

# Object-Centric Latent Action Learning

Albina Klepach<sup>1</sup>, Alexander Nikulin<sup>1,2</sup>, Ilya Zisman<sup>1</sup>, Denis Tarasov<sup>1</sup>, Alexander Derevyagin<sup>1,3</sup>,  
Andrei Polubarov<sup>1</sup>, Nikita Lyubaykin<sup>1,4</sup>, Igor Kiselev<sup>5</sup>, Vladislav Kurenkov<sup>1,4</sup>

<sup>1</sup>dunno1ab.ai

<sup>2</sup>Moscow State University

<sup>3</sup>Higher School of Economics

<sup>4</sup>Innopolis University

<sup>5</sup>Accenture

albina.klepach@gmail.com

## Abstract

Leveraging vast amounts of unlabeled internet video data for embodied AI is currently bottlenecked by the lack of action labels and the presence of action-correlated visual distractors. Although recent latent action policy optimization (LAPO) has shown promise in inferring proxy action labels from visual observations, its performance degrades significantly when distractors are present. To address this limitation, we propose a novel object-centric latent action learning framework that centers on objects rather than pixels. We leverage self-supervised object-centric pretraining to disentangle the movement of the agent and distracting background dynamics. This allows LAPO to focus on task-relevant interactions, resulting in more robust proxy-action labels, enabling better imitation learning and efficient adaptation of the agent with just a few action-labeled trajectories. We evaluated our method in eight visually complex tasks across the Distracting Control Suite (DCS) and Distracting MetaWorld (DMW). Our results show that object-centric pretraining mitigates the negative effects of distractors by **50%**, as measured by downstream task performance: average return (DCS) and success rate (DMW).

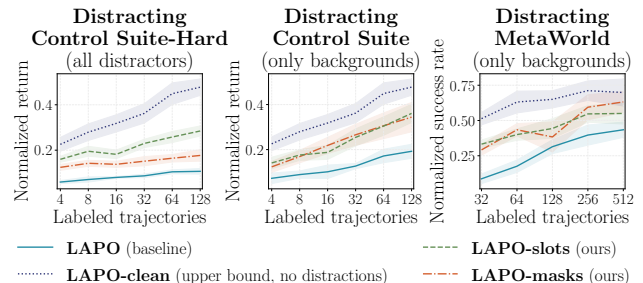
**Code** — <https://github.com/dunno1ab/object-centric-lapo>

**Extended version** — <https://arxiv.org/abs/2502.09680>

## 1 Introduction

In recent years, the scaling of model and data sizes has led to the creation of powerful and general foundation models (Bommasani et al. 2021) that have enabled many breakthroughs in understanding and generation of natural language (Achiam et al. 2023; Brown et al. 2020) and images (Dehghani et al. 2023; Radford et al. 2021). In contrast, the field of embodied AI has generally remained behind in terms of generalization and emergent abilities, being mostly limited by the lack of diverse high-quality data for pre-training (Guruprasad et al. 2024; Lin et al. 2024). The vast amount of video data on the Internet, covering a wide variety of human-related activities, can potentially fulfill the current data needs for training generalist agents (McCarthy et al. 2024). Unfortunately, such videos cannot be used directly due to lack of explicit action labels, which are essential for imitation learning and reinforcement learning algorithms.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.



**Figure 1: Main Results.** Our methods, **LAPO-slots** and **LAPO-masks**, leverage object-centric representations to significantly improve latent action learning under visual distractions. Compared to baseline (**LAPO** (Schmidt and Jiang 2024), trained on raw videos with distraction), our approaches reduce the performance gap toward the clean-data upper bound (**LAPO-clean**, trained on clean videos without distractions) by **50%**. Distractions include camera movements, color variations, and dynamic background videos. Downstream performance is normalized to a behavior cloning agent trained on all ground-truth action labels. Results averaged over three random seeds. See Section 4 for details.

In order to compensate for the lack of action labels, approaches based on Latent Action Models (LAMs) (Schmidt and Jiang 2024; Ye et al. 2024; Cui et al. 2024; Bruce et al. 2024), aim to infer latent actions from the sequence of visual observations. Such latent actions can later be used for behavioral cloning from large unlabeled datasets. However, a major limitation of current LAM approaches, is their susceptibility to action-correlated distractors, such as dynamic backgrounds, that falsely correlate with agent actions, and may lead models to overfit to non-causal patterns (Wang et al. 2024; Misra et al. 2024a; McCarthy et al. 2024; Nikulin et al. 2025). Existing methods, such as Latent Action Pretraining (LAPA) (Ye et al. 2024; Chen et al. 2024), often assume curated, distractor-free datasets or rely on costly annotations, severely limiting their scalability and applicability in realistic settings.

To address this bottleneck, we propose object-centric latent action learning, a framework that leverages object-centric

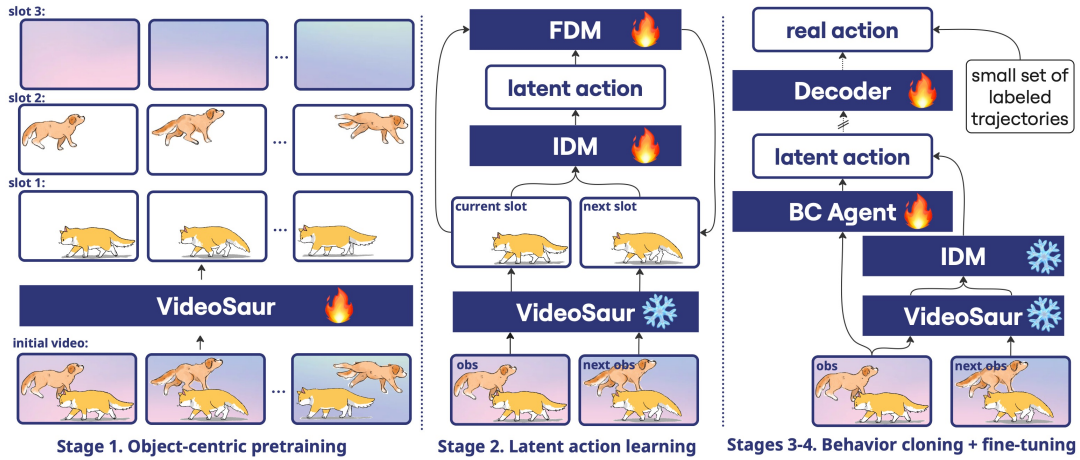


Figure 2: General overview of our method pre-training pipeline. **Object-Centric Pretraining:** We decompose video sequences into interpretable object slots. A linear probe trained on slot representations automatically selects task-relevant slots by identifying those most predictive of actions. **Latent Action Learning:** We train a latent action model based on LAPO, learning inverse and forward dynamics in the slot space. **Behavior Cloning and Fine-tuning:** A behavior cloning agent is trained on the inferred latent actions. The resulting policy is then fine-tuned using a small number of trajectories with ground-truth actions ( $\leq 2.5\%$  of total data), enabling strong downstream performance with minimal supervision.

