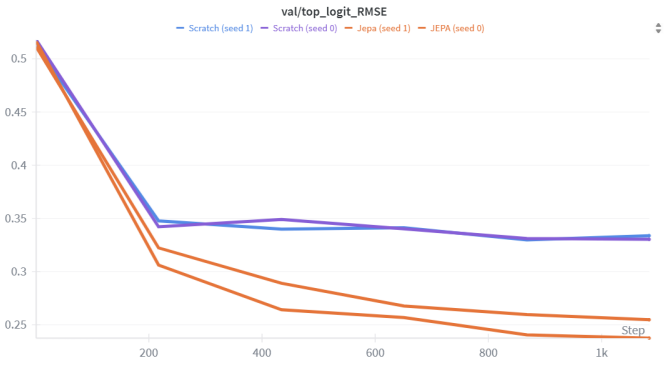


Fig. 6: **25% label budget: convergence and final error.** Validation *top–logit RMSE* (lower is better) versus training step for two seeds (0, 1). Across both seeds, JEPA pretraining converges faster and reaches a lower final error than training from scratch, highlighting its benefit in the moderate low-label regime.

C. Sensitivity to learning rates (10% grid)

We ran a **3×2 grid** over $(\mathrm{LR}_{\mathrm{backbone}}, \mathrm{LR}_{\mathrm{head}})$ on the **10%** split (time constraints precluded repeating per budget). Results (Appendix, Fig. 7) plot validation *top–logit RMSE* versus training step and show a broad plateau at convergence: final differences across the grid are small ($\lesssim 0.01$ rad). Larger backbone LRs destabilize early fine-tuning, while smaller head LRs slow convergence. We therefore fix $(\mathrm{LR}_{\mathrm{backbone}}, \mathrm{LR}_{\mathrm{head}}) = (1\times10^{-5}, 1\times10^{-3})$ for all budgets.

Figure 5 shows that the JEPA curve sits uniformly below scratch at 1–25% and the gap closes at 100%, corroborating Table I.

D. Inference-aware multimodality

Evaluation uses top–logit selection (no oracle "best-of-K"), matching the training objective (Eq. 2). Qualitative examples in Fig. 4 show distinct, stable hypotheses for the same object/pose; their mean would be infeasible, underscoring the need for a multi-hypothesis head.

Takeaway. Point–JEPA materially improves sample efficiency for grasp–joint prediction, with the largest gains in moderate low-label regimes; under full supervision, scratch training attains parity.

V. Conclusion

We investigated whether 3D self-supervised pretraining improves grasp-joint prediction under limited labels. Integrating a Point–JEPA backbone with a winner-takes-all multi-hypothesis head yields consistent gains in low-label regimes on DLR–Hand II, with the largest relative improvement at 25% data. At 100% labels, training from scratch attains parity. Diagnostics indicate that the selector is effective (smaller selection gap) and that Coverage@15° improves with pretraining, supporting the inference-aware design.

VI. Future work

We outline several directions to strengthen generalization and deployment relevance:

- **Cross-domain evaluation.** Run the pre-registered *test_category* suite and real-robot tests to assess transfer beyond object-level splits.
- **Heads and selection.** Sweep K and study diversity-encouraging variants, while monitoring selection gap and top–logit coverage as primary inference-aware metrics.
- **Geometry-aware patching.** Replace FPS + fixed-kNN with patches built from simple geometric cues, e.g., multi-scale radius neighborhoods, curvature-guided grouping, and small overlaps, so each patch better respects local shape.
- **Geometry-focused tokenizer.** Swap the PointNet patch encoder for one that uses explicit local geometry. The JEPA objective stays the same; only the patch encoder changes.
- **Objective design.** Jointly model grasp success/score and angles (multi-task), or explore structured probabilistic heads that remain stable in high-D while preserving top-1 inference.
- **Compute-efficient fine-tuning.** Evaluate lightweight adaptation (e.g., LoRA) and backbone LR schedules across label budgets.

References

[1] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "data2vec: A general framework for self-supervised learning in speech, vision and language," *arXiv preprint arXiv:2202.03555*, 2022.

[2] A. Saito, P. Kudeshia, and J. Poovvancheri, "Point-jepa: A joint embedding predictive architecture for self-supervised learning on point cloud," 2025.

[3] K. Abou Zeid, J. Schult, A. Hermans, and B. Leibe, "Point2vec for self-supervised representation learning on point clouds," 2023.

[4] G. Hess, J. Jaxing, E. Svensson, D. Hagerman, C. Petersson, and L. Svensson, "Masked autoencoder for self-supervised pre-training on lidar point clouds," in *2023 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*. IEEE, 2023, pp. 350–359.

[5] Z. Xie *et al.*, "Masked autoencoders for point cloud self-supervised learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[6] X. Yu *et al.*, "Point-bert: Pre-training 3d point cloud transformers with masked point modeling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[7] Y. Liu *et al.*, "Maskpoint: Masked autoencoders for point cloud self-supervised learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[8] M. Assran, Q. Duval, I. Misra, P. Bojanowski, P. Vincent, M. Rabbat, Y. LeCun, and N. Ballas, "Self-supervised learning from images with a joint-embedding predictive architecture," 2023.

[9] N. Hu, H. Cheng, Y. Xie, S. Li, and J. Zhu, "3d-jepa: A joint embedding predictive architecture for 3d self-supervised representation learning," 2024.

[10] J. Mahler *et al.*, "Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics," in *Robotics: Science and Systems (RSS)*, 2017.

[11] M. Gualtieri, A. Ten Pas, K. Saenko, and R. Platt, "High precision grasp pose detection in dense clutter," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2016.

