Contrastive Action Grounding for Object-Centric World Models

Thomas Deng, Zubin Carvalho
tdeng23@stanford.edu, zubinc@stanford.edu

***Abstract*—Object-centric world models promise structured, entity-level prediction, but do they actually use the actions they are given? We study this in a robotic manipulation setting with a temporal object-centric world model built on the Joint-Embedding Predictive Architecture (JEPA), operating over slots from a frozen VideoSAUR encoder. We uncover a consistent failure mode: standard objectives (Hungarian MSE, future action conditioning, and inverse dynamics) all learn to ignore actions, predicting nearly the same future whether given the true actions or none at all. We quantify this with the GT−Zero gap, the change in prediction accuracy between ground-truth and zeroed actions, which stays near zero for every standard objective throughout training. Our fix is a contrastive action loss that forces futures under different action sequences to diverge, trained with a higher learning rate on the action encoder. It is the only variant to develop real action sensitivity, reaching a +0.299 GT−Zero gap at rollout horizon $k = 10$, and this grounding does not cost prediction quality: the same model is the single best performer on a semantic VQA benchmark derived from LIBERO-100, beating every baseline at every horizon and by the widest margin in the hardest setting (50% predicted). The decisive factor is pressure from the training objective, not architectural access to action information.**

I. Introduction

Predicting how a visual scene evolves is fundamental to robotics and video understanding. Recent world models have demonstrated strong planning capabilities by learning latent dynamics from visual observations [5]. However, patch-level and pixel-space models treat all image regions equally, lacking explicit representations of the discrete, manipulable entities that govern scene semantics, causing them to fail precisely in the cases that matter most: object interactions, collisions, and occlusions.

Object-centric representations offer a principled remedy. Inspired by the long-standing idea that scenes are better represented as collections of objects than as raw pixels, models such as Slot Attention [9] decompose a scene into a small set of latent vectors corresponding to distinct entities, enabling downstream models to reason over individual objects rather than image patches. Prior work has extended this to video via temporal slot attention [6, 14], built transformer-based dynamics predictors over slots [13], and applied object-centric world models to robotic manipulation [10]. These approaches have been integrated with predictive architectures [11], and parallel work has explored predicting in the feature space of frozen visual backbones [16, 3]. A critical question, however, remains open: even when an object-centric world model is *given* robot action information, does it actually use it to shape predictions, or does it learn to ignore it and rely on scene inertia instead?

We investigate this failure mode in the context of robotic manipulation, where future states are fundamentally determined by object interactions induced by actions. Using LIBERO-100 [8], a diverse tabletop manipulation benchmark with RGB observations and action trajectories, we build an object-centric world model based on JEPA [1] operating over slots from a frozen VideoSAUR encoder. We find that standard training objectives, including Hungarian MSE, future action conditioning, and inverse dynamics losses, all produce nearly action-invariant predictions despite conditioning on robot actions. We introduce a contrastive action objective that explicitly forces futures generated under different action sequences to diverge, and evaluate both action sensitivity and semantic fidelity using a frozen ALOE [4] VQA head.

Our experiments confirm that action-grounding requires objective-level pressure, not just architectural access to action information, and that supplying this pressure yields the best model on *both* axes we measure. Specifically: (1) Hungarian MSE, future action conditioning, and inverse dynamics all fail to produce action-sensitive predictions, with GT−Zero gaps near zero throughout training; (2) a contrastive action loss paired with a high action-encoder learning rate achieves a GT−Zero gap of +0.299, far exceeding all baselines; and (3) this action sensitivity translates directly into downstream performance: the contrastive model is the single best-performing variant at every horizon on the ALOE VQA benchmark, with the largest gains at the long-horizon settings where causal structure matters most. Notably, a model that explicitly receives future action embeddings in its transformer (v6d) still fails, demonstrating that *what the training objective demands*, not what information is architecturally available, determines whether actions are used.