representations to isolate task-relevant entities from distracting visual noise. Object-centric models decompose scenes into discrete, interpretable object slots (Locatello et al. 2020; Dittadi et al. 2022) using self-supervised mechanisms that group pixels into coherent entities based on feature similarity. These slots encode spatio-temporal properties (Zadaianchuk, Seitzer, and Martius 2023) effectively disentangling complex visual inputs into semantically meaningful units.

Our core hypothesis is that object-centric representations provide the necessary structural priors to separate causal agent-object interactions from non-causal visual correlations. By operating on slot-based object features rather than raw pixels, our method enables LAMs to focus on the dynamics of relevant objects while filtering out distractions such as background motion.

We evaluated our approach in continuous control environments with dynamic distractors, such as Distracting Control Suite (DCS) (Stone et al. 2021) and previously unexplored Distracting Metaworld (DMW) (Yu et al. 2019), showing that self-supervised object-centric pretraining, built on VideoSAUR (Zadaianchuk, Seitzer, and Martius 2023), significantly improves robustness compared to standard LAM baselines. Our results suggest that object decomposition serves as an effective inductive bias for latent action learning, especially under realistic, noisy conditions. In addition, we propose a simple way to automatically choose slots relevant to control and analyze its effectiveness. We summarize our main results on Figure 1.

## 2 Preliminaries

In our experiments, we use environments with partial observability (POMDP) (Åström 1965; Kaelbling, Littman, and Cassandra 1998). A POMDP is defined by the tuple  $(\mathcal{S}, \mathcal{A}, T, R, \Omega, \gamma)$ , where:  $\mathcal{S}$  is the state space,  $\mathcal{A}$  is the ac-

tion space,  $T(s_{t+1} | s_t, a_t)$  defines the transition dynamics,  $R(s_t, a_t, s_{t+1})$  is the reward function,  $\Omega(o_t | s_t)$  is the observation model,  $\gamma \in [0, 1]$  is the discount factor. At each time step  $t$ , the agent receives an observation,  $o_t \sim \Omega(\cdot | s_t)$ , and selects an action  $a_t \sim \pi(\cdot | o_t)$  to maximize its expected discounted return:  $\mathbb{E}[\sum_{k=t}^{\infty} \gamma^{k-t} r_k]$ . The agent transitions to the next state  $s_{t+1} \sim T(\cdot | s_t, a_t)$  and receives a reward  $r_t = R(s_t, a_t, s_{t+1})$ .

We consider the setting in which we first learn a policy  $\pi$  from offline actionless demonstration data, then fine-tune it on real actions, introduced by (Schmidt and Jiang 2024). Suppose we are given a dataset  $\mathcal{D}$  of expert trajectories in the form of sequences of observations and actions:

$$\tau_{oa} = \{(o_0, a_0, o_1), \dots, (o_{|\tau|-1}, a_{|\tau|-1}, o_{|\tau|})\}$$

When such labeled data is available, Behavior Cloning (BC) (Pomerleau 1988) can be used to train a policy  $\pi_{BC}(a_t | o_t)$  through supervised learning. However, when using real-world video data the action labels are missing, and we only observe sequences of the form:

$$\tau_o = \{o_0, o_1, \dots, o_{|\tau|}\}$$

**Latent Action Modeling.** Edwards et al. (2019); Schmidt and Jiang (2024) proposed to convert classic BC into a pseudo-BC task by inferring latent actions  $z_t$  that explain each observed transition  $(o_t, o_{t+1})$ . These inferred actions are then used to train a latent action policy  $\tilde{\pi}(z_t | o_t)$ , bypassing the need for ground-truth action labels. To obtain latent actions, LAM trains an inverse-dynamics model (IDM)  $f_{IDM}(z_t | o_t, o_{t+1})$ , which infers latent actions from state transitions. To ensure that the IDM learns meaningful representations, it is trained jointly with a forward-dynamics model (FDM):  $f_{FDM}(o_{t+1} | o_t, z_t)$ , which predicts the next observation from the current observation and the inferred action. The models are optimized to perform next observation

prediction:

$$\mathcal{L}_{\text{MSE}} = \mathbb{E}_t \left[ \|f_{\text{FDM}}(f_{\text{IDM}}(\mathbf{o}_t, \mathbf{o}_{t+1}), \mathbf{o}_t) - \mathbf{o}_{t+1}\|^2 \right]$$

Without distractors LAM is able to recover ground-truth actions (Schmidt and Jiang 2024; Bruce et al. 2024), however, with distractions it does not work (Nikulin et al. 2025).

**Challenge of distractors.** Real-world videos inherently contain action-correlated distractors: environmental dynamics (e.g. moving backgrounds, camera jitter) that spuriously correlate with agent actions. However, existing learning from observation methods lack mechanisms to disentangle distractors (Efroni et al. 2022; Misra et al. 2024b), leading to overfitting in noisy settings. Recently, Nikulin et al. (2025) empirically demonstrated that LAPO-based methods suffer significant performance degradation when trained on data that contain distractors and suggested reusing available action labels to provide supervision during LAM pre-training. However, it is not always possible to provide such supervision, since action labels do not exist in principle in some domains (e.g. YouTube videos). On the other hand, object-centric decomposition can be applied in a meaningful way to any real-world video data, greatly expanding the applicability of latent action learning.

