Principled Latent Actions through Policy Discrimination

Anonymous authors
Paper under double-blind review

Abstract

Scaling sequential decision-making requires data, and video may well be the most abundant source of real-world data. Latent or world action models offer a promising approach for leveraging video data by learning a representation that succinctly captures variation across frames. However, this representation captures not only action information, but also dynamic noise, which can result in poor or conflated action representation spaces. When the noise-conflated latent action space is then mapped to the true action space, this limits the performance of learned policies. In this work, we introduce a formal characterization of this conflated action representation space. We then leverage advances in causal theory to illustrate how using policy indicators provide sufficient information to disentangle actions from noise. We then introduce a simple modification to the classic latent action model paradigm to learn this space. Compared with other latent action model methods, we show improved performance in action recovery and in several distracted MuJoCo locomotion settings.

1 Introduction

Recent advances in sequential decision-making have seen impressive policies learned in a host of domains, from games (Mnih et al., 2015; Fuchs et al., 2021; Fan et al., 2022) to challenging simulated environments for open-ended, embodied, and multi-agent decision-making (Küttler et al., 2020; Albrecht et al., 2022; Li et al., 2024), language models (Ouyang et al., 2022), and many more. Yet learning these policies is predicated on a substantial and often costly amount of interaction data, i.e., tuples of at least state, action, and next-state interactions, whether collected offline via demonstrators (Fu et al., 2020; Mandlekar et al., 2023; Walke et al., 2023) or online through an online reinforcement learning agent (Mnih et al., 2015; Haarnoja et al., 2018). However, there is much more abundant action-free data in the form of video, which could mitigate this cost (Seo et al., 2022; Nair et al., 2023; Ma et al., 2022).

Latent or World action models (Pai et al., 2025; Ye et al., 2026; Li et al., 2026; Garrido et al., 2026a) have seen substantial recent success in leveraging this data to train impressive robot policies. These models involve learning a latent action space that describes the dynamics of the state transition between two states alongside the forward dynamics or world model. Then, a small number of labeled trajectories to uncover a mapping between this latent action space and the true distribution of actions. Thus, these models transform video into action-labeled trajectories that can be used with traditional sequential decision-making algorithms.

While promising, these methods also suffer from a fundamental flaw: they cannot disentangle dynamic noise from the underlying actions. This is because the latent "action" vector is learned from the variation in the state-next state pair via information bottleneck only – any change, whether from the robot actions or from the background, will be captured in this latent space. However, the real world is full of time-correlated noise, whether it is people or objects moving in the background, variations in lighting over time, or distractions moving through the workspace. In this case, the latent

actions will capture all this variation, even though the agent does not have control over any of these. This introduces a significant possible source of error when recovering the mapping between latent and true actions, resulting in poor performance.

In this work, we first characterize this challenge through recent findings from causal representation learning that indicate the core assumptions necessary to identify a latent variable from observational data (Rodriguez-Sanchez et al., 2025)—that is, recover the true actions from state transitions. In particular, Lachapelle et al. (2022) and Xu et al. (2024) show that autoencoding is sufficient to recover the latent if it is sparse, and with a downstream variable that is intervened on. We combine this with findings from action-free learning (Levine et al., 2025), which indicate policy indicators as the key downstream variable that can be used to properly disentangle the true latent action.

This elucidates our core contribution: POLDISC , which learns principled latent actions from observational state by using policy discrimination as a joint objective for learning the state encoders (POLDISC-z) or the latent action itself (POLDISC-a). We then provide theoretical justification for policy discrimination as a sufficient variable to identify the true actions from the latent actions under the previously mentioned assumptions. Finally, we show empirically that our method is able to achieve better action recovery of the true actions in distracted Mujoco environments, Cheetah, Walker, and Hopper.

2 Preliminaries

Exogenous Markov Decision Processes. We consider environments that can be modeled as Exogenous Markov Decision Processes (Exo-MDPs), where the state space admits a decomposition $\mathcal{S} := \mathcal{X} \times \mathcal{E}$. Here, $\mathcal{X}$ denotes the endogenous state space, containing the components of the environment that are affected by the agent's actions, and $\mathcal{E}$ denotes the exogenous state space, containing components that evolve independently of the agent. A state is written as $\mathbf{s} = (\mathbf{x}, \mathbf{e})$, where $\mathbf{x} \in \mathcal{X}$ and $\mathbf{e} \in \mathcal{E}$. The action space is denoted by $\mathcal{A}$, with actions $\mathbf{a} \in \mathcal{A}$.

The transition dynamics of the Exo-MDP are given by $p(\mathbf{s}' \mid \mathbf{s}, \mathbf{a}) = \mathbb{P}(S' = \mathbf{s}' \mid S = \mathbf{s}, A = \mathbf{a})$. The defining assumption of an Exo-MDP is that these dynamics factorize into endogenous and exogenous components: $p(\mathbf{s}' \mid \mathbf{s}, \mathbf{a}) = p_{\mathcal{X}}(\mathbf{x}' \mid \mathbf{x}, \mathbf{a}) p_{\mathcal{E}}(\mathbf{e}' \mid \mathbf{e})$. Thus, the endogenous component evolves as a function of the current endogenous state and action, while the exogenous component evolves independently of both the action and the endogenous state. In practice, this decomposition is not observed directly: we only observe states or observations containing both endogenous and exogenous information. A central goal is therefore to learn representations that emphasize the controllable, endogenous factors while discarding irrelevant exogenous variation.