Our contributions are: (i) identifying the action-invariance failure mode in object-centric JEPA world models; (ii) a contrastive action objective that directly addresses it; and (iii) showing that contrastive training yields both the strongest action sensitivity and the best downstream task performance, surpassing Hungarian MSE baselines, future action conditioning, and inverse dynamics alternatives at every prediction horizon.

II. Methods

A. Overview

Our model follows a three-stage pipeline adapted from C-JEPA [11]. A frozen VideoSAUR encoder [14] produces $N = 7$ object slots of dimension $D = 128$ per frame. A transformer predictor takes a history window of $T_h = 5$ frames and predicts slots for the next $T_p = 3$ frames at a frameskip of 4 (about 8 Hz). At each timestep, a 7-dimensional end-effector action is encoded by a two-layer MLP and appended as an additional action slot, giving a per-frame token set of shape $(N{+}1, D) = (8, 128)$.

During training, $k = 3$ of 7 object slots are randomly masked at all future timestep positions; the predictor must infer them from visible slots, the action slot, and the full history. The action slot is never masked. Prediction targets are slot representations from an EMA copy of the encoder (decay 0.996), following the EMA target encoder convention of V-JEPA [3] and V-JEPA 2 [2]. The transformer uses depth 6, 16 heads, head dimension 64, and FFN hidden dimension 2048 with 0.1 dropout.

B. Training Objective

Because slots are unordered, we use Hungarian matching [7] to align predicted slots $\hat{Z} \in \mathbb{R}^{N \times D}$ with EMA targets $Z^* \in \mathbb{R}^{N \times D}$:

$$\mathcal{L}_{\text{pred}} = \frac{1}{N \cdot T_p} \sum_{t=1}^{T_p} \sum_{i=1}^{N} \|\hat{z}_{t,i} - z^*_{t,\sigma^*(i)}\|^2$$

where σ^* is the minimum-cost bijection. An equal-weight MSE on masked history slots is added following C-JEPA.

C. Contrastive Action Loss

Standard Hungarian MSE training produced nearly action-invariant predictions, since the action encoder received negligible gradient signal. Our approach draws inspiration from temporal action-contrastive methods in visual RL [15], adapted here to the object-centric slot prediction setting. We address this with two coupled changes.

For each batch we construct negative pairs by shuffling action sequences across elements. The predictor produces futures conditioned on the real sequence a and a shuffled sequence a' from a different trajectory. A margin-based contrastive loss requires the mean-slot representations to diverge:

$$\mathcal{L}_{\text{contrast}} = \max\big(0,\ m - \|\bar{z}(a) - \bar{z}(a')\|^2\big)$$

where $\bar{z}$ is the mean over object slots at the first predicted timestep and $m = 0.5$. The total loss is $\mathcal{L} = \mathcal{L}_{\text{pred}} + \lambda_{\text{contrast}} \cdot \mathcal{L}_{\text{contrast}}$ with $\lambda_{\text{contrast}} = 0.1$.

We assign the action encoder a learning rate five times higher than the base (base learning rate 2×10^{-4}, action-encoder learning rate 1×10^{-3}). Without this, the contrastive gradient at the action encoder is too small relative to the predictor to produce meaningful updates even when the loss is nonzero.

TABLE I
Action sensitivity at the latest checkpoint at horizon $k = 10$. GT: cosine similarity under ground-truth actions. Δ: the GT$-$Zero gap (GT minus the same similarity under zero actions). Pers.: persistence baseline.

Method	Step	GT	Δ	Pers.
OC-JEPA	45k	0.7695	-0.0026	0.9384
v6b (H-MSE)	100k	0.7746	-0.0011	0.9402
v6d (FAC)	100k	0.7710	-0.0002	0.9393
v6e (Contrastive)	40k	0.7663	**+0.2989**	0.9397
v6f (InvDyn)	30k	0.7720	$+0.0089$	0.9483

III. Experiments

A. Setup

We train and evaluate on LIBERO-100 [8], comprising 100 tabletop manipulation tasks with about 50 expert demonstrations each (about 5,000 total). We use the agentview RGB camera, subsample at frameskip 4, and reserve the final 10% of demonstrations per task as a held-out test set. All metrics are averaged over 100 randomly sampled test trajectories.