**Object-Centric Decomposition.** Object-centric learning (OCL) aims to decompose complex visual scenes into structured, interpretable representations by isolating individual entities over time. These entities are typically encoded as a set of  $K$  slot vectors  $\mathcal{S}_t = \{\mathbf{s}_t^{(1)}, \mathbf{s}_t^{(2)}, \dots, \mathbf{s}_t^{(K)}\}$ , representing a distinct entity or region within the scene and capture object-specific properties such as position, shape, color, and motion. A key strength of OCL models lies in their ability to encode the causal structure of the environment. By isolating coherent, temporally consistent entities, they provide a natural defense against distractors in real-world video data. For example, VideoSAUR (Zadaianchuk, Seitzer, and Martius 2023), a self-supervised video decomposition model, has been shown to effectively ignore spurious correlations introduced by moving backgrounds or viewpoint changes, focusing instead on task-relevant objects.

Although object slots provide disentangled and semantically meaningful representations, identifying which slots correspond to task-relevant entities remains challenging and necessitates additional mechanisms to identify and track relevant entities without manual supervision. Also, most OCL frameworks, require the number of slots  $K$  to be specified a priori, which limits flexibility in modeling scenes with varying numbers of objects and can lead to either under-representation or fragmentation when  $K$  is mismatched with scene complexity.

### 3 Method

**Object-Centric Representation Learning.** We employ the VideoSAUR model (Zadaianchuk, Seitzer, and Martius 2023) to decompose input video frames into spatio-temporal object slots. We initially experimented with the STEVE model (Singh, Wu, and Ahn 2022), a widely used and promising approach, but found that it failed to consistently isolate entities



Figure 3: Visuals from Distraction Control Suite. From top to bottom: cheetah-run, walker-run, hopper-hop, humanoid-walk. From left to right: the distracted observation (background video, color, and camera position variations), the non-distracted observation, the mixture of slot decoder masks obtained after object-centric pretraining, and the main object slot decoder mask selected after object-centric pretraining.

such as the hopper in our tasks (see Supplementary Material). VideoSaur’s self-supervised architecture isolates individual entities within a scene, providing structured representations that are less susceptible to background noise and incidental motion. At the end of this step, we obtain an encoder, that directly maps a trajectory of observations  $\tau_o = (o_1, \dots, o_T)$  to a corresponding trajectory of slots  $\tau_s = (s_1, \dots, s_T)$ . VideoSAUR encodes each observation  $o_t$  in a fixed number  $K$  of slot vectors  $s_t^{(k)} \in \mathbb{R}^d$ , where the parameter  $K$  is set prior to training. Due to its transformer-based decoder, VideoSAUR allows each slot  $s_t^{(k)}$  to be projected back into the original observation space using attention maps as alpha masks. This enables us to visualize and interpret the spatial support of each slot in the image. We denote these by  $m_t^{(k)}$  and refer to them as object masks throughout the paper (see Figure 3 and Supplementary Material for visualizations). To improve consistency between training runs and reduce slot permutation issues, we employ fixed initialization of slots (see details in Supplementary Material), which ensures that the same slot index corresponds to a similar semantic object across different episodes.

**Slot Selection via Linear Action Probe.** In our setting, control-relevant entities typically include the main agent (e.g., the cheetah in DCS tasks or the robotic arm and gripper in DMW) and task-critical objects (e.g., hammer, ball, or cubes). Depending on the environment and the number of slots  $K$ , these entities may be encoded across one or more slots. To automatically identify which slots capture action-relevant information, we adopt a probing approach inspired by Alain and Bengio (2016), measuring how well individual

| Environment & Metric             | Task & mean over tasks | LAPO        | LAPO-clean  | LAPO-masks  |             | LAPO-slots  |             |
|----------------------------------|------------------------|-------------|-------------|-------------|-------------|-------------|-------------|
|                                  |                        | baseline    | upper-bound | ours        | +gain%      | ours        | +gain%      |
| DCS-Hard<br>(Normalized Return)  | cheetah-run            | 0.24 ± 0.02 | 0.76 ± 0.04 | 0.41 ± 0.03 | +32%        | 0.55 ± 0.04 | +58%        |
|                                  | hopper-hop             | 0.03 ± 0.01 | 0.27 ± 0.03 | 0.08 ± 0.01 | +20%        | 0.15 ± 0.02 | +50%        |
|                                  | walker-run             | 0.04 ± 0.01 | 0.32 ± 0.07 | 0.06 ± 0.01 | +6%         | 0.12 ± 0.01 | +27%        |
|                                  | humanoid-walk          | 0.02 ± 0.01 | 0.06 ± 0.01 | 0.04 ± 0.02 | +47%        | 0.06 ± 0.01 | +105%       |
|                                  | <b>mean</b>            | 0.08 ± 0.01 | 0.35 ± 0.04 | 0.15 ± 0.02 | <b>+26%</b> | 0.22 ± 0.02 | <b>+52%</b> |
| DCS<br>(Normalized Return)       | cheetah-run            | 0.34 ± 0.06 | 0.76 ± 0.04 | 0.54 ± 0.09 | +48%        | 0.50 ± 0.07 | +38%        |
|                                  | hopper-hop             | 0.05 ± 0.01 | 0.27 ± 0.03 | 0.15 ± 0.01 | +46%        | 0.19 ± 0.02 | +64%        |
|                                  | walker-run             | 0.10 ± 0.01 | 0.32 ± 0.07 | 0.21 ± 0.05 | +52%        | 0.18 ± 0.03 | +36%        |
|                                  | humanoid-walk          | 0.02 ± 0.01 | 0.06 ± 0.01 | 0.05 ± 0.02 | +67%        | 0.09 ± 0.02 | +174%       |
|                                  | <b>mean</b>            | 0.13 ± 0.02 | 0.35 ± 0.04 | 0.24 ± 0.04 | <b>+50%</b> | 0.24 ± 0.03 | <b>+50%</b> |
| DMW<br>(Normalized Success Rate) | hammer                 | 0.75 ± 0.07 | 0.98 ± 0.01 | 0.96 ± 0.01 | +91%        | 0.99 ± 0.02 | +102%       |
|                                  | bin-picking            | 0.18 ± 0.08 | 0.74 ± 0.10 | 0.49 ± 0.10 | +56%        | 0.33 ± 0.08 | +27%        |
|                                  | basketball             | 0.17 ± 0.03 | 0.51 ± 0.07 | 0.34 ± 0.09 | +50%        | 0.37 ± 0.09 | +58%        |
|                                  | soccer                 | 0.14 ± 0.06 | 0.36 ± 0.08 | 0.21 ± 0.08 | +34%        | 0.23 ± 0.06 | +41%        |
|                                  | <b>mean</b>            | 0.31 ± 0.06 | 0.65 ± 0.06 | 0.50 ± 0.07 | <b>+55%</b> | 0.48 ± 0.06 | <b>+50%</b> |