[12] M. Sundermeyer, A. Mousavian, and D. Fox, "Contact-graspnet: Efficient 6-dof grasp generation in cluttered scenes," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2021.

[13] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[14] T. Phan-Minh *et al.*, "Covernet: Multimodal behavior prediction using trajectory sets," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2020.

APPENDIX A
IMAGE ARCHIVE

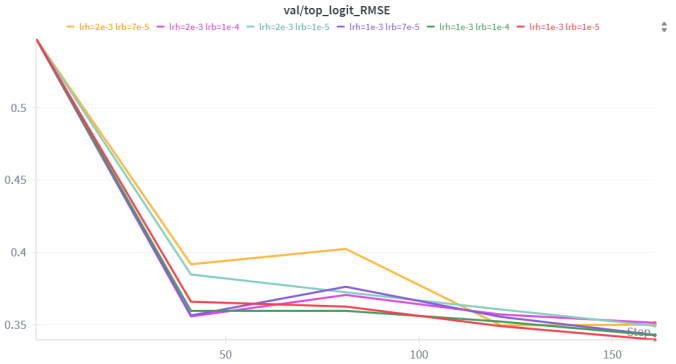


Fig. 7: **LR sensitivity on 10% data (3×2 grid).** Validation *top–logit RMSE* versus training step for $(\mathrm{LR}_{\mathrm{head}}, \mathrm{LR}_{\mathrm{backbone}}) \in \{1\times10^{-3}, 2\times10^{-3}\} \times \{1\times10^{-5}, 1\times10^{-4}, 7\times10^{-5}\}$. Curves are very close at convergence (differences $\lesssim 0.01$ rad), indicating a broad plateau. We select $(\mathrm{LR}_{\mathrm{backbone}}, \mathrm{LR}_{\mathrm{head}}) = (1\times10^{-5}, 1\times10^{-3})$ for a slightly lower final RMSE and stable early training; a larger backbone LR (7×10^{-5}) shows transient degradation early. The legend specifies $(\mathrm{LR}_{\mathrm{head}}, \mathrm{LR}_{\mathrm{backbone}})$ pairs.

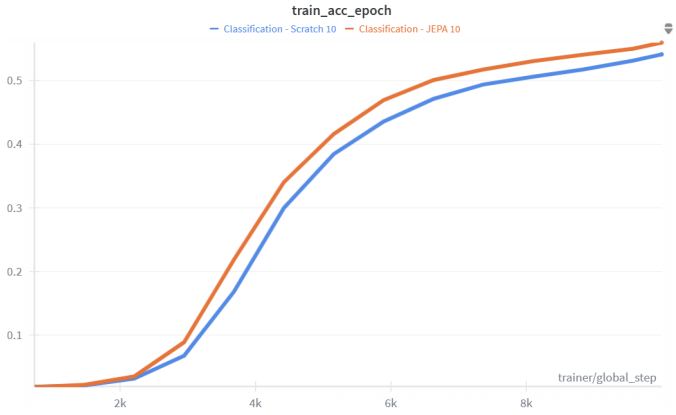


Fig. 8: **Classification sanity check.** Object classification on DLR–Hand II meshes using the same backbone. JEPA-pretrained features achieve marginally higher accuracy than training from scratch, reproducing the trend reported for Point–JEPA and validating our pipeline and JEPA checkpoint.

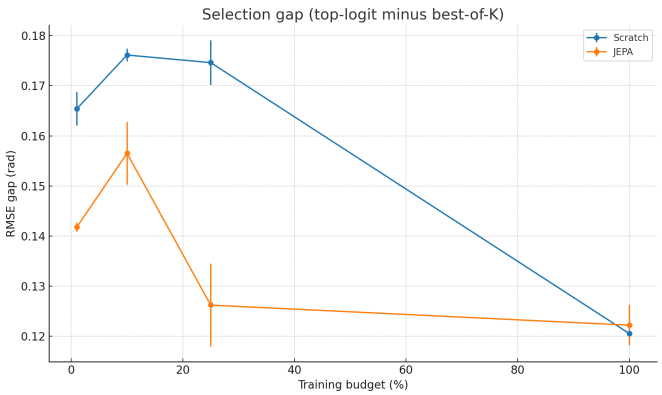


Fig. 9: **Selection gap.** $\Delta_{\mathrm{sel}} = \mathrm{RMSE}_{\mathrm{top-logit}} - \mathrm{RMSE}_{\mathrm{best\text{-}of\text{-}}K}$ versus label budget. Lower is better. JEPA shows smaller gaps at 1–10%, indicating more reliable mode selection. Points: means; error bars: ±1 SD.

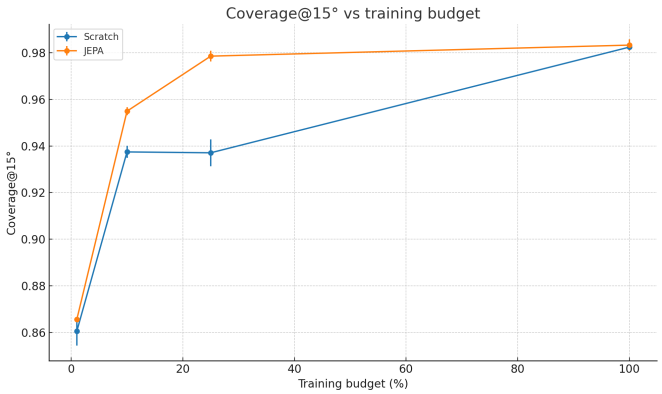


Fig. 10: **Coverage@15°.** Probability that at least one head lies within 15° of the ground truth, by label budget. Coverage increases with supervision and is higher with JEPA at 1–10%. Points: means; error bars: ±1 SD.