Action sensitivity is measured by the GT$-$Zero gap $\Delta = \text{CosSim}_{\text{GT}} - \text{CosSim}_{\text{Zero}}$. Here $\text{CosSim}_{\text{GT}}$ is the cosine similarity between the predicted future and the true future when the model is conditioned on the ground-truth (GT) actions, and $\text{CosSim}_{\text{Zero}}$ is the same quantity when the actions are replaced by zeros. The gap is therefore how much better the model predicts with the real actions than with no action information. A model that ignores actions produces the same future in both cases, giving a gap near zero.

Semantic fidelity is measured via an ALOE [4] VQA benchmark: for each observation fraction f in $\{0.9, 0.8, 0.7, 0.6, 0.5\}$, we use the first f fraction as ground-truth context, roll the predictor forward for the remaining $1-f$ fraction, and feed the resulting slot trajectory to a frozen ALOE VQA head. Lower f corresponds to harder settings.

We compare five variants that share the same VideoSAUR encoder [14] and transformer predictor backbone, differing only in training objective or masking strategy. **OC-JEPA** is a C-JEPA-style [11] model trained with *zero* slot masking, meaning all object slots are visible at every future timestep, so the model receives no pressure to infer occluded objects from context or actions. **v6b (H-MSE)** is the full C-JEPA recipe with random slot masking but only the base Hungarian MSE prediction loss and no auxiliary action objective; this is the primary baseline isolating the effect of the contrastive loss. **v6d (FAC)** augments v6b by additionally providing *future* action embeddings as extra transformer tokens at every predicted timestep, testing whether architectural access to future actions alone is sufficient. **v6f (InvDyn)** augments v6b with an inverse dynamics auxiliary loss [15]: an MLP must recover the action from the predicted slot transition, providing indirect action-grounding pressure without explicit counterfactual comparison. **v6e (Contrastive)** is our proposed method, described in Section II-C.

B. Action Sensitivity

Table I surfaces a subtlety that is central to interpreting these results: the persistence baseline achieves cosine similarity of about 0.94, substantially higher than all learned models (about 0.77), yet it is by construction incapable of action-grounded prediction. This apparent paradox arises because cosine similarity in slot space conflates two very different phenomena. Many LIBERO scenes evolve slowly, as tabletops, backgrounds, and stationary objects barely move between consecutive frames, so simply copying the previous latent state yields a representation highly aligned with the ground-truth slots in the L_2 sense. Learned predictors, by contrast, may *intentionally* diverge from the previous state when they predict object motion caused by the robot's action, reducing raw cosine similarity even while producing a causally more faithful trajectory.

This is precisely why the GT−Zero gap Δ is the more informative metric. A model that ignores actions entirely, whether trivially (as in persistence) or through learned scene inertia (as in the baseline predictors), produces the same output regardless of what action is provided, giving a gap near zero. A model that genuinely uses action information to shape its predictions will produce representations that are more aligned with the ground truth under correct actions than under zero actions, giving a positive gap. All learned baselines, OC-JEPA, v6b, v6d, and v6f, exhibit GT−Zero gaps within 0.003 of zero, indicating their predictions are essentially action-agnostic despite receiving action inputs. The contrastive model (v6e) is the sole exception, reaching a gap of +0.299 at horizon $k = 10$.

A further diagnostic comes from comparing each model's GT cosine similarity against persistence: all learned models score roughly 17 points *below* persistence (0.77 versus 0.94). For the action-agnostic baselines, this gap is purely a cost, not a benefit, since they predict worse than doing nothing while also ignoring actions. For v6e, the lower cosine similarity may partially reflect genuine prediction of action-induced motion; the ALOE horizon-sweep results in Section III-C confirm that this translates to better task-discriminative content, and in fact to the best task-discriminative content of any model we test.

Crucially, v6d (Future Action Conditioning) explicitly injects future action embeddings into the transformer at every step, yet its gap remains near zero. This isolates the failure: the model has architectural *access* to action information but no *objective-level incentive* to use it. The contrastive loss provides that incentive directly.

