CONTROL-CENTRIC REPRESENTATION LEARNING USING ACTION-FREE DATASETS WITH DISTINCT POLICIES

Anonymous authors
Paper under double-blind review

ABSTRACT

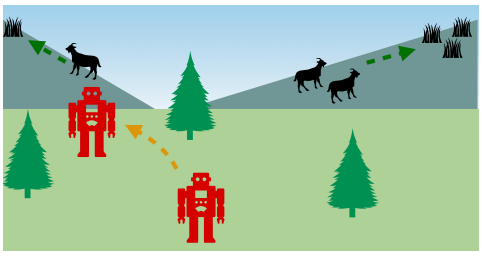
Action- and reward-free data are available in settings such as robot manipulation, video game play, and human egocentric recordings. Prior work uses these data to learn visual representations for downstream control with limited action-labeled data. However, time-correlated visual distractions can degrade the representations learned by these methods. Critically, in many common visual control tasks, a naive encoder cannot distinguish noise from controllable features, leading to poor generalization. In this work, we consider action-free trajectories grouped by known data-collection policy within a shared environment. We introduce `CARDPol`, a representation-learning objective that predicts policy identity from pairs of observations. Under appropriate assumptions, we show that the ground-truth endogenous representation minimizes our objective. Empirically, our proposed loss learns representations that exhibit reduced sensitivity to time-correlated exogenous noise. `CARDPol` improves distracted LIBERO success by 3–11 percentage points and DMC video-distraction returns by 25–53% over the strongest evaluated baseline in each setting. On human egocentric video from EgoDex, it achieves the lowest mean wrist-position decoding error, with the closest baseline exhibiting 15% worse error. Our results demonstrate that policy identity can provide useful supervision for learning control-relevant representations from high-dimensional observations without action labels.

1 INTRODUCTION

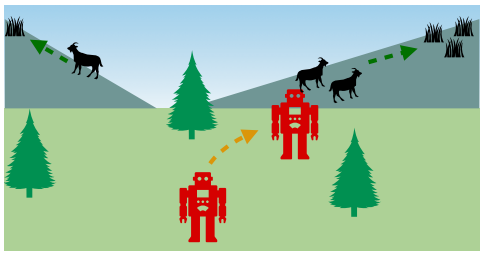
Learning to perform control tasks in high-dimensional or visual observation spaces can be challenging due to the difficulty of extracting the often relatively lower-dimensional space of relevant features. A key question in learning policies for the real world is how to efficiently represent these high-dimensional observations such that control can be robust to irrelevant features. One promising paradigm to solve this problem is to rely on offline interaction data in an environment (Schmidt & Jiang, 2024; Ma et al., 2022; Ghosh et al., 2023; Islam et al., 2023). For example, in robot navigation, large amounts of egocentric video data are available from real-world environments (Zhang et al., 2024), collected either by tele-operated robots or by humans. Similarly, in video game environments, countless hours of screen recordings are available from players streaming their gameplay (Baker et al., 2022; Ghosh et al., 2023). However, this data is usually passive in nature; the control inputs (i.e., actions) and objectives (i.e., scalar per-step rewards) are unknown. Learning representations for control in this data regime is difficult compared to traditional offline or online settings in which control inputs and objectives can help steer what features are important. This gap makes it difficult to distinguish between meaningful and noise features in passive observations (Misra et al., 2024). We consider a setting in which these action-free trajectories are grouped by the policies that generated them. Our goal is to use policy identity as supervision for learning control-relevant representations without observing actions or rewards.

To distinguish control-relevant state from uncontrollable variation, Efroni et al. (2022) introduced the Exogenous Block Markov Decision Process (Ex-BMDP) framework. In this model, observations contain an endogenous state, whose dynamics depend on the agent's actions, and an exogenous state, which evolves independently of those actions. We consider data collected by policies that depend only on the endogenous state. The representation-learning goal is to retain information about the endogenous state while reducing sensitivity to exogenous variation.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Depiction of two observations of a robot following Policy A



(b) Depiction of two observations of a robot following Policy B

Figure 1: The robot in Figure 1a follows policy A (go left), while the robot in Figure 1b follows policy B (go right). To classify a trajectory as coming from policy A or B, it suffices to observe the robot's location, ignoring the trees, sky, and goats.

Popular action-free representation learning methods such as value implicit pre-training (VIP) (Ma et al., 2022) and intention-conditioned value functions (ICVF) (Ghosh et al., 2023) do not explicitly distinguish endogenous dynamics from exogenous visual variation. This can be a major disadvantage in settings where most features of an observation are irrelevant to or uncontrolled by the agent (Islam et al., 2023; Nikulin et al., 2025; Garrido et al., 2026). In fact, in the *action-free* offline setting, Misra et al. (2024) show that without additional assumptions, learning an encoder to factor endogenous and exogenous components is not possible in Ex-BMDPs. However, Levine et al. (2025) show that with data from two behaviorally-distinct agents it is possible to learn ϕ in a provably sample-efficient manner. With inspiration from the theoretical guarantees of Levine et al. (2025), we develop a method for learning control-relevant representations from action-free trajectories collected under distinct policies. When policies act in the same environment, their different behaviors provide supervision about the underlying endogenous dynamics. For this supervision to be useful, the policies must act independently of exogenous variation, and distracting visual context must not provide a shortcut for identifying the policy. Figure 1 illustrates the intuition: policy A moves the robot left, whereas policy B moves it right. Observing the robot's movement is sufficient to distinguish these behaviors, without representing changes in the surrounding trees, sky, or animals.

We introduce Controllable Action-free Representations using Distinct Policies (`CARDPol`), which jointly trains a shared visual encoder and a classifier to predict policy identity from observation pairs. The objective is directly implementable with neural networks and does not require the timestep-specific encoders or discrete-state construction used by CRAFT. Under the assumptions specified in Section 4, we show that minimizing the objective preserves sufficient information to recover the endogenous state. This recovery guarantee concerns the information retained by the representation; we separately evaluate its sensitivity to exogenous variation.

We evaluate `CARDPol` through downstream control, endogenous-state decoding, and controlled changes to visual distractions. In DeepMind Control Suite (Tassa et al., 2018), we pretrain encoders on action-free trajectories and evaluate them through offline reinforcement learning with frozen representations. In LIBERO (Liu et al., 2023), we jointly optimize `CARDPol` with target-task behavior cloning, using task identity as a proxy for policy identity. Additional control results are provided for OGBench (Park et al., 2025). We also evaluate pose decoding from representations learned on task-annotated human video in EgoDex (Hoque et al., 2026). The LIBERO and EgoDex experiments test task labels as practical proxies for policy identity. Our results show that policy-based supervision can support downstream control while retaining endogenous information and reducing sensitivity to distracting visual variation.

2 Related Works

We present literature relevant to 1) learning policies from passive data in the context of reinforcement learning and 2) extracting controllable latent representations from reward-free or action-free data.

Learning from Action-free Trajectories. Pre-training a visual encoder for use in downstream RL tasks is a pipeline used in many settings (Seo et al., 2022; Nair et al., 2023; Schmidt & Jiang, 2024;

Schmeckpeper et al., 2021; Zheng et al., 2023; Misra et al., 2024; Ghosh et al., 2023; Majumdar et al., 2023). Of particular interest is training representations that are capable of either 1) capturing value functions (Ma et al., 2022; 2023; Ghosh et al., 2023) or 2) encoding inter-frame variance as a *latent* action (Schmidt & Jiang, 2024; Ye et al., 2025; Cui et al., 2024). These objectives provide supervision for learning representations without action labels, but do not explicitly distinguish endogenous dynamics from exogenous visual variation. For example, VIP (Ma et al., 2022) uses temporal relationships between observations to learn a goal-conditioned value representation, while ICVF (Ghosh et al., 2023) learns intention-conditioned value functions through a successor-measure formulation. Time-correlated distractions can also carry information about these temporal relationships, creating an incentive to represent features that are irrelevant to control. `CARDPol` instead uses policy identity as supervision, exploiting differences in behavior across policy-partitioned trajectories.

Extracting Controllable Representations. In many real-world visual control settings, many features evolve as an uncontrollable Markov chain. To formalize this problem, Efroni et al. (2022) introduced the Exogenous Block MDP (Ex-BMDP) framework, which has since been explored in the offline setting in numerous works (Misra et al., 2024; Islam et al., 2023; Lamb et al., 2023; Levine et al., 2024; 2025). Several of these works propose using a "multi-step inverse" model which provably extracts the endogenous state from Ex-BMDPs using a loss function similar to our proposed method (Lamb et al., 2023; Islam et al., 2023; Levine et al., 2024). To learn such a mapping, Islam et al. (2023) introduce ACRO which attempts to predict the action taken at time t conditioned on observations at time t and $t + k$ for some $0 < k < K$. Intuitively, ACRO learns a useful representation of Ex-BMDP observations because the only information necessary to predict the action taken by a policy is the endogenous state. Under coverage assumptions and appropriate information bottlenecking, optimizing the multi-step inverse dynamics loss leads to a minimal representation of relevant control information. However, these prior multi-step inverse approaches assume access to action labels, which `CARDPol` does not.

3 PRELIMINARIES

We consider a reward-free Exogenous Block Markov Decision Process (Ex-BMDP) with observation space $\mathcal{X}$ and action space $\mathcal{A}$. Each observation $x \in \mathcal{X}$ is generated from an endogenous state $s \in \mathcal{S}$ and an exogenous state $e \in \mathcal{E}$. The latent transition dynamics factor as

$$P(s', e' \mid s, e, a) = P_{\text{end}}(s' \mid s, a)\, P_{\text{exo}}(e' \mid e).$$

It is assumed that there exists a function $\phi^* : \mathcal{X} \to \mathcal{S}$ that can map from the observation space to the endogenous state space. We assume that the behavior policies are exo-free: their action distributions depend only on the endogenous state. Specifically, for any x_1, x_2 satisfying $\phi^*(x_1) = \phi^*(x_2)$, each behavior policy π_i satisfies $\pi_i(\cdot \mid x_1) = \pi_i(\cdot \mid x_2)$.

The learner receives N datasets $\{\mathcal{D}_i\}_{i=1}^N$, where $\mathcal{D}_i$ contains observation trajectories collected under behavior policy π_i. The policy label i is known for each trajectory, but the policies themselves need not be accessible. The `CARDPol` objective uses only observations and policy labels; it requires neither action or reward labels nor ground-truth endogenous states. $\mathcal{D}(\pi_i, k)$ denotes the distribution of *observation pairs*, separated by k steps, collected under π_i. That is, we can sample $(x_t, x_{t+k}) \sim \mathcal{D}(\pi_i, k)$. Our goal is to learn an encoder $\phi : \mathcal{X} \to \mathcal{Z}$ that preserves endogenous information useful for downstream control while reducing sensitivity to exogenous variation. Policy-partitioned trajectories are available in several commonly used control benchmarks. For example, datasets like D4RL (Fu et al., 2021) (e.g., 'medium', 'expert', 'medium-replay', etc.) , ExORL (Yarats et al., 2022) (e.g., 'RND', 'DIAYN', etc.) , and OGBench (Park et al., 2025) (e.g., 'play', 'noise', etc.) provide dataset partitions associated with different data-collection behaviors. Such partitions can provide candidate policy classes, although a partition may contain trajectories generated by multiple behavior policies. Human-video and robotics datasets also often provide task or behavior annotations (Grauman et al., 2022; Damen et al., 2020; Mandlekar et al., 2021; Liu et al., 2023) that can serve as trajectory partitions, although these labels do not necessarily correspond exactly to policies in the Ex-BMDP sense.

CRAFT (Levine et al., 2025). Our method `CARDPol` is inspired by and improves upon the method outlined in Levine et al. (2025), called CRAFT. At a high level, CRAFT operates on a pair of datasets, each containing action-free trajectories that have been collected by behaviorally distinct agents (called

agent A and agent B). CRAFT learns a set of transition-conditioned classifiers $\{f_h\}_{h=1}^{H-1}$, one for each timestep h, to predict the log-odds ratio of the transition having been generated by agent A or agent B. Then, the values of each f_h are used to cluster different observation pairs together. Next, CRAFT learns a series of binary classifiers to de-duplicate the endogenous states in each cluster. Lastly, once the latent endogenous states have been identified, CRAFT learns an encoder ϕ_h for each timestep h that takes in an observation and predicts a discrete endogenous state. Further details on CRAFT can be found in Section F.