Table 1: Object-centric pretraining substantially improves robustness to visual distractors, where standard imitation learning from observation methods like LAPO struggle. In environments with visual distractors: dynamic background videos (DCS and DMW) and additional camera motion and color variation (DCS-Hard), LAPO’s performance is significantly degraded compared to the clean-data upper bound. We report normalized evaluation returns (DCS) and success rates (DMW) for a behavior cloning (BC) agent trained on latent actions, where a score of 1.0 corresponds to a BC agent trained on the full set of ground-truth actions. Each value averages performance across varying amounts of fine-tuning data and three random seeds. The relative performance gain is defined as:  $\frac{\text{LAPO-slots/masks} - \text{LAPO}}{\text{LAPO-clean} - \text{LAPO}} \times 100\%$ . This measures the fraction of the recoverable performance gap between the distracted baseline (LAPO) and the clean-data performance (LAPO-clean) that our method recovers. LAPO-slots and LAPO-masks achieve large improvements over LAPO, closing up to  $\sim 50\%$  of the gap, demonstrating that object-centric representations effectively mitigate the negative effects of visual distractions.

slot representations predict ground-truth actions.

We first apply Principal Component Analysis (PCA) to slot encodings computed from a small set of labeled trajectories, reducing dimensionality and noise. Then, a linear regressor is trained on resulting representations to predict the true action. To ensure robustness, we evaluate performance using 5-fold cross-validation and report the average test Mean Squared Error (MSE), which we refer to as the **Linear Action Probe score**. Lower MSE indicates stronger action predictivity and thus higher relevance to control.

We select the most predictive slots as the final set of relevant slots  $\mathcal{S}^* = \{\mathbf{s}^{(k)} \mid k \in \mathcal{K}^*\}$ , where  $\mathcal{K}^*$  denotes the selected slot indices. This selection is performed once per dataset after OC pretraining, leveraging the fixed slot initialization, ensuring stable and consistent slot interpretation across episodes.

**Latent Action Modeling.** Utilizing the selected object-centric representations, we train a latent action model inspired by LAPO (Schmidt and Jiang 2024).

**LAPO-slots** operates purely in the latent space. Its inverse-dynamics model  $\mathbf{z}_t \sim f_{\text{IDM}}^{\text{s}}(\cdot \mid \mathbf{s}_t, \mathbf{s}_{t+1})$  and forward-dynamics model  $\hat{\mathbf{s}}_{t+q} \sim f_{\text{FDM}}^{\text{s}}(\cdot \mid \mathbf{s}_t, \mathbf{z}_t)$  are trained to reconstruct the next slot embeddings, minimizing  $\|\hat{\mathbf{s}}_{t+1} - \mathbf{s}_{t+1}\|^2$ .

**LAPO-masks** operates in pixel space. It first creates a

filtered image by applying the object masks from the selected slots to the input frame. The dynamics models are then trained to reconstruct this filtered image at the next timestep. Visualizations of the filtered images can be found in Figures 3 and 5b and the Supplementary Material.

**Behavior Cloning and Finetuning.** The inferred latent actions serve as proxies for actual action labels. We train a behavior cloning (BC) agent to predict these latent actions, using the same dataset as for latent action learning. To evaluate the pre-training effectiveness, as a final stage, we fine-tune the resulting agent on a limited set of trajectories with ground-truth action labels (no more than 2.5% of total data), in line with (Schmidt and Jiang 2024; Ye et al. 2024; Nikulin et al. 2025). The finetuned BC agent scores for different numbers of labeled trajectories are presented on Figure 4.

## 4 Experiments

Our experiments are designed to address the following core questions: Can object-centric representations disentangle task-relevant motion from distracting visual elements? How much supervision is needed to train effective latent action models under high visual distraction? Can we automate slot selection to reduce the dependency on domain expertise or manual annotation?

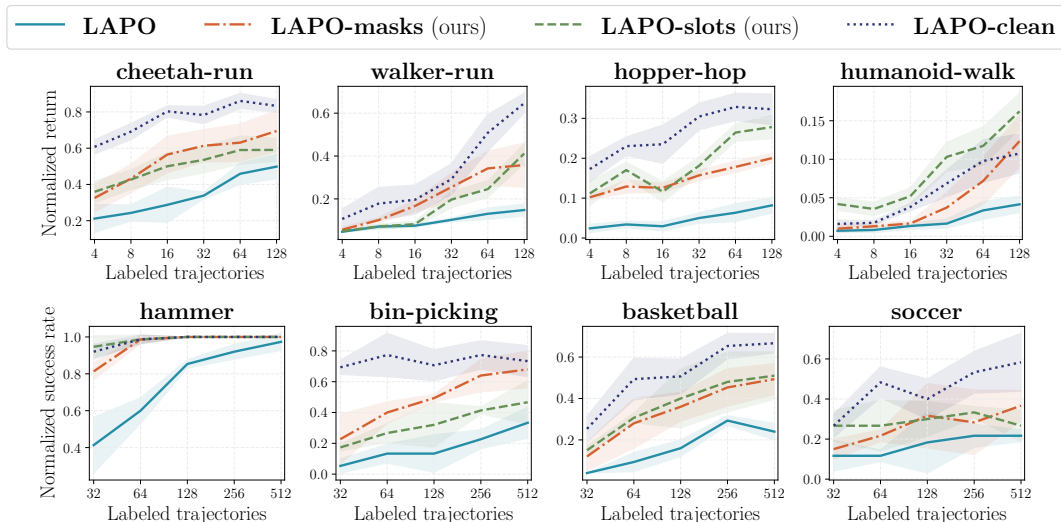


Figure 4: Normalized evaluation returns and success rates of the BC agent trained on latent actions for varying numbers of fine-tuning labeled trajectories. TL;DR: Object-centric learning improves evaluation returns in DCS tasks and success rate in goal-based DMW task for all tasks. The plots are arranged from left to right in order of increasing task complexity. The values are averaged across three random seeds. The BC agent trained with access to the full dataset of ground-truth actions would return a score of 1 for each task.