Figure 1 shows the GT−Zero gap throughout training. All baselines remain within 0.01 of zero, while v6e diverges monotonically from the earliest checkpoint, confirming sustained gradient pressure from the contrastive objective.

C. Semantic Fidelity: Horizon Sweep

Figure 2 shows ALOE VQA accuracy across prediction horizons. H-MSE (v6b) degrades most severely, collapsing to 19.8% at 50% predicted, near chance (25%). OC-JEPA recovers to 34.0%, showing that the temporal architecture contributes beyond the loss alone. Inverse Dynamics performs comparably (32.8%), suggesting similar inductive bias without additive benefit. The Contrastive model achieves 39.0% at 50% predicted and 96.6% at 10% predicted, the highest accuracy at every horizon. Its advantage over H-MSE widens from 5 points at 10% predicted to 19 points at 50% predicted, the expected behavior of a model that has internalized action causality. The takeaway is that the only model that genuinely conditions on actions is also the most accurate at recovering task-relevant content: action grounding and downstream fidelity reinforce one another rather than trading off, and the contrastive objective is therefore the best choice on both metrics simultaneously.

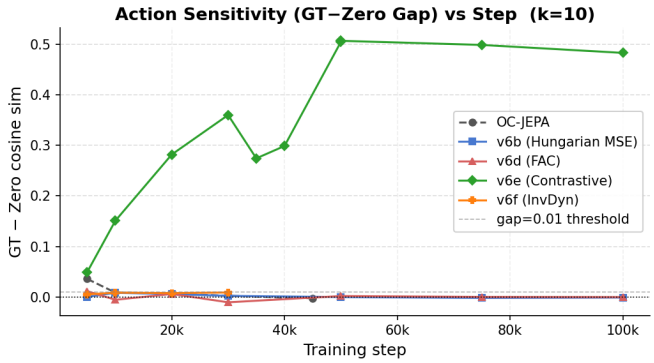

Fig. 1. GT−Zero gap at horizon $k = 10$ throughout training. All baselines remain near zero; the Contrastive Action model diverges monotonically, demonstrating strong action dependence.

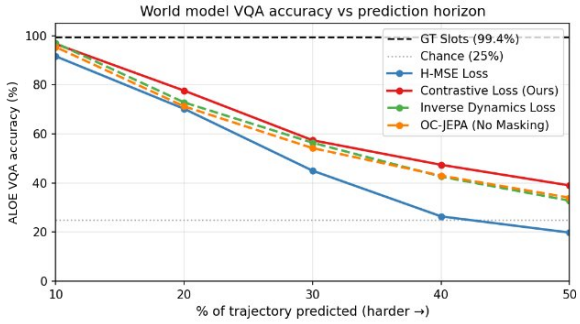

Fig. 2. ALOE VQA accuracy versus prediction horizon. Larger x-values indicate harder settings. Ground-truth slots achieve 99.4%; random guessing yields 25%. The Contrastive Action Loss consistently outperforms all baselines, with the largest margin at long horizons.

D. Qualitative Analysis

Across all methods, large static structures (tables, backgrounds) are predicted reliably. However, small manipulable objects (mugs, books, tools) remain difficult to track and frequently collapse into background slots within a few rollout steps (Figure 3). This failure persists even in v6e, indicating that improved action sensitivity alone does not guarantee accu-