In this work we relax several assumptions of CRAFT that make direct application to the continuous-control settings considered here difficult. Accordingly, CRAFT is predicated on the following:

1. Each trajectory must start at identical endogenous states, *which is seldom possible to ensure with passively collected data in continuous state spaces.*
2. CRAFT requires discrete and deterministic endogenous dynamics, *which does not hold in the continuous-control environments considered in this work.*
3. CRAFT requires learning a separate encoder and classifier at each timestep in the trajectories, *which would require training a separate model at each timestep and does not directly match the shared-encoder formulation used in our experiments.*

One distinct advantage of CRAFT is that it provides explicit sample complexity guarantees, which `CARDPol` does not. However, these guarantees depend linearly on $\log(|\Phi|)$, where Φ is the hypothesis class of encoders. When applied to modern deep neural network encoders, such "parameter counting" bounds tend to be vacuously large, and do not usefully bound expected performance.

4 CARDPOL: CONTROLLABLE ACTION-FREE REPRESENTATIONS USING DISTINCT POLICIES

`CARDPol` learns representations by predicting which behavior policy generated a pair of temporally separated observations. We jointly train a visual encoder $\phi_\theta : \mathcal{X} \to \mathcal{Z}$, where $\mathcal{Z} \subseteq \mathbb{R}^d$, and a policy classifier f_ψ. The same encoder processes both observations, and the classifier uses their embeddings to predict the policy label associated with the trajectory. To construct a training example, we sample a policy label i uniformly from $\{1, \dots, N\}$, a temporal offset k uniformly from $\{1, \dots, K\}$, and an observation pair $(x_t, x_{t+k}) \sim D(\pi_i, k)$ from a trajectory collected under that policy. We optimize the encoder and classifier jointly using the cross-entropy objective

$$\mathcal{L}_{\texttt{CARDPol}}\left(f_\psi, \phi_\theta\right) = \mathop{\mathbb{E}}_{\substack{k\sim\mathcal{U}(1,K)\\ i\sim\mathcal{U}(1,N)}} \mathop{\mathbb{E}}_{(x_t,x_{t+k})\sim\mathcal{D}(\pi_i,k)} \left[-\log f_\psi\left(i \mid \phi_\theta(x_t), \phi_\theta(x_{t+k}); k\right)\right]. \tag{1}$$

Here, K is the maximum temporal offset and N is the number of policy classes. Uniform sampling over policy labels gives each class equal weight regardless of its dataset size. Gradients from the classification loss update both the classifier and the shared encoder.

Our method is related to multi-step inverse dynamics prediction, which has been shown to be an effective technique for filtering out exogenous noise and learning endogenous representations (Islam et al., 2023; Misra et al., 2024). Similarly, but without requiring actions, the `CARDPol` objective encourages representations that retain information about the endogenous state. Our analysis establishes endogenous-state recoverability under the finite-space, coverage, stationarity, noise-independence, and policy-diversity assumptions formalized in Section E. The diversity condition requires policies to induce distinguishable multi-step endogenous transition distributions. Under these assumptions:

Theorem 1. *The ground-truth encoder ϕ^* associated with an Ex-BMDP M minimizes $\mathcal{L}_{\texttt{CARDPol}}$. Furthermore, any encoder ϕ' which minimizes $\mathcal{L}_{\texttt{CARDPol}}$ is at least as representationally powerful as ϕ^*, in the sense that there exists a function σ such that $\forall x,\ \sigma(\phi'(x)) = \phi^*(x)$.*

Theorem 1 guarantees sufficient information for endogenous-state recovery under its assumptions, but does not require exclusion of exogenous information. Preserving the full observation can also permit optimal classification. The guarantee assumes finite spaces; our continuous-control experiments evaluate the objective empirically. We constrain representation capacity through dimensional bottlenecks of 256 in DMC and 64 in LIBERO, and separately evaluate endogenous decodability and exogenous sensitivity. We evaluate both frozen-encoder offline RL and joint optimization with behavior cloning;

only the latter additionally updates the encoder using action labels. Formal assumptions and proofs are provided in Section E.

5 EXPERIMENTS

Through our experiments we aim to evaluate the general capabilities of `CARDPol` in action-free settings. We evaluate downstream control (Section 5.1), endogenous-state decoding (Section 5.2), and sensitivity to exogenous variation (Section 5.3). DMC uses frozen encoders with offline RL, while LIBERO jointly optimizes representation learning and target-task behavior cloning. Appendix D.1 reports additional frozen-encoder control results on OGBench.

Baselines. We compare `CARDPol` with reconstruction-based objectives (VAE; Kingma & Welling (2013), VQVAE; van den Oord et al. (2018)), contrastive learning (CURL; Srinivas et al. (2020)), and value-based objectives (VIP Ma et al. (2022) and ICVF Ghosh et al. (2023)). These methods use the same visual data but do not use policy labels. To distinguish the contribution of the objective from that of policy-label supervision, we also evaluate PolCon, a supervised contrastive baseline that treats observations from the same policy as positives and observations from different policies as negatives. Its objective is detailed in Appendix A.4.

5.1 DOWNSTREAM CONTROL

We demonstrate downstream control in a locomotion domain, the DeepMind Control Suite (DMC; Tassa et al. (2018)), and a manipulation domain, Libero (Liu et al., 2023).

5.1.1 FROZEN REPRESENTATIONS FOR OFFLINE RL

Dataset Creation. Due to the incompleteness or unavailability of data in related works (Islam et al., 2023; Lu et al., 2023) we construct our own datasets of action-free trajectories for representation pre-training that closely mimic prior works. For each DMC domain, we first train state-based policies using soft actor-critic (SAC) (Haarnoja et al., 2018) and save five checkpoints spanning performance levels from near-random to expert; specific checkpoint performance can be found in Table 3. We roll out each checkpoint and render the resulting trajectories under four distraction conditions: 1) *Clean* uses the default rendering; 2) *Static* holds a sampled background image from a subset of DAVIS (Pont-Tuset et al., 2018) dataset fixed throughout an episode; 3) *Video* plays a sampled background video from a subset of the DAVIS dataset; and 4) *Hard Video* plays a sampled background video from a larger subset of the DAVIS dataset. The distractions are sampled independently. Encoder pretraining uses the rendered observations with actions and rewards withheld. For `CARDPol` , each trajectory is additionally labeled by the checkpoint that generated it. For these results, the pre-training dataset in each domain contains a total of 1.25M transitions by combining 5 individual checkpoint datasets of 250k each. We report results for different dataset compositions in Appendix C.2

Experimental Methodology. We follow the experimental protocol of Islam et al. (2023), where applicable. We first pre-train the representation on the action-free visual dataset and then freeze the learned encoder for downstream policy training. For all methods, we use an encoder network based on the DrQ+BC (Lu et al., 2023) architecture, followed by a linear projection to a 256-dimensional representation; this 256-dimensional embedding constitutes the information bottleneck of our model. After pretraining, we freeze the encoder and train DrQ+BC on 250k expert transitions per domain with action and reward labels. We choose to use 250k expert transitions because prior work (Lu et al., 2023; Islam et al., 2023) has shown that between 100k and 500k expert transitions demonstrate strong performance. These expert observations are also included in the 1.25M-transition pretraining dataset, where their action and reward labels are withheld. All methods use the same datasets and downstream training budget, providing a controlled comparison of representation objectives under this protocol.

Results. Table 1 shows that `CARDPol` achieves the highest mean return in every evaluated condition. Its advantage varies across settings: CURL performs similarly in static Cheetah, and ICVF performs similarly in clean Walker, whereas `CARDPol` has a larger advantage in clean Cheetah and several distracted settings. Across the Video and Hard-video conditions, `CARDPol` achieves 25–53% higher mean return than the strongest evaluated baseline in each condition. These results support the use of policy-based supervision for learning representations that remain useful for control under

Table 1: Evaluation performance of an agent trained with an offline RL algorithm using frozen pre-trained encoders. The best result in each row and any result whose ± standard-error (over 4 seeds) interval overlaps it are bold.

Env.	Exo.	CARDPol	PolCon	VIP	ICVF	VAE	VQVAE	CURL
CHEETAH	Clean	**523±5**	402±40	411±28	412±13	364±27	131±10	337±17
	Static	**614±4**	425±15	462±5	510±12	201±11	271±19	**602±11**
	Video	**449±16**	336±26	257±15	301±9	174±9	133±4	357±8
	Hard-Video	**293±11**	196±13	117±7	218±16	88±0	82±2	150±6
WALKER	Clean	**885±17**	726±21	313±18	**883±6**	140±11	148±6	647±34
	Static	**649±33**	384±44	139±3	254±17	72±3	76±2	132±7
	Video	**578±8**	378±9	134±7	129±3	101±4	95±2	102±3
	Hard-Video	**443±5**	354±18	86±5	123±8	81±6	72±1	74±2

Table 2: LIBERO success rates (%), averaged over ten tasks and five seeds per task. Errors denote standard errors across tasks; bold marks the highest mean in each row.

Suite	Setting	No Rep.	CARDPol	VIP	VAE	VQ-VAE	CURL	ICVF
Goal	Video	69.6±9.1	**77.4**±5.3	66.4±8.8	50.8±7.1	57.4±7.4	65.8±7.3	71.8±4.4
	Distract	57.8±8.4	**69.2**±7.6	59.6±8.2	52.0±7.1	52.2±6.2	3.0±1.5	62.0±7.9
Object	Video	68.2±8.0	71.2±6.9	65.1±5.5	49.2±4.6	51.8±6.2	**73.2**±5.8	67.6±7.0
	Distract	51.4±3.7	**64.6**±3.0	54.1±4.9	49.6±4.7	46.6±4.7	0.8±0.3	54.0±4.8
Spatial	Video	50.8±9.0	52.6±9.6	**55.4**±9.9	34.8±5.5	33.4±5.8	52.8±8.9	49.2±7.7
	Distract	46.8±9.6	**51.6**±7.7	45.6±8.0	35.4±6.9	33.8±9.1	4.4±1.5	48.8±8.8

dynamic visual distractions. CARDPol uses policy labels to identify features relevant to the observed agents' decision-making, which helps separate controllable structure from distracting visual dynamics. CARDPol also achieves higher mean return than PolCon in all eight conditions. Because PolCon uses the same policy labels, encoder, pretraining data, and optimization settings, this comparison supports the benefit of CARDPol's objective beyond access to policy-label supervision alone in this setting.

5.1.2 JOINT REPRESENTATION LEARNING AND BEHAVIOR CLONING

We next evaluate whether CARDPol improves policy learning when optimized jointly with behavior cloning, using action-free demonstrations from other tasks as additional representation-learning data.

Datasets. We evaluate on three task suites from the LIBERO benchmark: LIBERO-Spatial, LIBERO-Object, and LIBERO-Goal, each containing 10 simulated robot manipulation tasks. The suites emphasize different forms of generalization: Spatial varies spatial relationships, Object varies the manipulated objects, and Goal varies the desired manipulation goals. For our experiments, we group demonstrations by task and use task identity as a proxy for policy identity. We use the RGB observations from LIBERO's `agentview_image`. For the **Clean** setting, we use the nominal LIBERO observation. For the **Video** setting, we play a random video from the EPIC-KITCHENS (Damen et al., 2020) dataset beside (i.e., horizontally concatenated to) the main visual observation.

Experimental Methodology. We use the ViLT policy from the LIBERO benchmark suite (Liu et al., 2023) as our backbone: a per-timestep spatial transformer over image patches followed by a causal temporal transformer, mapping stacked camera observations to actions. We treat the spatial transformer's per-timestep output (i.e., its representation prior to temporal aggregation) as our representation $\phi(x_t)$; in these experiments, we use the default embedding size of 64. Each training step draws two batches from the same demonstration pool: a focused batch, restricted to a

target task, trains the standard behavior-cloning loss $\mathcal{L}_{\text{BC}}$; a mixed batch, drawn across the full task distribution, supplies a representation loss $\mathcal{L}_{\text{rep}}$ computed on the 64-dimensional output of spatial representations $\phi(x_t) \in \mathbb{R}^{64}$ produced by the same spatial transformer. The total training objective is $\mathcal{L} = \mathcal{L}_{\text{BC}} + \lambda \mathcal{L}_{\text{rep}}$.