## Datasets

We evaluate on two benchmark environments: Distracting Control Suite (DCS) and Distracting Meta-World (DMW). Details on data collection are in Supplementary Material.

**Distracting Control Suite (DCS).** DCS (Stone et al. 2021) extends DeepMind Control with three visual distractions: (1) dynamic backgrounds (DAVIS 2017 videos), (2) color variations (hue/saturation shifts), and (3) camera perturbations. We focus on dynamic backgrounds (**DCS**) and report full results on all distractions (**DCS-hard**) in Figure 1 and table 1 and Supplementary Material. Tasks: cheetah-run, walker-run, hopper-hop, humanoid-walk (increasing complexity). Experts were trained on privileged state; datasets consist of observation-action pairs. Behavior cloning, trained on privileged true actions, is able to achieve expert performance.

**Distracting Meta-World (DMW).** DMW extends Meta-World (Yu et al. 2019) with DCS-style dynamic backgrounds. Tasks (hammer, bin-picking, basketball, soccer) involve multi-object interactions and compositional reasoning. Experts use Meta-World’s oracle policies.

## Models

To conduct the experiments, 4 models were trained: **LAPO**, **LAPO-clean**, **LAPO-slots** and **LAPO-masks**. All models were trained on the same datasets. More details on training can be found in Supplementary Material.

The baseline model is LAPO, which is trained on observations with distractors, following the Schmidt and Jiang (2024) procedure (**LAPO** in the figures). We use it as a baseline to demonstrate the currently existing limitations of latent action pretraining.

We also trained LAPO on clean data without distractors as our upper bound to illustrate the performance gap caused by distractors (**LAPO-clean** in the figures).

LAPO-slots and LAPO-masks are the models that follow the object-centric latent action pre-training pipeline described in Section 3 (respectively, **LAPO-slots** and **LAPO-masks** in the figures). LAPO-slots uses slot embeddings extracted from VideoSAUR as input to the latent action model, LAPO-masks uses slot masks, treating each mask as an attention map applied to the original frame.

## Evaluation performance

To assess the quality of learned latent actions, we train a BC agent  $\pi_{BC}(a_t | z_t)$  on the inferred latent actions  $z_t$ , then fine-tune it on a small number of trajectories (0.1% – 2.5% from the entire dataset) with true action labels. The small size of the action-labeled sample mirrors real-world deployment, where labeling is expensive. The final performance is measured as **normalized episodic return** for DCS and **normalized success rate** for DMW, where a score of 1 corresponds to an oracle BC policy (see Supplementary Material) trained on the entire dataset with ground-truth actions. As shown in Table 1, both LAPO-slots and LAPO-masks strongly outperform the standard LAPO baseline for both the DCS and DMW domains. Notably: In DCS, LAPO-slots improves the gap between LAPO performance with and without distractors by 50% on average, with even greater gains in complex tasks like ‘humanoid-walk‘ (+174%). In DMW, the improvements are consistent, with LAPO-slots achieving a +50% improvement on average over the baseline. These results support our hypothesis: object-centric representations help filter out distractors and retain actionable information. Figure 4 shows evaluation returns across increasing numbers