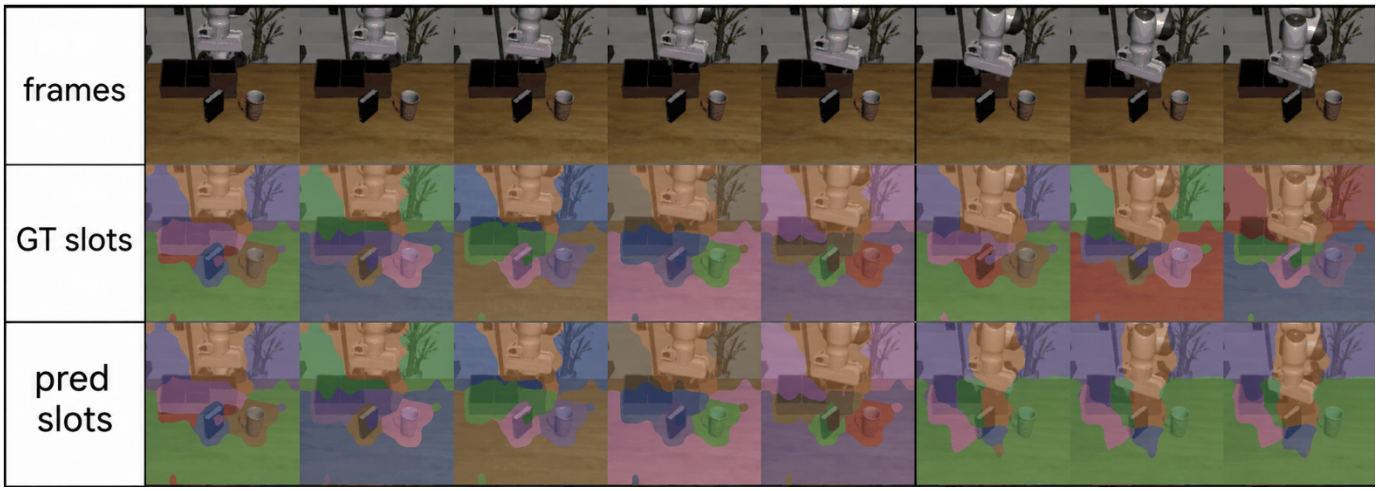

Fig. 3. Qualitative slot prediction rollout on a held-out LIBERO trajectory. Rows show RGB observations, ground-truth VideoSAUR slots, predicted slots, and the persistence baseline. The dark line marks the boundary between observed and predicted frames. Large scene structures remain coherent, while small manipulable objects (e.g., the mug) gradually disappear or merge into surrounding slots, a failure mode shared across all models.

rate object-centric prediction and pointing to open challenges in tracking small objects under occlusion.

IV. Conclusion

We presented an object-centric JEPA world model for robotic manipulation and identified a key failure mode: standard training objectives produce nearly action-invariant predictions despite explicit action conditioning. Our contrastive action loss, paired with a high action-encoder learning rate, is the only method that achieves substantial action sensitivity (a +0.299 GT$-$Zero gap at $k = 10$), and it is simultaneously the best-performing model on the downstream ALOE VQA benchmark, beating every baseline at every prediction horizon. Action grounding and downstream task performance therefore go hand in hand rather than trading off. These results demonstrate that learning action-grounded dynamics requires explicit objective-level pressure, not merely providing action information to the model. Future work should address the persistent challenge of tracking small manipulable objects over long horizons.

References

[1] 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. *arXiv preprint arXiv:2301.08243*, 2023. URL https://arxiv.org/abs/2301.08243.

[2] Mahmoud Assran et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025. URL https://arxiv.org/abs/2506.09985.

[3] Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*, 2024. URL https://arxiv.org/abs/2404.08471.

[4] David Ding, Felix Hill, Adam Santoro, Malcolm Reynolds, and Matthew Botvinick. Attention over learned object embeddings enables complex visual reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[5] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023. URL https://arxiv.org/abs/2301.04104.

[6] Thomas Kipf, Gamaleldin F. Elsayed, Aravindh Mahendran, Austin Stone, Sara Sabour, Georg Heigold, Rico Jonschkowski, Alexey Dosovitskiy, and Klaus Greff. Conditional object-centric learning from video. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2111.12594.

[7] Harold W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2 (1–2):83–97, 1955.

[8] Bo Liu et al. Libero: Benchmarking knowledge transfer for lifelong robot learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2306.03310.

[9] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In *Advances in Neural Information Processing Systems*, volume 33, pages 11525–11538, 2020.

[10] Malte Mosbach, Jan Niklas Ewertz, Angel Villar-Corrales, and Sven Behnke. SOLD: Slot object-centric latent dynamics models for relational manipulation learning from pixels. *arXiv preprint arXiv:2410.08822*, 2024.