The two losses are backpropagated through the network. Gradients from the behavior-cloning loss $\mathcal{L}_{\text{BC}}$ flow into the transformer backbone and the action head. The representation loss $\mathcal{L}_{\text{rep}}$ updates the spatial encoder and its associated auxiliary head, but does not update the behavior-cloning action head. For each dataset, we train separate BC policies focused on each task where action labels are only available for that one task. This is done across each task in each dataset. For each task and training seed, we evaluate the trained policy over 50 rollouts. Table 2 reports success rates averaged across the ten tasks in each suite, with standard errors computed across the tasks. We use five training seeds per task, giving 2,100 training runs across seven methods and six suite variants. We selected λ per method from a compute-limited logarithmic grid over $[0.001, 1.0]$, choosing the highest success on distracted tasks among settings with visibly decreasing auxiliary loss. Zero weight corresponds to the separately reported No Rep. baseline. See Appendix A.2 for training settings.

Results. In the **Video** conditions, adding `CARDPol` improves mean success over the No Rep. baseline by 11.4, 13.2, and 4.8 percentage points on Goal, Object, and Spatial, respectively. It also achieves the highest mean success among the representation objectives, exceeding the strongest alternative by 7.2, 10.5, and 2.8 percentage points. In the **Clean** conditions, performance remains competitive with the strongest baselines. These results support using `CARDPol` as an auxiliary objective during action-supervised policy learning.

5.2 Endogenous Information Recovery

5.2.1 State Decoding in DMC

We test how well `CARDPol` extracts the controllable, proprioceptive state from visual observations. Following Nikulin et al. (2025), we freeze each learned encoder and train a state decoder, implemented as a 2-layer MLP, to predict the normalized ground-truth proprioceptive state from its corresponding observation's embedding. In Figure 2a, we report the mean squared error on held-out validation images, where an error of 0 indicates perfect decoding. Across distraction settings, `CARDPol` (in green) consistently achieves among the lowest errors. While most baselines achieve similarly low decoding validation loss in the `clean` setting, their decoding performance degrades as the distractions become more dynamic. These results indicate that endogenous information remains more readily decodable from `CARDPol` representations under dynamic distractions using the chosen probe. They do not determine whether the representations also retain exogenous information, which we evaluate separately in Section 5.3.1.

5.2.2 Human-Pose Decoding in EgoDex

We next test whether the objective remains useful on human egocentric video, where task annotations provide policy-class proxies rather than identities of fixed data-collection policies. We apply `CARDPol` as a representation learning objective on the EgoDex dataset (Hoque et al., 2026), a large-scale video dataset of 829 hours of real-world human manipulation with ground truth 3D hand and finger tracking. The dataset includes 338k trajectories covering 194 language-labeled tasks. Our experiments use a subset covering 111 task classes, with task annotations supplying policy-class labels. Appendix A.3 describes the subset and data partitions. We train ResNet18 (He et al., 2015) encoders using `CARDPol` and baselines. After representation pretraining, we freeze each encoder and train an MLP pose decoder on a separate labeled partition. We evaluate this decoder on the observations used for encoder pretraining, whose pose labels were not used to train either the encoder or the decoder. Thus, evaluation observations are unseen by the decoder but have been observed during encoder pretraining. In Figure 3c, we report mean squared error for predicting 3D wrist positions in the camera frame of the action-free training data. `CARDPol` achieves the lowest mean pose-decoding error among the evaluated methods, with VAE providing the closest baseline. These results support the application of `CARDPol` to human egocentric video containing natural scene variation and camera motion. The experiment does not isolate the contribution of camera motion to the differences between methods.

(a) Mean squared error of state decoding on validation set of unseen observations.

(b) Exogenous-variance ratio under matched video distractions.

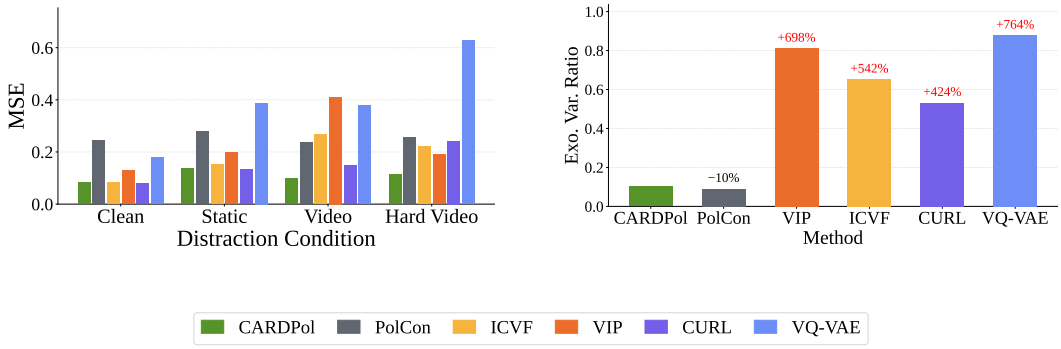


Figure 2: Endogenous-state decoding and sensitivity to exogenous distractions in DMC. Percentages indicate changes relative to `CARDPol`.

5.3 Sensitivity to Exogenous Variation

5.3.1 Matched-Distractor Representation Analysis

Accurate state decoding does not establish invariance to distractions. A representation may retain both endogenous and exogenous information. We therefore measure representation changes across visual distractors while holding the endogenous trajectory fixed.

Experimental Methodology. We use matched trajectories in which the same endogenous state sequence is rendered under multiple video distractors. Let $\tilde{z}_{d,t}$ denote the standardized embedding at timestep t under distractor d, where each embedding dimension is standardized over all distractors and timesteps. We measure the ratio of distractor-induced variance to within-trajectory temporal variance as

$$\text{ExoVarRatio} = \frac{\frac{1}{T}\sum_{t=1}^{T} \text{Tr}(\text{Cov}_d[\tilde{z}_{d,t}])}{\frac{1}{D}\sum_{d=1}^{D} \text{Tr}(\text{Cov}_t[\tilde{z}_{d,t}])}. \tag{2}$$

The numerator measures representation variance across distractors while holding the endogenous state fixed, while the denominator measures the overall temporal variance of the representation along the trajectory. A lower ratio indicates less variation across distractors relative to temporal variation within trajectories. Because the numerator and denominator measure different sources of variation, this ratio is not necessarily bounded by one. Additional details are provided in Appendix B.1

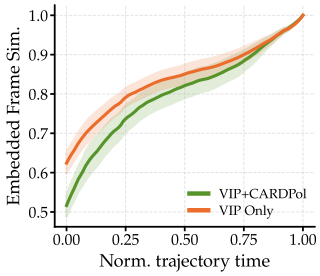
Results. In Figure 2b, we plot the variance defined in eq. (2), aggregated across the `Hard` datasets for Cheetah and Walker. Figure 2b shows that `CARDPol` and PolCon have substantially lower exogenous-variance ratios than the other evaluated methods, with PolCon achieving a slightly lower ratio than `CARDPol`. However, `CARDPol` yields lower state-decoding error than PolCon in every distraction condition (fig. 2a) and higher downstream returns (table 1). Together, these results show the importance of evaluating distractor sensitivity alongside retention of control-relevant information. To assess sensitivity to the dimensional bottleneck, we vary the embedding dimension from 8 to 2048 in Appendix C.3 and show that `CARDPol` maintains the lowest exogenous-variance ratio and state-decoding error throughout this range among the tested methods, suggesting that its advantage is not confined to the default representation size.

5.3.2 Goal-Similarity Analysis with VIP and CARDPol

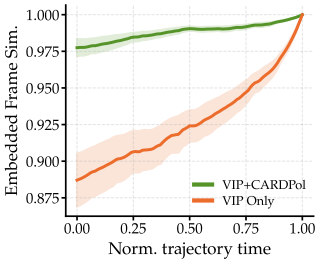
Finally, we test whether `CARDPol` can reduce distractor sensitivity when added to an existing representation-learning objective.

Experimental Methodology. `CARDPol` can be combined with existing representation-learning methods to encourage invariance to exogenous noise. To demonstrate this, we train two VIP

(a) Changes in goal similarity due to only robot movement.



(b) Changes in goal similarity due to only exogenous movement.



(c) Mean squared error from decoding representations to human pose on EgoDex dataset.

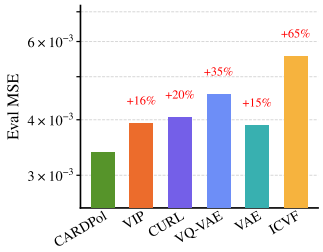


Figure 3: Goal-similarity responses to robot motion (fig. 3a) and distractor motion (fig. 3b), and human-pose decoding error on EgoDex (fig. 3c). Percentages indicate changes relative to `CARDPol`.

encoders on the same distracted LIBERO dataset: one with the standard VIP objective and one with `CARDPol` added as an auxiliary loss. Again, VIP is an action-free representation-learning method that learns an embedding space in which cosine similarity is correlated with temporal distance. Training settings and controlled evaluation procedures are provided in Appendix B.2.

Dataset. We use the LIBERO-Spatial dataset, but each trajectory is rendered with dynamic visual distractions, and the resulting distracted, action-free trajectories are used for encoder training. Since `CARDPol` requires policy labels, we use the language task label of each trajectory as a proxy for its policy class.

Results. In Figure 3, we plot the cosine similarity between each observation embedding and the embedding of the goal image, aggregated over 10 tasks with 5 trajectories per task. Both panels use encoders trained on the same distracted LIBERO data, but modify the evaluation trajectories to isolate different sources of visual change. In fig. 3a, we hold the distractor region fixed while the original demonstration evolves. VIP with (green) and without (orange) the `CARDPol` auxiliary loss produces similar goal-similarity curves, suggesting that adding `CARDPol` does not interfere with VIP's ability to represent task-relevant robot motion. In fig. 3b, we instead keep the robot static while allowing only the distraction to move. Standard VIP assigns importance to the dynamic but task-irrelevant visual features, causing the predicted goal-similarity to change throughout the trajectory. In contrast, VIP trained with `CARDPol` as an auxiliary loss produces a largely unchanged goal-similarity estimate, indicating reduced sensitivity of the goal-similarity readout to the exogenous distraction.

6 Limitations and Conclusion

Our results show that policy identity provides useful supervision for action-free representation learning, and that the way this supervision is used matters. `CARDPol` improves downstream control over the evaluated policy-contrastive baseline, preserves decodable endogenous information, and exhibits reduced sensitivity to visual distractions. Its applications to joint behavior cloning, human video, and VIP pretraining demonstrate its use across different training settings. `CARDPol` requires informative trajectory partitions and can exploit visual shortcuts when exogenous context correlates with policy identity. Task annotations provide useful empirical supervision in our LIBERO and EgoDex experiments, but need not satisfy the theoretical assumptions. Our human-video evidence concerns pose decodability from annotated recordings. Robustness to policy-correlated visual context and methods for correcting such correlations remain untested.

7 AI Use Statement

We used AI tools to assist with manuscript organization and plot styling. The authors reviewed the resulting suggestions and verified quantitative statements against the experimental outputs. The code

used to run the experiments, analyze the results, and for baselines was all written without usage of AI. The ideation, math, and theory for this project was also done without the usage of AI tools.

8 Ethics statement

In this paper we introduce a novel objective for training representations from realistic data. The work has applications to robotics and learning from human data. Broadly, these fields can have both positive and negative impacts on society, but our method does not directly connect to one particular downstream externality.

9 Reproducibility

We include code with our submission for full transparency. We describe all theoretical assumptions required to re-derive the theory presented in this paper. We include all non-obvious hyperparameters in the appendix required to reproduce our experiments.

References

Bowen Baker, Ilge Akkaya, Peter Zhokhov, Joost Huizinga, Jie Tang, Adrien Ecoffet, Brandon Houghton, Raul Sampedro, and Jeff Clune. Video pretraining (vpt): Learning to act by watching unlabeled online videos, 2022. URL https://arxiv.org/abs/2206.11795.