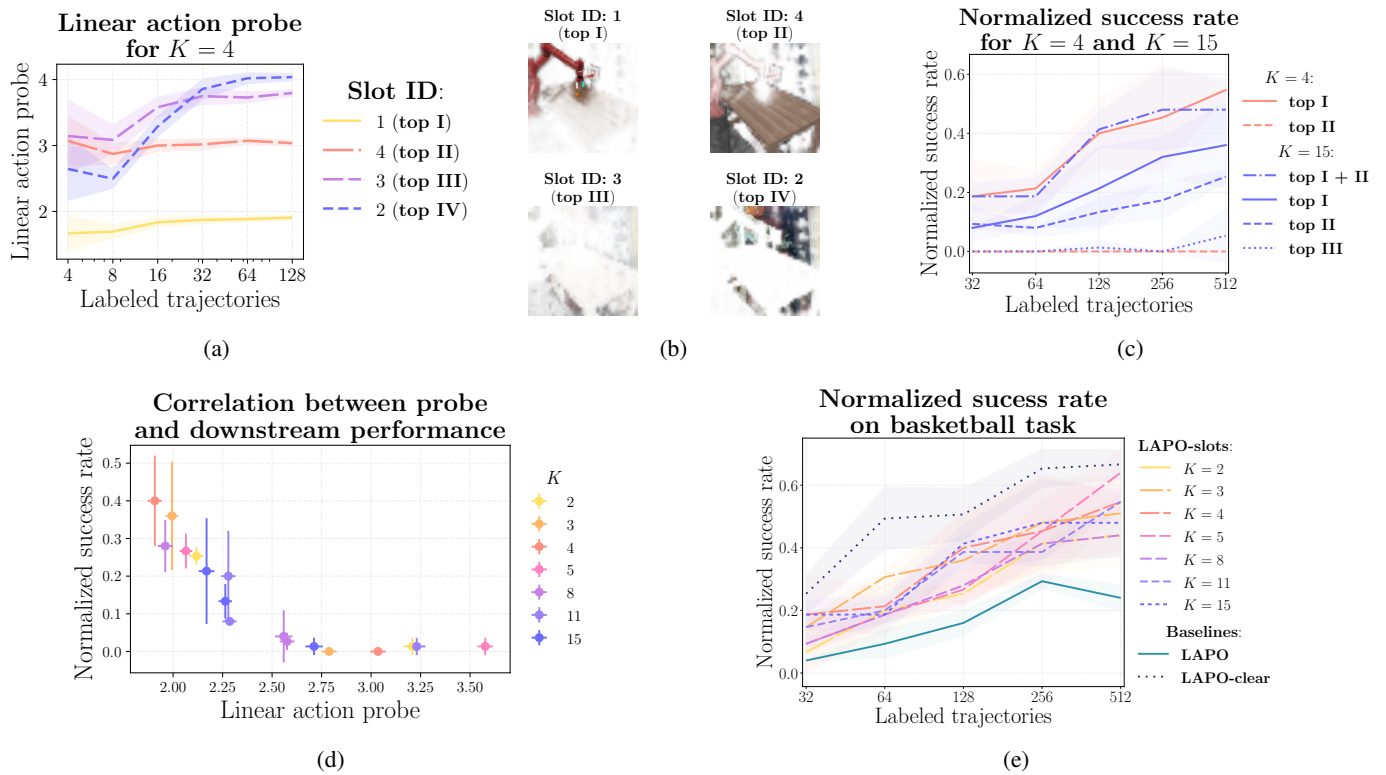


Figure 5: Slot selection study on basketball task. (a,b) Linear action probes under varying labeled trajectories budget and corresponding examples of slot masks for basketball task for  $K = 4$ . (c) BC performance on different slots for  $K = 4$  and  $K = 15$  on basketball task. (d) Linear action probe scores vs. normalized BC success rates (on 128 trajectories). TL;DR: Probe-based selection correlates with downstream performance. (e) BC performance on varying  $K$  parameter on basketball task. Full-action BC achieves score of 1. TL;DR: Regardless of the  $K$  parameter LAPO-slots outperform baseline LAPO.

of fine-tuning trajectories. Both slot-based methods show better generalization from fewer labeled examples.

### Slot Selection Investigation

Object-centric models enable interpretable, modular control by attributing dynamics to individual entities, but identifying which slots encode control-relevant information remains challenging.

We show that linear action probes, trained to predict actions from individual slot representations, can automatically identify relevant slots (method in Section 3). Probe scores align with manual analysis (Figures 5a and 5b) and strongly correlate with downstream BC performance (Figure 5d), validating them as a reliable relevance metric.

Probing reveals that useful information concentrates in either a small set of dominant slots (one slot for  $K \leq 8$  or a tightly coupled pair for  $K > 8$ ), with sharply worse scores for all others. Only these top-ranked slots yield non-zero BC performance (Figure 5c for  $K = 4$  and  $K = 15$  cases); the rest are ineffective. For  $K > 8$ , concatenating the top two slots improves performance, confirming they capture complementary information. Thus, probe score distributions guide both selection and fusion of meaningful slots.

Additional visualizations are presented in the Supplementary Material.

### Varying Number of Slots

The number of slots  $K$  is a key hyperparameter in object-centric models like VideoSAUR (Zadaianchuk, Seitzer, and Martius 2023), often set based on scene complexity. While adaptive allocation has been explored (Fan et al. 2024), most methods use a fixed  $K$ , making robustness to its choice crucial for real-world deployment. We evaluate LAPO-slots for  $K = 2$  to 15 on DMW tasks. As shown in Figure 5e, performance remains consistently strong across all  $K$ , demonstrating robustness to the choice of  $K$ , which is a practical advantage when the number of relevant entities is unknown. Too few slots ( $K = 2$ ) cause collapsed representations, merging multiple entities into one. In contrast, larger  $K$  introduces duplicated slots with nearly identical representations. For  $K = 5$ , only two unique slots emerge; for  $K = 8$ , four; for  $K = 11$  or 15, seven. Despite redundancy, downstream performance remains stable, as slot selection focuses on the most informative ones.

Interestingly, when applied to DMW, we expected a clear separation of distinct objects such as the robotic arm, hammer, ball, etc. Instead, VideoSAUR grouped these task-relevant entities into a single slot even when the number of slots  $K$  was greater than the amount of objects. We attribute this to limited trajectory diversity: variations arise from the manip-

ulated object’s initial position, with little change in arm or target configuration. Without sufficient variability, the model lacks incentive to disentangle entities. This reflects a general tendency in object-centric models: they favor functional over geometric decomposition, merging semantically distinct but behaviorally coupled entities. Importantly, this behavior mirrors real-world conditions, where data is often imbalanced and incomplete.

### Increasing distractions difficulty

To better understand how different types of visual distractions impact latent action learning, we evaluate our methods under increasingly challenging conditions in the Distracting Control Suite (DCS) (Stone et al. 2021) environment. In this work, we primarily focus on dynamic backgrounds as the core challenge, but we also test combinations of all three distractor types (backgrounds, camera position, and color variations) on DCS (marked DCS-hard on Figure 1 and Table 1) to assess robustness.

As shown in Figure 1, all methods suffer performance degradation under additional distractors, for example, LAPO performance drops by x2.7 under background distractions and by x4.4 under additional distractors, compared to the performance without distractors. However, LAPO-slots maintains a consistent relative improvement over baseline (50% on DCS  $\rightarrow$  52% on DCS-hard), while LAPO-masks degrades (50% on DCS  $\rightarrow$  26% on DCS-hard). This difference probably stems from VideoSAUR’s use of the DINOv2 encoder (Oquab et al. 2023), which captures high-level semantic features and is more robust to appearance changes than the CNN-based encoder in LAPO and LAPO-masks. These findings suggest that slot-based representations not only isolate task-relevant entities but also encode objects using features that generalize better under real-world visual perturbations.

## 5 Related work

Vision-Language-Action (VLA) models like LAPA (Ye et al. 2024), Genie (Bruce et al. 2024), and UniVLA (Bu et al. 2025) have shown impressive capabilities by training on large-scale video data and deploying on real robots. However, scale can make it difficult to isolate specific failure modes. Our work focuses on one such critical failure: the sensitivity of the underlying Latent Action Models to visual distractors. To enable a controlled study of this problem, we use the Latent Action Policy Optimization (LAPO) framework (Schmidt and Jiang 2024) within simulated environments. While effective in clean settings, LAPO’s performance degradation under distraction makes it an ideal baseline to rigorously evaluate robustness improvements.