URL https://arxiv.org/abs/2410.08822.

[11] Heejeong Nam, Quentin Le Lidec, Lucas Maes, Yann LeCun, and Randall Balestriero. Causal-jepa: Learning world models through object-level latent interventions. *arXiv preprint arXiv:2602.11389*, 2026. URL https://arxiv.org/abs/2602.11389.

[12] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy 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 (TMLR)*, 2024. arXiv:2304.07193.

[13] Ziyi Wu, Nikita Dvornik, Klaus Greff, Thomas Kipf, and Animesh Garg. SlotFormer: Unsupervised visual dynamics simulation with object-centric models. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.05861.

[14] Andrii Zadaianchuk, Matthaeus Kleindessner, Francesco Locatello, Thomas Brox, and Gabriele Abbati. Object-centric learning for real-world videos by predicting temporal feature similarities. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2306.04829.

[15] Ruijie Zheng, Xiyao Wang, Yanchao Sun, Shuang Ma, Hal Daumé, III, Huazhe Xu, and Furong Huang. TACO: Temporal latent action-driven contrastive loss for visual reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2306.13229.

[16] Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. DINO-WM: World models on pre-trained visual features enable zero-shot planning. *arXiv preprint arXiv:2411.04983*, 2024. URL https://arxiv.org/abs/2411.04983.

APPENDIX

A. Object Tokenization with VideoSAUR

VideoSAUR [14] extends Slot Attention [9] to video building on temporal slot attention ideas from SAVi [6], via a self-supervised objective encouraging temporally consistent slot assignments, using a frozen DINOv2-small [12] backbone as the patch-feature source. Each frame is resized to 196×196 and encoded into $N = 7$ slots of dimension $D = 128$. The encoder runs two iterations of slot attention per frame (three on the first frame). Encoder parameters are locked after pretraining on LIBERO-100; only the produced slot sequences are consumed by the predictor.

B. Transformer Architecture

The predictor is a non-causal transformer over the flattened spatiotemporal token sequence of shape $(T_h{+}T_p) \times (N{+}1)$. We use depth 6, 16 attention heads, head dimension 64, FFN hidden dimension 2048, and 0.1 dropout, with learned time positional embeddings. This architecture is adapted from C-JEPA's Masked Slot AP Predictor [11], modified to exclude only the single action slot from masking.

C. Action Encoding

The raw action at each timestep is a 7-dimensional vector $[\Delta x, \Delta y, \Delta z, \Delta\text{roll}, \Delta\text{pitch}, \Delta\text{yaw}, \text{gripper}]$. At frameskip 4, four consecutive raw actions are concatenated into a 28-dimensional vector. A two-layer MLP with LayerNorm and GELU projects this to $D = 128$ and appends it as an extra slot. Dataset-wide mean and standard deviation are computed over all actions before training; actions are standardized to zero mean and unit variance per dimension.

D. Alternative Loss Variants

Inverse dynamics. Given mean object slots $\bar{z}_t$ and $\bar{z}^{\text{pred}}_{t+1}$, an MLP g predicts the action: $\hat{a} = g(\bar{z}_t,\ \bar{z}^{\text{pred}}_{t+1} - \bar{z}_t)$, with $\mathcal{L}_{\text{inv}} = \|\hat{a} - a\|^2$ and $\lambda_{\text{inv}} = 0.05$.

OC-JEPA baseline. An identical model trained with zero masked slots, providing no pressure to infer masked objects from context.

E. Dataset and Preprocessing

LIBERO-100 comprises 100 tabletop tasks with about 50 expert demonstrations each. Raw demonstrations (100–300 timesteps) are subsampled at frameskip 4; four intervening actions are concatenated per step. Each subsampled demonstration is sliced into overlapping clips of length $T_h{+}T_p = 8$ with stride 1. Frames are resized to 196×196 with bicubic interpolation and normalized with ImageNet statistics. Slots are L2-normalized to unit vectors before being passed to the predictor, concentrating training on predicting slot direction rather than magnitude.