Zichen Cui, Hengkai Pan, Aadhithya Iyer, Siddhant Haldar, and Lerrel Pinto. Dynamo: In-domain dynamics pretraining for visuo-motor control. *Advances in Neural Information Processing Systems*, 37:33933–33961, 2024.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. The epic-kitchens dataset: Collection, challenges and baselines, 2020. URL https://arxiv.org/abs/2005.00343.

Yonathan Efroni, Dipendra Misra, Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Provably filtering exogenous distractors using multistep inverse dynamics. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=RQLLzMCefQu.

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4rl: Datasets for deep data-driven reinforcement learning, 2021. URL https://arxiv.org/abs/2004.07219.

Quentin Garrido, Tushar Nagarajan, Basile Terver, Nicolas Ballas, Yann LeCun, and Michael Rabbat. Learning latent action world models in the wild, 2026. URL https://arxiv.org/abs/2601.05230.

Dibya Ghosh, Chethan Anand Bhateja, and Sergey Levine. Reinforcement learning from passive data via latent intentions. In *International Conference on Machine Learning*, pp. 11321–11339. PMLR, 2023.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina Gonzalez, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jachym Kolar, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbelaez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna

Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the world in 3,000 hours of egocentric video, 2022. URL https://arxiv.org/abs/2110.07058.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor, 2018. URL https://arxiv.org/abs/1801.01290.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. URL https://arxiv.org/abs/1512.03385.

Ryan Hoque, Peide Huang, David J. Yoon, Mouli Sivapurapu, and Jian Zhang. Egodex: Learning dexterous manipulation from large-scale egocentric video, 2026. URL https://arxiv.org/abs/2505.11709.

Riashat Islam, Manan Tomar, Alex Lamb, Yonathan Efroni, Hongyu Zang, Aniket Rajiv Didolkar, Dipendra Misra, Xin Li, Harm Van Seijen, Remi Tachet Des Combes, et al. Principled offline rl in the presence of rich exogenous information. In *International Conference on Machine Learning*, pp. 14390–14421. PMLR, 2023.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning, 2021. URL https://arxiv.org/abs/2004.11362.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2013. URL https://arxiv.org/abs/1312.6114.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit q-learning, 2021. URL https://arxiv.org/abs/2110.06169.

Alex Lamb, Riashat Islam, Yonathan Efroni, Aniket Rajiv Didolkar, Dipendra Misra, Dylan J Foster, Lekan P Molu, Rajan Chari, Akshay Krishnamurthy, and John Langford. Guaranteed discovery of control-endogenous latent states with multi-step inverse models. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=TNocbXm5MZ.

Alexander Levine, Peter Stone, and Amy Zhang. Multistep inverse is not all you need. *Reinforcement Learning Journal*, 2:884–925, 2024.

Alexander Levine, Peter Stone, and Amy Zhang. Offline action-free learning of ex-bmdps by comparing diverse datasets. In *Reinforcement Learning Conference*, 2025.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning, 2023. URL https://arxiv.org/abs/2306.03310. MIT License.

Cong Lu, Philip J Ball, Tim GJ Rudner, Jack Parker-Holder, Michael A Osborne, and Yee Whye Teh. Challenges and opportunities in offline reinforcement learning from visual observations. *Transactions on Machine Learning Research*, 2023.

Yecheng Jason Ma, Shagun Sodhani, Dinesh Jayaraman, Osbert Bastani, Vikash Kumar, and Amy Zhang. Vip: Towards universal visual reward and representation via value-implicit pre-training. *arXiv preprint arXiv:2210.00030*, 2022.

Yecheng Jason Ma, Vikash Kumar, Amy Zhang, Osbert Bastani, and Dinesh Jayaraman. LIV: Language-image representations and rewards for robotic control. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 23301–23320. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/ma23b.html.

Arjun Majumdar, Karmesh Yadav, Sergio Arnaud, Yecheng Jason Ma, Claire Chen, Sneha Silwal, Aryan Jain, Vincent-Pierre Berges, Pieter Abbeel, Jitendra Malik, Dhruv Batra, Yixin Lin, Oleksandr Maksymets, Aravind Rajeswaran, and Franziska Meier. Where are we in the search for an artificial visual cortex for embodied intelligence? 2023.

Ajay Mandlekar, Danfei Xu, Josiah Wong, Soroush Nasiriany, Chen Wang, Rohun Kulkarni, Li Fei-Fei, Silvio Savarese, Yuke Zhu, and Roberto Martín-Martín. What matters in learning from offline human demonstrations for robot manipulation. In *arXiv preprint arXiv:2108.03298*, 2021.

Dipendra Misra, Akanksha Saran, Tengyang Xie, Alex Lamb, and John Langford. Towards principled representation learning from videos for reinforcement learning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=3mnWvUZIXt.

Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3m: A universal visual representation for robot manipulation. In *Conference on Robot Learning*, pp. 892–909. PMLR, 2023.

Alexander Nikulin, Ilya Zisman, Denis Tarasov, Nikita Lyubaykin, Andrei Polubarov, Igor Kiselev, and Vladislav Kurenkov. Latent action learning requires supervision in the presence of distractors, 2025. URL https://arxiv.org/abs/2502.00379.

Seohong Park, Kevin Frans, Benjamin Eysenbach, and Sergey Levine. Ogbench: Benchmarking offline goal-conditioned rl. In *International Conference on Learning Representations (ICLR)*, 2025. MIT License.

Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation, 2018. URL https://arxiv.org/abs/1704.00675. Creative Commons Attribution 4.0 (CC BY 4.0).

Karl Schmeckpeper, Oleh Rybkin, Kostas Daniilidis, Sergey Levine, and Chelsea Finn. Reinforcement learning with videos: Combining offline observations with interaction. In Jens Kober, Fabio Ramos, and Claire Tomlin (eds.), *Proceedings of the 2020 Conference on Robot Learning*, volume 155 of *Proceedings of Machine Learning Research*, pp. 339–354. PMLR, 16–18 Nov 2021. URL https://proceedings.mlr.press/v155/schmeckpeper21a.html.

Dominik Schmidt and Minqi Jiang. Learning to act without actions. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=rvUq3cxpDF.

Younggyo Seo, Kimin Lee, Stephen L James, and Pieter Abbeel. Reinforcement learning with action-free pre-training from videos. In *International Conference on Machine Learning*, pp. 19561–19579. PMLR, 2022.

Aravind Srinivas, Michael Laskin, and Pieter Abbeel. Curl: Contrastive unsupervised representations for reinforcement learning, 2020. URL https://arxiv.org/abs/2004.04136.

Austin Stone, Oscar Ramirez, Kurt Konolige, and Rico Jonschkowski. The distracting control suite–a challenging benchmark for reinforcement learning from pixels. *arXiv preprint arXiv:2101.02722*, 2021. Licensed under the Apache License, Version 2.0.

Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, et al. Deepmind control suite. *arXiv preprint arXiv:1801.00690*, 2018. Apache License 2.0.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning, 2018. URL https://arxiv.org/abs/1711.00937.

Denis Yarats, David Brandfonbrener, Hao Liu, Michael Laskin, Pieter Abbeel, Alessandro Lazaric, and Lerrel Pinto. Don't change the algorithm, change the data: Exploratory data for offline reinforcement learning. In *ICLR 2022 Workshop on Generalizable Policy Learning in Physical World*, 2022.

Seonghyeon Ye, Joel Jang, Byeongguk Jeon, Sejune Joo, Jianwei Yang, Baolin Peng, Ajay Mandlekar, Reuben Tan, Yu-Wei Chao, Bill Yuchen Lin, Lars Liden, Kimin Lee, Jianfeng Gao, Luke Zettlemoyer, Dieter Fox, and Minjoon Seo. Latent action pretraining from videos, 2025. URL https://arxiv.org/abs/2410.11758.

Arthur Zhang, Chaitanya Eranki, Christina Zhang, Ji-Hwan Park, Raymond Hong, Pranav Kalyani, Lochana Kalyanaraman, Arsh Gamare, Arnav Bagad, Maria Esteva, and Joydeep Biswas. Toward robust robot 3-d perception in urban environments: The ut campus object dataset. *IEEE Transactions on Robotics*, 40:3322–3340, 2024. doi: 10.1109/TRO.2024.3400831.

Qinqing Zheng, Mikael Henaff, Brandon Amos, and Aditya Grover. Semi-supervised offline reinforcement learning with action-free trajectories. In *International conference on machine learning*, pp. 42339–42362. PMLR, 2023.



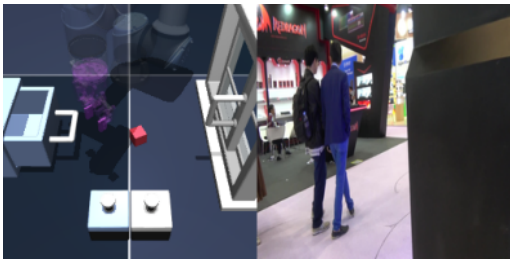
(a) LIBERO Environment



(b) Distracting Control Suite Cheetah



(c) Distracting Control Suite Walker



(d) OGBench Scene Environment



(e) EgoDex Data

Figure 4: Visualization of distracted environments.

Table 3: Checkpoint returns and membership in the three-, five-, and seven-policy pretraining datasets. Bullets indicate included checkpoints.

Checkpoint	Cheetah Run Return	3	5	7	Walker Walk Return	3	5	7
1	4±1		•	•	30±5		•	•
2	195±72		•	•	409±95		•	•
3	542±19	•	•	•	733±30			•
4	577±22			•	810±40	•	•	•
5	634±44	•	•	•	838±21			•
6	699±15			•	885±58	•	•	•
7	804±16†	•	•	•	968±12†	•	•	•

A Experimental Setup

In this section, we discuss experimental setups for DMC, LIBERO, and EgoDex as well as formally introduce the Policy-Contrastive Baseline, PolCon.

A.1 DeepMind Control Suite

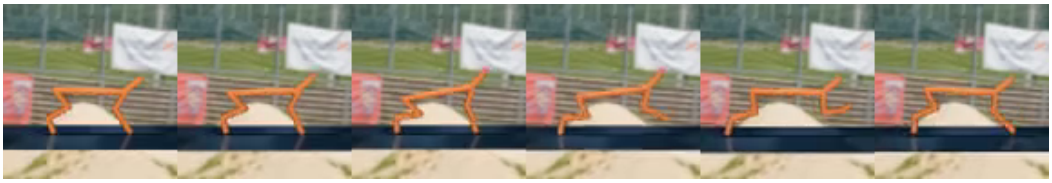
Trajectory collection. We collect trajectories in the Cheetah-run and Walker-walk tasks using state-based SAC policies saved at different stages of training. Each checkpoint contributes 250k transitions. Table 3 reports checkpoint performance and identifies the subsets used to construct the three-, five-, and seven-policy datasets, containing 750k, 1.25M, and 1.75M transitions, respectively. The five-policy configuration is used for the main results. Checkpoint identity provides the policy label for `CARDPol`; actions and rewards are withheld during representation pretraining. Downstream offline RL uses 250k expert transitions with action and reward labels. These transitions are the expert trajectories also included in pretraining.

Visual observations and distractions. Observations are rendered as 84×84 RGB images and processed into stacks of three consecutive frames, with an action repeat of two. We evaluate four

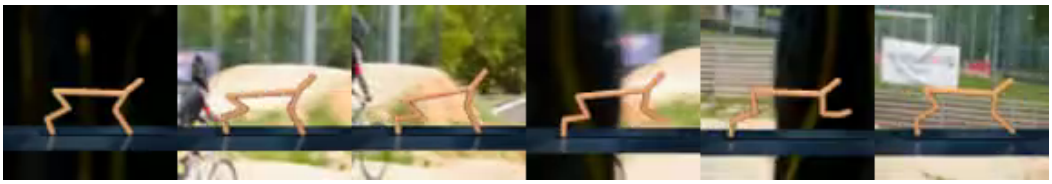
visual conditions: Clean uses the nominal environment rendering; Static samples a background image and holds it fixed throughout an episode; Video plays a sampled background video; and Hard-video plays a sampled background video from a larger candidate set. Backgrounds are drawn from the Easy and Hard DAVIS partitions described by Stone et al. (2021). Distractors are sampled uniformly and independantly at the start of each episode. Evaluation uses held-out videos defined by Stone et al. (2021). See examples in Figure 5.