Our work confronts this distractor challenge by integrating object-centric learning (OCL). OCL has shown significant promise for compositional generation (Akan and Yemez 2025), building structured world models (Heins et al. 2025; Barcellona et al. 2024) and compositional video prediction (Lee et al. 2025). The application of OCL to latent action learning has been explored by Villar-Corrales and Behnke (2025), who use object-centric models for latent action learning in robotic simulations. However, their work, along with

most OCL-based world models, does not investigate performance under heavy visual distraction, a key focus of our paper. Furthermore, their reliance on the SAVi architecture (Kipf et al. 2021) for object discovery proves less robust in complex, distracting environments than the VideoSAUR model (Zadaianchuk, Seitzer, and Martius 2023) we employ (see Supplementary Material).

Another work that directly addresses latent action learning with distractors is Nikulin et al. (2025). While their goal is similar, their method is orthogonal: they propose a reconstruction-free framework that requires a small set of labeled trajectories to provide supervision during pre-training. Together, these works highlight the growing interest in robust latent action learning under visual distractions. Our contribution uniquely bridges object-centric perception and latent action modeling, emphasizing slot stability, interpretability, and reduced supervision requirements.

## 6 Discussion and Limitations

Our results demonstrate that object-centric (OC) representations can significantly mitigate the impact of distractors when learning latent actions from video. By disentangling scenes into meaningful slots, our approach allows latent action models to focus on causal dynamics rather than spurious correlations, providing a strong inductive bias for learning in noisy environments.

The effectiveness of our method is inherently linked to the capabilities of the underlying OCL model. We believe that as OCL methods continue to improve and scale from simulated to more complex, real-world data, such as VideoSAUR improved upon previous models, so too will the performance and applicability of our approach. Our contribution is to show that integrating object decomposition is a viable path for robust latent action learning, thereby motivating further investment in the underlying OCL technologies.

The reliance on current OCL frameworks introduces several limitations. While powerful segmentation models could isolate objects, they are not inherently designed to handle the dynamic, action-correlated distractors (e.g., camera motion, color shifts) which are explored in our work. The challenges become even greater in scenarios with heavy occlusions or multiple camera viewpoints. Furthermore, current slot-based models like VideoSAUR lack memory mechanisms to handle objects entering or leaving the scene, and often assume a fixed number of objects  $K$ . While recent work has explored solutions like dictionary-based architectures and adaptive slot allocation (Đukić, Lebaillly, and Tuytelaars 2025; Fan et al. 2024), these remain open challenges.

Finally, we observed that the unsupervised nature of OCL models, particularly under limited data diversity, can lead to non-intuitive collapsed decompositions due to lack of direct control over the learned representations. A promising direction to mitigate this is to augment training data with generative approaches (Luo et al. 2025; Sikchi et al. 2024) or develop weakly-supervised OCL methods to learn more controllable representations.

## Acknowledgments

This work was supported by the The Ministry of Economic Development of the Russian Federation in accordance with the subsidy agreement (agreement identifier 000000C313925P4H0002; grant No 139-15-2025-012).

## References

- Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Akan, A. K.; and Yemez, Y. 2025. Slot-Guided Adaptation of Pre-trained Diffusion Models for Object-Centric Learning and Compositional Generation. In *International Conference on Learning Representations*.
- Alain, G.; and Bengio, Y. 2016. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*.
- Åström, K. J. 1965. Optimal control of Markov processes with incomplete state information I. *Journal of mathematical analysis and applications*, 10: 174–205.
- Barcellona, L.; Zadaianchuk, A.; Allegro, D.; Papa, S.; Ghidoni, S.; and Gavves, E. 2024. Dream to manipulate: Compositional world models empowering robot imitation learning with imagination. *arXiv preprint arXiv:2412.14957*.
- Bommasani, R.; Hudson, D. A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M. S.; Bohg, J.; Bosselut, A.; Brunskill, E.; et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.
- Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.
- Bruce, J.; Dennis, M. D.; Edwards, A.; Parker-Holder, J.; Shi, Y.; Hughes, E.; Lai, M.; Mavalankar, A.; Steigerwald, R.; Apps, C.; et al. 2024. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*.
- Bu, Q.; Yang, Y.; Cai, J.; Gao, S.; Ren, G.; Yao, M.; Luo, P.; and Li, H. 2025. UniVLA: Learning to Act Anywhere with Task-centric Latent Actions. *arXiv preprint arXiv:2505.06111*.
- Chen, X.; Guo, J.; He, T.; Zhang, C.; Zhang, P.; Yang, D. C.; Zhao, L.; and Bian, J. 2024. Igor: Image-goal representations are the atomic control units for foundation models in embodied ai. *arXiv preprint arXiv:2411.00785*.
- Cui, Z. J.; Pan, H.; Iyer, A.; Haldar, S.; and Pinto, L. 2024. DynaMo: In-Domain Dynamics Pretraining for Visuo-Motor Control. *arXiv preprint arXiv:2409.12192*.
- Dehghani, M.; Djolonga, J.; Mustafa, B.; Padlewski, P.; Heek, J.; Gilmer, J.; Steiner, A. P.; Caron, M.; Geirhos, R.; Alabdulmohsin, I.; et al. 2023. Scaling vision transformers to 22 billion parameters. In *International Conference on Machine Learning*, 7480–7512. PMLR.
- Dittadi, A.; Papa, S.; De Vita, M.; Schölkopf, B.; Winther, O.; and Locatello, F. 2022. Generalization and Robustness Implications in Object-Centric Learning. In *International Conference on Machine Learning*.
- Edwards, A.; Sahni, H.; Schroecker, Y.; and Isbell, C. 2019. Imitating latent policies from observation. In *International conference on machine learning*, 1755–1763. PMLR.
- Efroni, Y.; Misra, D.; Krishnamurthy, A.; Agarwal, A.; and Langford, J. 2022. Provable RL with Exogenous Distractors via Multistep Inverse Dynamics. *arXiv:2110.08847*.
- Fan, K.; Bai, Z.; Xiao, T.; He, T.; Horn, M.; Fu, Y.; Locatello, F.; and Zhang, Z. 2024. Adaptive slot attention: Object discovery with dynamic slot number. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23062–23071.
- Guruprasad, P.; Sikka, H.; Song, J.; Wang, Y.; and Liang, P. P. 2024. Benchmarking Vision, Language, & Action Models on Robotic Learning Tasks. *arXiv preprint arXiv:2411.05821*.
- Heins, C.; Van de Maele, T.; Tschantz, A.; Linander, H.; Markovic, D.; Salvatori, T.; Pezzato, C.; Catal, O.; Wei, R.; Koudahl, M.; et al. 2025. AXIOM: Learning to Play Games in Minutes with Expanding Object-Centric Models. *arXiv preprint arXiv:2505.24784*.
- Kaelbling, L. P.; Littman, M. L.; and Cassandra, A. R. 1998. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2): 99–134.
- Kipf, T.; Elsayed, G. F.; Mahendran, A.; Stone, A.; Sabour, S.; Heigold, G.; Jonschkowski, R.; Dosovitskiy, A.; and Greff, K. 2021. Conditional object-centric learning from video. *arXiv preprint arXiv:2111.12594*.
- Lee, Y.-C.; Lu, E.; Rumbley, S.; Geyer, M.; Huang, J.-B.; Dekel, T.; and Cole, F. 2025. Generative omnimatte: Learning to decompose video into layers. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 12522–12532.
- Lin, F.; Hu, Y.; Sheng, P.; Wen, C.; You, J.; and Gao, Y. 2024. Data scaling laws in imitation learning for robotic manipulation. *arXiv preprint arXiv:2410.18647*.
- Locatello, F.; Weissenborn, D.; Unterthiner, T.; Mahendran, A.; Heigold, G.; Uszkoreit, J.; Dosovitskiy, A.; and Kipf, T. 2020. Object-centric learning with slot attention. *Advances in neural information processing systems*, 33: 11525–11538.
- Luo, C.; Zeng, Z.; Du, Y.; and Sun, C. 2025. Solving New Tasks by Adapting Internet Video Knowledge. *arXiv preprint arXiv:2504.15369*.
- McCarthy, R.; Tan, D. C.; Schmidt, D.; Acero, F.; Herr, N.; Du, Y.; Thuruthel, T. G.; and Li, Z. 2024. Towards Generalist Robot Learning from Internet Video: A Survey. *arXiv preprint arXiv:2404.19664*.
- Misra, D.; Saran, A.; Xie, T.; Lamb, A.; and Langford, J. 2024a. Towards Principled Representation Learning from Videos for Reinforcement Learning. *arXiv preprint arXiv:2403.13765*.
- Misra, D.; Saran, A.; Xie, T.; Lamb, A.; and Langford, J. 2024b. Towards Principled Representation Learning from Videos for Reinforcement Learning. *arXiv:2403.13765*.