(a) `Clean` Environment



(b) `Static`: the background remains fixed within each episode.



(c) `video`: the background changes throughout the episode. The `video` setting uses a subset of the videos in the DAVIS (Pont-Tuset et al., 2018) dataset as background distractions.

Figure 5: Example visual observations in the Cheetah domain

A.1.1 DMC Hyperparameters

Tables 5 and 4 report downstream RL and representation-pretraining settings, respectively.

A.2 Libero

We use the LIBERO-Spatial, LIBERO-Object, and LIBERO-Goal suites, each containing ten manipulation tasks. For each task, we use 50 demonstrations, partitioned into training and validation sets at the trajectory level using a $0.9/0.1$ train/val split. Each target-task policy receives action labels only from that task. The representation objective additionally uses observations from all ten tasks, including the target task, with task identity serving as the policy-class label for `CARDPol`. Mixed-task batches are sampled uniformly across tasks.

We use agent-view RGB observations, resized to $128 \times 128 pixels$. The Clean condition uses the original observation. In the Video condition, an EPIC-KITCHENS clip is horizontally concatenated with the agent-view image (making the observation 128×256), after appropriate resizing. Clips are sampled from the `P01` partition independently of task identity. Training and evaluation distractors both use `P01` videos. The policy receives no additional information.

A.2.1 LIBERO Hyperparameters

Tables 6 and 7 report LIBERO training settings and method-specific hyperparameters.

Table 4: Representation pretraining hyperparameters. Every method uses the same encoder architecture, data, and optimization settings; only the objective differs.

Hyperparameter	Value
Shared by all methods	
Encoder	DrQ CNN: 4 conv layers (32 ch, 3×3, first stride 2) + linear
Representation dim.	256
Input	3 stacked RGB frames, 84×84, scaled to $[-1, 1]$
Pretraining data	5 policies × 500 episodes (+ 50 val. episodes each), no actions or rewards
Optimizer	Adam (β=$(0.9, 0.999)$), weight decay 10^{-4}
Learning rate	10^{-4}
Batch size	512
Epochs	4
Method-specific	
`CARDPol`	policy-ID classifier on pairs of frame-stacks $\leq K$=10 steps apart; MLP 2×512; LayerNorm + tanh output
VIP	γ=0.98
ICVF	γ=0.99, expectile 0.9, shared encoder
CURL	InfoNCE with bilinear similarity; positives = two augmentations of the same frame-stack
VQ-VAE	codebook 512 × 256, 16 latents, commitment 0.25, EMA decay 0.99; CNN decoder
AE	CNN decoder, pixel reconstruction

Table 5: Downstream offline RL (DrQ+BC) hyperparameters, shared by every pretrained encoder.

Hyperparameter	Value
Algorithm	DrQ+BC (V-D4RL), frozen pretrained encoder
Offline dataset	500 episodes (250k transitions) from one policy
Training steps	500k (1 gradient update every 2 steps)
Batch size	256
Optimizer / learning rate	Adam, 3×10^{-4} (actor and critic)
Discount γ	0.99
n-step return	3
Critic target update τ	0.01
Actor/critic trunk	linear 256 + LayerNorm + tanh
Actor/critic MLP	2 hidden layers of 1024, twin Q
Exploration std. schedule	linear $1.0 \rightarrow 0.1$ over 250k steps; clip 0.3
BC weight α	2.5 (TD3+BC normalization $\lambda = \alpha/\lvert\overline{Q}\rvert$)
Image augmentation	random shift ($\pm$4 px)
Frame stack / action repeat	3 / 2
Evaluation	every 10k steps, 10 episodes, deterministic policy
Reported metric	mean return over the last 5 evaluations; mean $\pm$ s.e. over 4 seeds

A.3 EGODEX

We use all the trajectories from EgoDex training dataset, covering 111 task classes. Task annotations define the policy classes used by `CARDPol`. The encoder-pretraining partition uses 334k trajectories, and the separate pose-decoder training partition contains 3.2k. Partitions are constructed using EgoDex's own `train` and `test` groups. Pose labels are withheld during encoder pretraining. After freezing the encoder, we train the pose decoder on the separate labeled partition and evaluate it on all the encoder-pretraining observations. **Their pose labels are used only for evaluation.** The decoder predicts 3D wrist positions in the camera frame, which is included in the dataset.

A.4 POLICY-CONTRASTIVE BASELINE

To test whether `CARDPol`'s policy-identification objective can be replaced by a contrastive one that uses the same supervision, we train a policy contrastive baseline that uses exactly the same inputs as

Table 6: Training settings shared by all methods. Every method uses the same ViLT backbone and training recipe; only the auxiliary objective differs.

Setting	Value
Backbone	ViLT (embed size 128; spatial token downsampled to 64)
Observations	Agent-view RGB, 128×128 (no proprioception, wrist camera, or language)
Suites	LIBERO-Goal, LIBERO-Object, LIBERO-Spatial
Tasks / seeds	10 tasks $\times$ 5 seeds per suite
Sequence length	10
Epochs	50
Batch size	64
Optimizer	AdamW ($\beta = (0.9, 0.999)$)
Learning rate	1×10^{-4}, cosine annealing
Weight decay	1×10^{-4}
Gradient clipping	100
Data augmentation	Enabled
Train split	0.9

Table 7: Method-specific hyperparameters. λ is the weight on the auxiliary representation loss ($\mathcal{L} = \mathcal{L}_{\text{BC}} + \lambda \mathcal{L}_{\text{rep}}$). All auxiliary heads take the same 64-dimensional pooled visual representation as input.

	CARDPol	VIP	VAE	VQ-VAE	CURL	ICVF
Rep. loss weight λ	0.01	0.01	1.0	1.0	1.0	0.01
Embedding / latent dim	64	64	64	64	64	64
Inputs per sample	(o_t, o_{t+k})	(o_0, g, o_t, o_{t+1})	o_t	o_t	2 aug. views of o_t	(s, s', s_+, z)
Method-specific						
Future offset k	$[1, 10]$	–	–	–	–	–
Discount γ	–	0.98	–	–	–	0.99
Temperature τ	–	–	–	–	0.1	–
Codebook size	–	–	–	512	–	–
Commitment cost	–	–	–	0.25	–	–
KL weight	–	–	1.0	–	–	–
L1 / L2 emb. reg.	–	0 / 0	–	–	–	–
Expectile	–	–	–	–	–	0.9
Target EMA rate	–	–	–	–	–	0.005
Goal sampling	–	–	–	–	–	0.3 / 0.5 / 0.2
Same-goal prob.	–	–	–	–	–	0.5

CARDPol : action- and reward-free trajectories, each labeled only with the index $c \in \{1, \ldots, N\}$ of the data-collection policy that generated it. It also share CARDPol 's encoder ϕ, frame-stack input o_t (three stacked RGB frames), pretraining data, image augmentation, and optimization settings (Table 4). Only the pretraining loss differs. The loss operates on ℓ_2-normalized embeddings $z = \phi(o)/\|\phi(o)\|_2$ with no projection head, uses temperature $\tau = 0.1$, and the unnormalized encoder output $\phi(o)$ is the representation used downstream.

Trajectory-agnostic policy contrastive. This baseline treats the policy label as a class and applies the supervised contrastive loss of Khosla et al. (2021). Each batch element is a single frame-stack o^i with policy label c^i, sampled independently of its position in the trajectory. Every other frame-stack in the batch from the same policy is a positive; every frame-stack from a different policy is a negative. For an anchor i with positive set $P(i) = \{p \neq i : c^p = c^i\}$, the loss is

$$\mathcal{L}_{\text{agn}} = -\frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{V}} \frac{1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp\left(z^{i\top} z^p / \tau\right)}{\sum_{a \neq i} \exp(z^{i\top} z^a / \tau)}, \tag{3}$$

where $\mathcal{V}$ is the set of anchors with at least one positive in the batch. The objective clusters all observations produced by the same policy, regardless of the trajectory or time step they come from.

This method does not explicitly select temporally related positive pairs; its positive set is determined solely by policy identity.

B Representation Evaluation Protocols

B.1 Matched-Distractor Representation Analysis

Matched evaluation observations. We select 100 trajectories from the expert datasets in each of Cheetah-run and Walker-walk. For each trajectory, we construct $D = 60$ observation sequences with identical endogenous states and different video distractors, using re-rendering saved simulator states. Each sequence contains $T = 500$ timesteps. Distractors are sampled from the Hard-Video setting from DAVIS. This construction varies the exogenous visual content while preserving the underlying agent motion.

Embedding extraction. We embed an observation $x_{d,t}$ which denotes the timestep t in the trajectory with distraction d:

$$z_{d,t} = \phi_\theta(x_{d,t}) \in \mathbb{R}^{d_z}, \qquad d = 1, \dots, D,\ t = 1, \dots, T. \tag{4}$$

Standardization. Each embedding dimension is z-scored using the mean and standard deviation pooled over all $D \times T$ observations of each matched trajectory.:

$$\tilde{z}_{d,t} = \frac{z_{d,t} - \mu}{\sigma}, \qquad \mu, \sigma \in \mathbb{R}^{d_z} \text{ computed over all } (d, t). \tag{5}$$

This puts all dimensions on comparable scale before computing covariance traces, so the result isn't dominated by a few high-magnitude dimensions.

Temporal (total) variance. For each distractor instance d, we compute the covariance of the embedding across time and take its trace:

$$\text{tr-cov}_{\text{temporal}}(d) = \operatorname{tr}\big(\operatorname{Cov}_t(\tilde{z}_{d,1}, \dots, \tilde{z}_{d,T})\big), \qquad \overline{\text{tr-cov}}_{\text{temporal}} = \frac{1}{D}\sum_{d=1}^{D} \text{tr-cov}_{\text{temporal}}(d). \tag{6}$$

This captures total representation variance driven by the trajectory unfolding over time (task-relevant dynamics *and* any exogenous drift within a single background).

Exogenous variance. For each timestep t, holding the true state fixed, we compute the covariance of the embedding *across* the D distractor instances (i.e., across background variation only) and take its trace:

$$\text{tr-cov}_{\text{exo}}(t) = \operatorname{tr}\big(\operatorname{Cov}_d(\tilde{z}_{1,t}, \dots, \tilde{z}_{D,t})\big), \qquad \overline{\text{tr-cov}}_{\text{exo}} = \frac{1}{T}\sum_{t=1}^{T} \text{tr-cov}_{\text{exo}}(t). \tag{7}$$

Since the true state is held fixed across d at a given t, any variance here is attributable purely to nuisance background changes, not to task-relevant state changes.

Exogenous variance fraction. The final metric is the ratio of exogenous to total temporal variance:

$$\text{ExoVarRatio} = \frac{\overline{\text{tr-cov}}_{\text{exo}}}{\overline{\text{tr-cov}}_{\text{temporal}}}. \tag{8}$$

A lower ratio indicates less variation across distractors relative to temporal variation within trajectories. Because temporal variation can also contain exogenous changes, the ratio should be interpreted alongside endogenous-state decoding.

B.2 VIP with CARDPol: Controlled Goal-Similarity Analysis

Representation training. We train two encoders on the same distracted LIBERO-Spatial demonstrations: one using the VIP objective and one using

$$\mathcal{L} = \mathcal{L}_{\text{VIP}} + \lambda_{\text{aux}} \mathcal{L}_{\texttt{CARDPol}},$$

where $\lambda_{\text{aux}} = 0.1$. Both conditions use the same ResNet18 encoder, training data, preprocessing, augmentation, and optimization budget. Task identity supplies the policy-class label for `CARDPol`. No action labels are used.

Table 8: CARDPol endogenous state recovery loss (MSE, lower is better) as a function of the maximum temporal offset K. Each distraction column is the mean $\pm$ standard error over 3 seeds; Average pools all 9 runs of that environment.