Nikulin, A.; Zisman, I.; Tarasov, D.; Lyubaykin, N.; Polubarov, A.; Kiselev, I.; and Kurenkov, V. 2025. Latent Action Learning Requires Supervision in the Presence of Distractors. *arXiv preprint arXiv:2502.00379*.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H. V.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; Howes, R.; Huang, P.-Y.; Xu, H.; Sharma, V.; Li, S.-W.; Galuba, W.; Rabbat, M.; Assran, M.; Ballas, N.; Synnaeve, G.; Misra, I.; Jegou, H.; Mairal, J.; Labatut, P.; Joulin, A.; and Bojanowski, P. 2023. DINOv2: Learning Robust Visual Features without Supervision.

Pomerleau, D. A. 1988. Alvin: An autonomous land vehicle in a neural network. *Advances in neural information processing systems*, 1.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Schmidt, D.; and Jiang, M. 2024. Learning to Act without Actions. In *The Twelfth International Conference on Learning Representations*.

Sikchi, H.; Agarwal, S.; Jajoo, P.; Parajuli, S.; Chuck, C.; Rudolph, M.; Stone, P.; Zhang, A.; and Niekum, S. 2024. RL Zero: Zero-Shot Language to Behaviors without any Supervision. *arXiv preprint arXiv:2412.05718*.

Singh, G.; Wu, Y.-F.; and Ahn, S. 2022. Simple unsupervised object-centric learning for complex and naturalistic videos. *Advances in Neural Information Processing Systems*, 35: 18181–18196.

Stone, A.; Ramirez, O.; Konolige, K.; and Jonschkowski, R. 2021. The Distracting Control Suite – A Challenging Benchmark for Reinforcement Learning from Pixels. *arXiv:2101.02722*.

Đukić, N.; Lebailly, T.; and Tuytelaars, T. 2025. OCEBO: Object-Centric Pretraining by Target Encoder Bootstrapping. In *The Thirteenth International Conference on Learning Representations*.

Villar-Corrales, A.; and Behnke, S. 2025. PlaySlot: Learning Inverse Latent Dynamics for Controllable Object-Centric Video Prediction and Planning. *arXiv:2502.07600*.

Wang, Y.; Wan, S.; Gan, L.; Feng, S.; and Zhan, D.-C. 2024. AD3: Implicit Action is the Key for World Models to Distinguish the Diverse Visual Distractors. *arXiv preprint arXiv:2403.09976*.

Ye, S.; Jang, J.; Jeon, B.; Joo, S.; Yang, J.; Peng, B.; Mandlkar, A.; Tan, R.; Chao, Y.-W.; Lin, B. Y.; Liden, L.; Lee, K.; Gao, J.; Zettlemoyer, L.; Fox, D.; and Seo, M. 2024. Latent Action Pretraining from Videos. *arXiv:2410.11758*.

Yu, T.; Quillen, D.; He, Z.; Julian, R.; Hausman, K.; Finn, C.; and Levine, S. 2019. Meta-World: A Benchmark and Evaluation for Multi-Task and Meta Reinforcement Learning. In *Conference on Robot Learning (CoRL)*.

Zadaianchuk, A.; Seitzer, M.; and Martius, G. 2023. Object-Centric Learning for Real-World Videos by Predicting Temporal Feature Similarities. In *Thirty-seventh Conference on Neural Information Processing Systems*.