Env.	K	Clean	Static	Video	Average
CHEETAH	0	0.097 ± 0.002	0.108 ± 0.002	0.115 ± 0.002	0.107 ± 0.003
	1	0.103 ± 0.001	0.111 ± 0.002	0.120 ± 0.002	0.111 ± 0.003
	2	0.104 ± 0.002	0.111 ± 0.001	0.121 ± 0.003	0.112 ± 0.003
	4	0.102 ± 0.001	0.111 ± 0.001	0.121 ± 0.003	0.111 ± 0.003
	8	0.105 ± 0.002	0.111 ± 0.002	0.125 ± 0.004	0.113 ± 0.003
	12	0.105 ± 0.001	0.116 ± 0.001	0.124 ± 0.003	0.115 ± 0.003
	16	0.107 ± 0.002	0.113 ± 0.002	0.125 ± 0.002	0.115 ± 0.003
	24	0.114 ± 0.002	0.116 ± 0.002	0.129 ± 0.003	0.119 ± 0.003
	32	0.115 ± 0.003	0.120 ± 0.002	0.128 ± 0.004	0.121 ± 0.003
WALKER	0	0.267 ± 0.004	0.274 ± 0.005	0.278 ± 0.003	0.273 ± 0.003
	1	0.273 ± 0.008	0.279 ± 0.005	0.286 ± 0.003	0.279 ± 0.003
	2	0.266 ± 0.005	0.274 ± 0.002	0.283 ± 0.004	0.275 ± 0.003
	4	0.270 ± 0.005	0.274 ± 0.004	0.277 ± 0.004	0.274 ± 0.002
	8	0.275 ± 0.005	0.279 ± 0.008	0.278 ± 0.001	0.277 ± 0.003
	12	0.265 ± 0.005	0.275 ± 0.004	0.279 ± 0.003	0.273 ± 0.003
	16	0.280 ± 0.003	0.282 ± 0.005	0.287 ± 0.002	0.283 ± 0.002
	24	0.285 ± 0.001	0.291 ± 0.004	0.291 ± 0.004	0.289 ± 0.002
	32	0.284 ± 0.004	0.294 ± 0.004	0.295 ± 0.004	0.291 ± 0.003

Controlled evaluation sequences. We evaluate five trajectories from each of the ten LIBERO-Spatial tasks. These trajectories are included in representation training. From each trajectory, we construct two evaluation sequences. In the robot-motion condition, the original LIBERO-Spatial demonstration evolves while the added distractor region is held at a fixed video frame. In the distractor-motion condition, the underlying scene is fixed, while the distractor video evolves.

Goal similarity and aggregation. For each evaluation sequence, the goal image is the final frame in the original demonstration. We compute cosine similarity between each observation embedding and the goal embedding:

$$s_t = \frac{\phi_\theta(x_t)^\top \phi_\theta(x_g)}{\|\phi_\theta(x_t)\|_2 \, \|\phi_\theta(x_g)\|_2}.$$

We map trajectory timesteps to normalized time in $[0, 1]$ because not all demonstrations have the exact same number of frames. Curves are aggregated by averaging within tasks, then across tasks. The shaded regions indicate confidence interval across trajectories and tasks.

C CARDPOL ABLATIONS

C.1 SENSITIVITY TO K

In Table 8, we evaluate sensitivity to the maximum temporal offset K, sampling observation pairs with offsets $k \in \{1, \ldots, K\}$. We additionally consider $K = 0$, where the same stacked observation is provided to both encoder inputs. Table 8 shows that small offsets, including $K = 0$, yield competitive endogenous-state decoding, while larger offsets generally increase decoding error. Since each observation contains multiple consecutive frames, a single frame stack may already capture enough local motion to distinguish the data-collection policies. This offers a possible explanation for the effectiveness of $K = 0$. Additional temporal separation may provide limited benefit when temporal information is already present within each observation.

C.2 SENSITIVITY TO DATASET COMPOSITION

We evaluate sensitivity to dataset composition by pretraining on trajectories from three, five, or seven policy checkpoints. Table 3 identifies the checkpoints included in each configuration. Each checkpoint contributes 250k transitions, yielding pretraining datasetsof 750k, 1.25M, and 1.75M

transitions, respectively. The downstream expert dataset and frozen-encoder evaluation protocol remain unchanged. This experiment varies both the number of behavior policies and the amount of pretraining data.

Tables 1, 9, and 10, show that CARDPol remains effective across these dataset compositions. Expanding from three to seven policies improves its mean downstream return in all evaluated conditions, although the trend is not uniformly monotonic across the intermediate five-policy configuration. For example, Cheetah Hard-video achieves returns of 204, 293, and 276 with three, five, and seven policies, respectively. These results suggest that broader pretraining datasets can benefit downstream control, while the particular policy mixture also matters. Because policy count and dataset size vary together, this experiment does not isolate the contribution of behavioral diversity.

We additionally evaluate state decoding across subsets containing two to five policies in Figure 6. Each included policy contributes 250k transitions, so policy count and dataset size vary together. Decoding error decreases on average with policy count, whereas higher action-distribution diversity is associated with higher decoding error across the evaluated subsets. This is likely due to the fact that if the endogenous state distributions are highly distinguishable (as would be the case random vs expert data), it becomes extremely easy to classify transitions.

Table 9: Downstream DMC returns with encoders pretrained on three policies (750k transitions).

Env.	Exo.	CARDPol	VIP	ICVF	VAE	VQVAE	CURL
CHEETAH	Clean	$\mathbf{518 \pm 20}$	374 ± 31	437 ± 16	211 ± 11	123 ± 10	480 ± 14
	Static	$\mathbf{559 \pm 9}$	300 ± 24	464 ± 17	147 ± 17	321 ± 21	$\mathbf{565 \pm 7}$
	video	$\mathbf{404 \pm 40}$	188 ± 12	286 ± 12	128 ± 8	114 ± 1	344 ± 11
	hard-video	$\mathbf{204 \pm 10}$	143 ± 5	106 ± 7	100 ± 3	84 ± 4	142 ± 7
WALKER	Clean	$\mathbf{921 \pm 8}$	196 ± 7	746 ± 12	181 ± 7	127 ± 9	687 ± 53
	Static	$\mathbf{696 \pm 27}$	104 ± 3	114 ± 13	64 ± 1	71 ± 2	108 ± 9
	video	$\mathbf{576 \pm 34}$	115 ± 5	100 ± 5	103 ± 4	114 ± 2	91 ± 4
	hard-video	$\mathbf{348 \pm 9}$	86 ± 6	123 ± 4	67 ± 1	77 ± 2	70 ± 3

Table 10: Downstream DMC returns with encoders pretrained on seven policies (1.75M transitions).

Env.	Exo.	CARDPol	VIP	ICVF	VAE	VQVAE	CURL
CHEETAH	Clean	$\mathbf{553 \pm 9}$	440 ± 14	$\mathbf{553 \pm 3}$	203 ± 14	128 ± 6	506 ± 19
	Static	$\mathbf{616 \pm 12}$	436 ± 15	559 ± 11	206 ± 14	256 ± 12	518 ± 23
	video	$\mathbf{499 \pm 19}$	212 ± 14	368 ± 5	115 ± 5	94 ± 2	374 ± 10
	hard-video	$\mathbf{276 \pm 10}$	121 ± 6	231 ± 15	94 ± 5	83 ± 5	176 ± 5
WALKER	Clean	$\mathbf{933 \pm 4}$	372 ± 23	$\mathbf{919 \pm 11}$	221 ± 6	110 ± 12	767 ± 16
	Static	$\mathbf{833 \pm 19}$	130 ± 5	386 ± 50	66 ± 2	65 ± 2	154 ± 15
	video	$\mathbf{733 \pm 10}$	96 ± 3	140 ± 9	86 ± 2	89 ± 4	101 ± 9
	hard-video	$\mathbf{532 \pm 6}$	76 ± 3	150 ± 6	80 ± 6	62 ± 2	80 ± 4

C.3 SENSITIVITY TO REPRESENTATION DIMENSION

We evaluate whether the representation quality of CARDPol depends on the chosen embedding dimension. We vary the encoder output dimension over $d \in \{8, 16, 32, 64, 128, 256, 512, 1024, 2048\}$ and evaluate both endogenous-state decoding and sensitivity to exogenous distractions. For each dimension, we retrain each representation-learning method on each DMC domain (Cheetah and Walker) across the Clean, Static, and Video distraction settings, then freeze the encoder for evaluation. Results are aggregated across all domains and distraction conditions.

Figure 7 shows that CARDPol achieves the lowest state-decoding error and Exogenous Variance ratio across the evaluated dimensions. Increasing the dimension generally improves state decoding, while CARDPol 's exogenous sensitivity remains low. The two measurements are complementary:

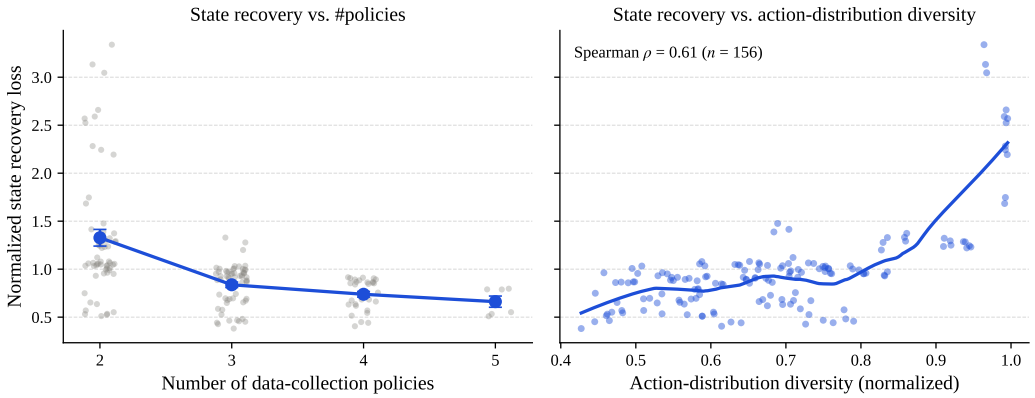


Figure 6: Normalized state-decoding error versus policy count (left) and action-distribution diversity (right). Each included policy contributes 250k transitions. Lower decoding error is better. We quantify action-distribution diversity by fitting a Gaussian kernel density estimate to each policy's observed actions and classifying held-out actions by their most likely source policy. For a subset of m policies, we normalize classification accuracy A as $(A - 1/m)/(1 - 1/m)$, mapping chance accuracy to zero and perfect classification to one. Higher scores indicate more distinguishable marginal action distributions, rather than policy differences conditioned on a shared state.

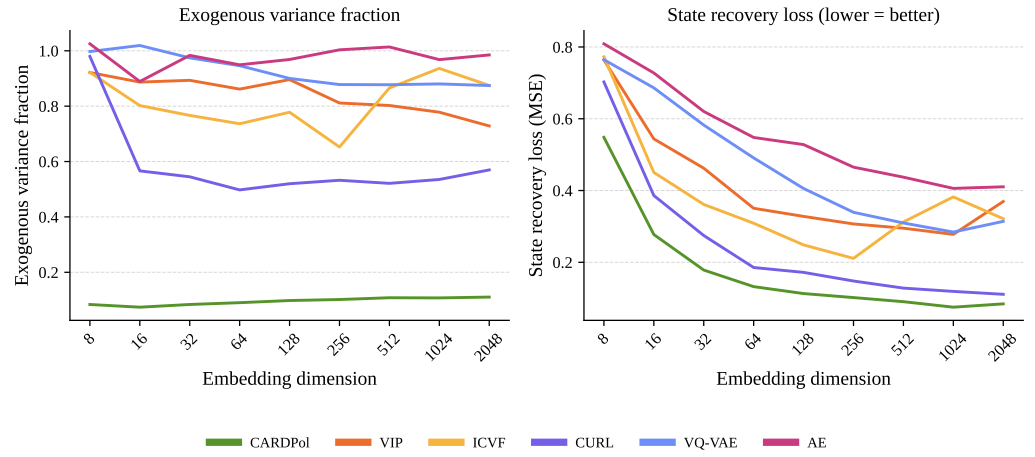


Figure 7: Effect of embedding dimension on exogenous sensitivity (left; lower is better) and endogenous-state decoding (right; lower is better). `CARDPol` maintains the lowest plotted values across the evaluated dimensions.

the decoding results show that the representation retains endogenous information, while the matched-distractor analysis measures its sensitivity to exogenous variation. Together, these results suggest that `CARDPol` 's advantage is not confined to the default embedding dimension of 256 used in the main DMC experiments.

D Additional Control Experiments: OGBench

D.1 Downstream Control in OGBench

We additionally evaluate frozen representations on visual manipulation tasks from OGBench. These experiments extend the DMC evaluation to a different task family and use Implicit Q-Learning (IQL; Kostrikov et al. (2021)) as the downstream offline RL algorithm. We evaluate five tasks in `visual-cube-single` and the Open Drawer task in `visual-scene`, under clean and video-distracted observations.

In the OGBench tasks, we construct datasets using the default data generation pipelines from the OGBench code base. Additionally, we add distractions to the environments by concatenating a video to the right of the original observation (see Figure 4d). Each policy dataset contains 1,000,000 transitions total. For the datasets across all the OGBench domains, we use three policies: `play`, `noisy` ($\xi = 0.1$), and `noisy` ($\xi = 0.3$), using OGBench's own generation code. Behavior identity supplies the policy label for `CARDPol`. Representation pretraining uses the three datasets, with actions and rewards withheld. Downstream IQL uses the `noisy` ($\xi = 0.1$) as has shown to be most performant on `visual-cube` in Park et al. (2025). Distracted observations are constructed by concatenating DAVIS dataset videos to the original observation.

Experimental protocol. We pretrain each encoder on the observation-only datasets described above, withholding actions and rewards. We then freeze the encoder and train IQL on the `Noisy` $\xi = 0.1$ dataset, using their action and reward labels. All representation methods use the same pretraining data and downstream evaluation protocol. We report results over 12 training seeds. For `visual-cube-single`, success rates are aggregated across the five evaluation tasks. The table reports final-checkpoint success.

Table 11: Evaluation performance of an agent trained with IQL (Kostrikov et al., 2021) using frozen pre-trained visual encoders. Each result represents the mean and standard error over 12 seeds.

Env.	Exo.	`CARDPol`	VIP	ICVF	AE	VQVAE	CURL
CUBE (5 tasks)	clean	$\mathbf{0.80 \pm 0.01}$	0.59 ± 0.03	0.16 ± 0.02	0.71 ± 0.03	0.00 ± 0.00	0.41 ± 0.03
	hard-video	$\mathbf{0.41 \pm 0.02}$	0.15 ± 0.01	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
SCENE (1 task)	clean	$\mathbf{0.29 \pm 0.03}$	0.02 ± 0.01	0.03 ± 0.01	0.05 ± 0.02	0.00 ± 0.00	0.02 ± 0.01
	hard-video	$\mathbf{0.17 \pm 0.02}$	0.02 ± 0.01	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.01 ± 0.00

Results. Table 11 reports final-checkpoint success rates. `CARDPol` achieves the highest mean success rate in both environments and observation conditions. In `visual-cube-single`, it achieves success rates of 0.80 with clean observations and 0.41 with video distractions, compared with 0.71 and 0.15 for the strongest respective baselines. On the Open Drawer task in `visual-scene`, `CARDPol` achieves 0.29 and 0.17, compared with 0.05 and 0.02 for the strongest respective baselines. Performance decreases under distractions, but `CARDPol` retains higher success rates than the evaluated alternatives. These results extend the evidence for useful frozen representations to visual manipulation with IQL.

E Theoretical Guarantees

In this section, we show that, under certain assumptions and in the limit of infinite data, the set of encoders ϕ_θ which minimize $\mathcal{L}_{\texttt{CARDPol}}$ as given in Equation 1 is exactly the set of encoders which provide complete information about the endogenous state s_t given x_t.

E.1 Proof of Theorem 1

Suppose that M is an Ex-BMDP that has *connected, aperiodic* latent dynamics, in the sense that: (1) every latent state in $\mathcal{S}$ can be reached from one another (eventually, with nonzero probability, with some sequence of actions) under $\mathcal{T}_s$, and further, (2) the lengths of possible self-loops of any latent state $s \in \mathcal{S}$ have a greatest common denominator of 1. In order to simplify our theoretical analysis, we make several assumptions:

Assumption 1. *$|\mathcal{S}|$, $|\mathcal{E}|$, and $|\mathcal{X}|$ are finite, and that the Markov chain induced by $\mathcal{T}_e$ on $\mathcal{E}$ has all positive entries; in other words, that*

$$\forall e, e', \quad \Pr_{\mathcal{T}_e}(e_{t+1} = e' | e_t = e) > 0. \tag{9}$$

Assumption 2. *For the set of policies $\pi_1, ..., \pi_N : \mathcal{X} \to \mathcal{A}$ on M from which our data is collected:*

$$\forall i, x, x' \in \mathcal{X}, a \in \mathcal{A}, \;\; \phi(x) = \phi(x') \Rightarrow \Pr_{\pi_i}(a_t | x_t = x) = \Pr_{\pi_i}(a_t | x_t = x'), \tag{10}$$

or, in other words, that the policies "only pay attention to" the controllable latent states of the observations, and ignore the noise.

Assumption 2 stems from the "noise-independent policy" assumption from Lamb et al. (2023); Misra et al. (2024); Levine et al. (2024); Islam et al. (2023). Additionally, we assume that the policies differ from each other in a meaningful way:

Assumption 3. *For some constant hyperparameter K, we assume that for all $s, s' \in \mathcal{S}$, there exists some $s'' \in \mathcal{S}$, $k \leq K$, such that s and s' are both reachable from s'' in exactly k steps, and some pair of policies π_i, π_j, such that:*

$$\frac{\Pr_{\pi_i}(s_{t+k} = s | s_t = s'')}{\Pr_{\pi_i}(s_{t+k} = s' | s_t = s'')} \neq \frac{\Pr_{\pi_j}(s_{t+k} = s | s_t = s'')}{\Pr_{\pi_j}(s_{t+k} = s' | s_t = s'')}. \tag{11}$$

However, we also assume that each policy π_i, in any given state s, will take each possible action with nonzero probability: this assumption ensures coverage, and also ensures that both sides of Equation 11 are defined. Because, for each π_i, actions depend only on the endogenous latent state s, the endogenous state transition probabilities

$$\Pr_{\pi_i}(s_{t+1}|s_t) \ (= \sum_a \Pr_{\mathcal{T}_f}(s_{t+1}|s_t, a) \cdot \Pr_{\pi_i}(a|s_t)) \tag{12}$$

are Markovian. Moreover, this endogenous latent Markov chain is irreducible and aperiodic (by the assumption that each policy will take each possible action with nonzero probability, and the assumption that $\mathcal{T}_s$ is connected and aperiodic). Therefore, the endogenous latent state s converges at a *unique stationary distribution* μ_i^s. Furthermore, because $\mathcal{T}_e$ is a positive Markov chain, e also converges at a unique stationary distribution μ^e. We assume that data is collected for a sufficiently long time, such that $\mathcal{D}(\pi_i, k)$ represents these stationary distributions. In other words, if $(x_t, x_{t+k}) \sim \mathcal{D}(\pi_i, k)$, then

$$x_t \sim \mathcal{Q}(s_t, e_t), \text{ where } s_t \sim \mu_i^s \text{ and } e_t \sim \mu^e. \tag{13}$$

Furthermore:

$$\begin{aligned} x_{t+k} &\sim \mathcal{Q}(s_{t+k}, e_{t+k}), \text{ where} \\ \Pr(s_{t+k}|s_t) &= \sum_{s_{t+1},\ldots,s_{t+k-1} \in \mathcal{S}} \prod_{t'=t}^{t+k-1} \Pr_{\pi_i}(s_{t'+1}|s_{t'}) \text{ and} \\ \Pr(e_{t+k}|e_t) &= \sum_{e_{t+1},\ldots,e_{t+k-1} \in \mathcal{E}} \prod_{t'=t}^{t+k-1} \Pr_{\mathcal{T}_e}(e_{t'+1}|e_{t'}). \end{aligned} \tag{14}$$

Note that $\Pr_{\pi_i}(s_{t'+1}|s_{t'})$ in Equation 14 is defined in Equation 12. We see that Equations 13 and 14 fully describe the distribution $\mathcal{D}(\pi_i, k)$. Now, recall the `CARDPol` loss function (Equation 1):

$$\mathcal{L}_{\texttt{CARDPol}}(\phi_\theta) := \mathop{\mathbb{E}}_{\substack{k \sim \mathcal{U}(1,K) \\ i \sim \mathcal{U}(1,N)}} \mathop{\mathbb{E}}_{(x_t, x_{t+k}) \sim \mathcal{D}(\pi_i, k)} -\log(f_i(\phi_\theta(x_t), \phi_\theta(x_{t+k}); k)). \tag{15}$$

Under the above assumptions, we present and prove Theorem 1.

Theorem 1. *The ground-truth encoder ϕ^* associated with an Ex-BMDP M minimizes $\mathcal{L}_{\texttt{CARDPol}}$. Furthermore, any encoder ϕ' which minimizes $\mathcal{L}_{\texttt{CARDPol}}$ is at least as representationally powerful as ϕ^*, in the sense that there exists a function σ such that $\forall x,\ \sigma(\phi'(x)) = \phi^*(x)$.*

Proof. We first show that ϕ minimizes $\mathcal{L}_{\texttt{CARDPol}}$. Note that the cross-entropy loss in Equation 15 is minimized by the function f^* defined as:

$$\forall x_t, x_{t+k} \in \mathcal{X}, k \in [K], \ \ f_i^*(\phi_\theta(x_t), \phi_\theta(x_{t+k}); k) := \Pr(\pi_i | x_t, x_{t+k}). \tag{16}$$

Furthermore, $\Pr(\pi_i|x_t, x_{t+k})$ is the *unique* minimizer of the cross-entropy loss function, everywhere that the pair (x_t, x_{t+k}) is visited with nonzero probability.

Note that by Equations 10, 13, and 14, we have:

$$\begin{aligned}
\Pr(\pi_i|x_t, x_{t+k}) &= \frac{\Pr_{\pi_i}(x_{t+k}|x_t)\cdot\Pr(\pi_i|x_t)}{\sum_{\pi_j}\frac{1}{N}\Pr_{\pi_j}(x_{t+k}|x_t)} \\
&= \frac{\Pr_{\mathcal{Q}}(x_{t+k}|\phi(x_{t+k}),\phi_e(x_{t+k}))\cdot\Pr(\phi_e(x_{t+k})|\phi_e(x_t))\cdot\Pr_{\pi_i}(\phi(x_{t+k})|\phi(x_t))\cdot\Pr(\pi_i|x_t)}{\sum_{\pi_j}\frac{1}{N}\Pr_{\mathcal{Q}}(x_{t+k}|\phi(x_{t+k}),\phi_e(x_{t+k}))\cdot\Pr(\phi_e(x_{t+k})|\phi_e(x_t))\cdot\Pr_{\pi_j}(\phi(x_{t+k})|\phi(x_t))} \\
&= \frac{\Pr_{\mathcal{Q}}(x_{t+k}|\phi(x_{t+k}),\phi_e(x_{t+k}))\cdot\cancel{\Pr(\phi_e(x_{t+k})|\phi_e(x_t))}}{\Pr_{\mathcal{Q}}(x_{t+k}|\phi(x_{t+k}),\phi_e(x_{t+k}))\cdot\cancel{\Pr(\phi_e(x_{t+k})|\phi_e(x_t))}}\frac{\Pr_{\pi_i}(\phi(x_{t+k})|\phi(x_t))\cdot\Pr(\pi_i|x_t)}{\sum_{\pi_j}\frac{1}{N}\Pr_{\pi_j}(\phi(x_{t+k})|\phi(x_t))} \\
&= \frac{\Pr_{\pi_i}(\phi(x_{t+k})|\phi(x_t))\cdot\Pr(\pi_i|x_t)}{\sum_{\pi_j}\frac{1}{N}\Pr_{\pi_j}(\phi(x_{t+k})|\phi(x_t))} \\
&= \frac{\Pr_{\pi_i}(\phi(x_{t+k})|\phi(x_t))}{\sum_{\pi_j}\frac{1}{N}\Pr_{\pi_j}(\phi(x_{t+k})|\phi(x_t))}\cdot\frac{\frac{1}{N}\Pr_{\pi_i}(x_t)}{\sum_{\pi_j}\frac{1}{N}\Pr_{\pi_j}(x_t)} \\
&= \frac{\Pr_{\pi_i}(\phi(x_{t+k})|\phi(x_t))}{\sum_{\pi_j}\frac{1}{N}\Pr_{\pi_j}(\phi(x_{t+k})|\phi(x_t))}\cdot\frac{\Pr_{\mathcal{Q}}(x_t|\phi(x_t),\phi_e(x_t))\mu^e(\phi_e(x_t))\mu^s_i(\phi(x_t))}{\sum_{\pi_j}\Pr_{\mathcal{Q}}(x_t|\phi(x_t),\phi_e(x_t))\mu^e(\phi_e(x_t))\mu^s_j(\phi(x_t))} \\
&= \frac{\Pr_{\pi_i}(\phi(x_{t+k})|\phi(x_t))}{\sum_{\pi_j}\frac{1}{N}\Pr_{\pi_j}(\phi(x_{t+k})|\phi(x_t))}\cdot\frac{\mu^s_i(\phi(x_t))}{\sum_{\pi_j}\mu^s_j(\phi(x_t))}.
\end{aligned}$$

The final line depends on x_t and x_{t+k} only through $\phi(x_t)$ and $\phi(x_{t+k})$, respectively, so if $\phi_\theta = \phi$, then a function f^* can be constructed such that $f^*_i(\phi_\theta(x_t), \phi_\theta(x_{t+k}); k) = \Pr(\pi_i|x_t, x_{t+k})$, as desired. We now prove that *any* encoder ϕ' which minimizes $\mathcal{L}_{\texttt{CARDPol}}$ is at least as representationally powerful as ϕ, in the sense that there exists a function σ such that $\forall x,\ \sigma(\phi'(x)) = \phi(x)$.

Consider any ϕ' such that no such σ exists. Then, for some pair $x, x' \in \mathcal{X}, \phi'(x) = \phi'(x')$, but $\phi(x) \neq \phi(x')$. Let $s = \phi(x), s' = \phi(x'), e = \phi_e(x), e' = \phi(e')$. Then, by Equation 11, we have that for some $i, j \in [N]$, and some $s'' \in \mathcal{S}$:

$$\frac{\Pr_{\pi_i}(s_{t+k} = s|s_t = s'')}{\Pr_{\pi_i}(s_{t+k} = s'|s_t = s'')} \neq \frac{\Pr_{\pi_j}(s_{t+k} = s|s_t = s'')}{\Pr_{\pi_j}(s_{t+k} = s'|s_t = s'')}. \tag{17}$$

Choose any $e'' \in \mathcal{E}$, and $x'' \in$ Support$(\mathcal{Q}(s'', e''))$. By the assumption that $\mathcal{T}_e$ has a positive transition matrix, $\Pr(e_{t+k} = e|e_t = e'')$ and $\Pr(e_{t+k} = e'|e_t = e'')$ are nonzero; multiplying both sides of Equation 17 by the same nonzero value yields:

$$\begin{aligned}
&\frac{\Pr_{\pi_i}(s_{t+k} = s|s_t = s'')}{\Pr_{\pi_i}(s_{t+k} = s'|s_t = s'')}\cdot\frac{\Pr_{\mathcal{Q}}(x|s,e)\cdot\Pr(e_{t+k} = e|e_t = e'')}{\Pr_{\mathcal{Q}}(x'|s',e')\cdot\Pr(e_{t+k} = e'|e_t = e'')} \neq \\
&\frac{\Pr_{\pi_j}(s_{t+k} = s|s_t = s'')}{\Pr_{\pi_j}(s_{t+k} = s'|s_t = s'')}\frac{\Pr_{\mathcal{Q}}(x|s,e)\cdot\Pr(e_{t+k} = e|e_t = e'')}{\Pr_{\mathcal{Q}}(x'|s',e')\cdot\Pr(e_{t+k} = e'|e_t = e'')}.
\end{aligned} \tag{18}$$

Applying the definition of Ex-BMDP dynamics, this simplifies to:

$$\frac{\Pr_{\pi_i}(x_{t+k} = x|x_t = x'')}{\Pr_{\pi_i}(x_{t+k} = x'|x_t = x'')} \neq \frac{\Pr_{\pi_j}(x_{t+k} = x|x_t = x'')}{\Pr_{\pi_j}(x_{t+k} = x'|x_t = x'')}. \tag{19}$$

By the irreducibility of the endogenous state Markov chains for both policies, s'' has nonzero probability in μ^i_s and μ^j_s, and consequently $\Pr_{\pi_i}(x_t = x''), \Pr_{\pi_j}(x_t = x'') > 0$. Multiplying the left-hand side of Equation 19 by $\frac{\Pr_{\pi_i}(x_t = x'')}{\Pr_{\pi_i}(x_t = x'')}(= 1)$ and the right-hand side by $\frac{\Pr_{\pi_j}(x_t = x'')}{\Pr_{\pi_j}(x_t = x'')}(= 1)$ yields:

$$\frac{\Pr_{\pi_i}(x_{t+k} = x \wedge x_t = x'')}{\Pr_{\pi_i}(x_{t+k} = x' \wedge x_t = x'')} \neq \frac{\Pr_{\pi_j}(x_{t+k} = x \wedge x_t = x'')}{\Pr_{\pi_j}(x_{t+k} = x' \wedge x_t = x'')}. \tag{20}$$

Applying Bayes' rule to both sides and cancelling like terms gives:

$$\frac{\Pr(\pi_i|x_{t+k} = x \wedge x_t = x'')}{\Pr(\pi_i|x_{t+k} = x' \wedge x_t = x'')} \neq \frac{\Pr(\pi_j|x_{t+k} = x \wedge x_t = x'')}{\Pr(\pi_j|x_{t+k} = x' \wedge x_t = x'')}. \tag{21}$$

Multiplying both sides by $\frac{\Pr(\pi_i|x_{t+k}=x'\wedge x_t=x'')}{\Pr(\pi_j|x_{t+k}=x\wedge x_t=x'')}$ (which is nonzero by the coverage assumption):

$$\frac{\Pr(\pi_i|x_{t+k}=x\wedge x_t=x'')}{\Pr(\pi_j|x_{t+k}=x\wedge x_t=x'')}\neq\frac{\Pr(\pi_i|x_{t+k}=x'\wedge x_t=x'')}{\Pr(\pi_j|x_{t+k}=x'\wedge x_t=x'')}. \tag{22}$$

Then either $\Pr(\pi_i|x_{t+k}=x\wedge x_t=x'')\neq\Pr(\pi_i|x_{t+k}=x'\wedge x_t=x'')$; or $\Pr(\pi_j|x_{t+k}=x\wedge x_t=x'')\neq\Pr(\pi_j|x_{t+k}=x'\wedge x_t=x'')$; or both. Without loss of generality, if $\Pr(\pi_i|x_{t+k}=x\wedge x_t=x'')\neq\Pr(\pi_i|x_{t+k}=x'\wedge x_t=x'')$ but $\phi'(x)=\phi'(x')$, then $\forall f$, either $f_i(\phi'(x''),\phi'(x);k)\neq\Pr(\pi_i|x_t=x'',x_{t+k}=x)$ or $f_i(\phi'(x''),\phi'(x');k)\neq\Pr(\pi_i|x_t=x'',x_{t+k}=x')$. By construction, the pairs $(x_t=x'',x_{t+k}=x')$ and $(x_t=x'',x_{t+k}=x)$ both occur with nonzero probability, so $f_i(\phi'(x_t),\phi'(x_{t+k});k)$ cannot equal $\Pr(\pi_i|x_t,x_{t+k})$ over the entire support. Then no choice of f allows the cross-entropy loss to achieve its minimum value.

□

F Description of CRAFT (Levine et al., 2025) Algorithm

`CARDPol` is inspired by CRAFT (Levine et al., 2025) but significantly extends its applicability by relaxing core assumptions. For comparison, we give a high-level description of CRAFT.

Assumption 4. *We are given two datasets, τ_A and τ_B, containing observation-only trajectories of an Ex-BMDP with discrete and deterministic endogenous dynamics, one collected by agent A and B, respectively.*

Assumption 5. *Each trajectory contained in τ_A and τ_B start in the same endogenous state.*

Each trajectory is treated as a sequence of consecutive observation pairs, i.e., $(x_1,x_2),(x_2,x_3),\ldots,(x_{H-1},x_H)$. At each time step h, the possible latent endogenous states are denoted by $s_h, s'_h, s''_h, \ldots$.

Assumption 6. *For $\alpha>0$ with fixed s_h and distinct s_{h+1} and s'_{h+1}, we assume:*

$$\left|\ln\left(\frac{|\mathcal{D}_A(s_h,s_{h+1})|}{|\mathcal{D}_B(s_h,s_{h+1})|}\right)-\ln\left(\frac{|\mathcal{D}_A(s_h,s'_{h+1})|}{|\mathcal{D}_B(s_h,s'_{h+1})|}\right)\right|\geq\alpha.$$

Conditioned on each observation pair (x_h,x_{h+1}), learn a classifier f_h to predict the probability of being collected by either agent A or agent B. The classification represented is not represented as binary, but rather via a *log-odds ratio* of the probability of being collected by agent A or agent B. Formally, f_h predicts the following:

$$\ln\left(\frac{\Pr[(x_h,x_{h+1})\in(\tau_A)_{h,h+1}\mid(x_h,x_{h+1})\in(\tau_A\uplus\tau_B)_{h,h+1}]}{\Pr[(x_h,x_{h+1})\in(\tau_B)_{h,h+1}\mid(x_h,x_{h+1})\in(\tau_A\uplus\tau_B)_{h,h+1}]}\right).$$

This is achieved by training each classifier f_h to minimize the following loss function (Levine et al., 2025):

$$\mathcal{L}(f_h):=\sum_{(x_h,x_{h+1})\in(\tau_A)_{h,h+1}}\ln\Big(1+e^{-f(x_h,x_{h+1})}\Big)+\sum_{(x_h,x_{h+1})\in(\tau_B)_{h,h+1}}\ln\Big(1+e^{f(x_h,x_{h+1})}\Big). \tag{23}$$

In the limit of infinite data, the optimal f_h will return:

$$f_h^*(x_h,x_{h+1})\to\ln\left(\frac{\left|\mathcal{D}_A^*\left(\phi_h^*(x_h),\phi_{h+1}^*(x_{h+1})\right)\right|}{\left|\mathcal{D}_B^*\left(\phi_h^*(x_h),\phi_{h+1}^*(x_{h+1})\right)\right|}\right).$$

After minimizing eq. (23) at each timestep h, Levine et al. (2025) shows that transitions (x_h,x_{h+1}) corresponding to the same endogenous states will produce distinct clusters of values of $f(x_h,x_{h+1})$. Due to this clustering and Assumption 5, the clusters associated with observation pairs at $h=1$, (x_1,x_2) , directly and uniquely represent each possible endogenous state at $h=2$ (e.g., $s_2, s'_2, s''_2, \ldots$). CRAFT then progresses to the next timestep $h=2$ by now considering only observation pairs that

start a particular s_2 (denoted as $\texttt{pairs}(s_2)$). CRAFT analyzes the spectrum of values produced by the classification of $\texttt{pairs}(s_2)$. We expect to see distinct clusters of values of $f_2(x_2, x_3)$ for all pairs in $\texttt{pairs}(s_2)$. This clustering is done for each subset $\texttt{pairs}(s_2)$, $\texttt{pairs}(s_2')$, etc. After identifying the possible latent successors of each $s_2' \in S_2$ separately, CRAFT determines whether successor states associated with distinct states $s_2, s_2', s_2'' \in S_2$ correspond to the same latent state $s_3 \in S_3$. This can be accomplished by learning binary classifiers between the observations x_3 appearing in the corresponding observation-pair sets. If the observations are indistinguishable, then the observation-pair sets correspond to the same latent state. If they are perfectly distinguishable, then they correspond to distinct latent states. Now, all the distinct latent states at $h = 3$ have been identified and we can continue to $h = 4$ in the same manner. Finally, encoders ϕ_h are trained at each step to predict the newly identified latent endogenous states.

Both the application and sample efficiency guarantees of CRAFT necessitate both identical starting states for all trajectories and discrete, deterministic endogenous MDP. In this work, we relax these core assumptions. A complete and formal description of CRAFT can be found in Levine et al. (2025).