Action-Free Multi-Policy Data. We assume access to an action-free dataset of trajectories collected in the same Exo-MDP under a finite set of behavior policies

$$\Pi := \pi_1, \ldots, \pi_N.$$

Each trajectory is labeled with a one-hot policy class indicating which policy generated it:

$$\mathcal{D} = (\tau_j, \mathbf{y}_j)_{j=1}^{M}, \qquad \tau_j = (\mathbf{s}_0^j, \ldots, \mathbf{s}_{T_j}^j), \qquad \mathbf{y}_j \in 0, 1^N.$$

The policies are assumed to be endogenous: each π_i selects actions only as a function of the endogenous state x_t, and does not depend on the exogenous state e_t.

Learning Latent Actions. Given action-free transitions $(\mathbf{s}, \mathbf{s}')$, latent action learning aims to infer a compact variable a_l that explains the change from $\mathbf{s}$ to $\mathbf{s}'$. Rather than learning this model directly in the observation space, recent work often learns latent actions in a representation space. Let $\phi : \mathcal{S} \to \mathcal{Z}$ be an encoder, with $\mathbf{z} = \phi(\mathbf{s})$ and $\mathbf{z}' = \phi(\mathbf{s}')$. A latent inverse model predicts

$$a_l = f_\psi(\mathbf{z}, \mathbf{z}'),$$

and a latent forward model predicts the next representation from the current representation and inferred latent action:

$$\mathcal{L}_{\text{dyn}} = \mathop{\mathbb{E}}_{\mathbf{s},\mathbf{s}'\sim\mathcal{D}} \left[\left\| \bar{\mathbf{z}}' - g_\theta\left(\mathbf{z}, a_l\right) \right\|^2 \right], \tag{1}$$

where $\bar{\mathbf{z}}$ represents a stop-gradient operation. Because f_ψ receives both the current and target representations as input, this objective requires an information bottleneck to prevent the inferred latent action from encoding arbitrary details of the next observation. Prior work imposes this bottleneck using discretization (Schmidt & Jiang, 2023), sparsity regularization (Garrido et al., 2026b), noise regularization (Gao et al., 2025), or action supervision (Nikulin et al., 2025). In this work, we instead use policy identity as an action-free supervisory signal for learning representations that emphasize controllable changes.

3 Policy Discrimination as Supervision

We use policy identity as action-free supervision of both representations and latent actions. Given datasets of trajectories generated by policies $\{\pi_i\}_{i=1}^N$, each transition is labeled by the policy $i \in \{1, \ldots, N\}$ that generated it. Given a trajectory generated by policy π_i, we sample state pairs $(\mathbf{s}_t, \mathbf{s}_{t+k})$, where k is sampled uniformly from $\{1, \ldots, K\}$. Since different policies induce different controllable transitions in the same environment, the policy label provides weak supervision for identifying transition information relevant to control. We consider two separate variants of this supervision. The first applies policy discrimination in representation space, while the second applies policy discrimination in latent-action space. Each variant is added independently to the latent-action dynamics loss in Equation (1).

Representation-Space Policy Discrimination. Our first method, denoted PolDisc-z, trains a classifier to predict the generating policy from a pair of representations. Given a sampled transition pair $(\mathbf{s}_t, \mathbf{s}_{t+k})$, we compute

$$\mathbf{z}_t = \phi(\mathbf{s}_t), \qquad \mathbf{z}_{t+k} = \phi(\mathbf{s}_{t+k}).$$

We define a policy classifier

$$h_\omega^\phi : \mathcal{Z} \times \mathcal{Z} \to \Delta^{N-1},$$

which predicts the policy index i from $(\mathbf{z}_t, \mathbf{z}_{t+k})$. The representation-space policy-discrimination loss is

$$\mathcal{L}_{\text{pol}}^z = \mathop{\mathbb{E}}_{(\tau,i)\sim\mathcal{D},\ t,\ k\sim\text{Unif}(\{1,\ldots,K\})} \left[-\ln h_\omega^z\left(i|\phi(\mathbf{s}_t), \phi(\mathbf{s}_{t+k})\right), \right], \tag{2}$$

This objective encourages the encoder to preserve transition information that distinguishes the behavior policies. Since policy-dependent variation arises through the agent's actions, this biases the representation toward controllable factors and away from exogenous variation that is independent of the policy.

The full objective for this variant is

$$\mathcal{L}^z = \mathcal{L}_{\text{dyn}} + \lambda_\phi \mathcal{L}_{\text{pol}}^z. \tag{3}$$

Latent-Action Policy Discrimination. Our second method, denoted PolDisc-ℓ, trains a classifier to predict the generating policy from the inferred latent action. For the same sampled transition pair, the latent inverse model produces

$$\mathbf{a}_\ell = f_\psi\left(\phi(\mathbf{s}_t), \phi(\mathbf{s}_{t+1})\right).$$

We define a policy classifier

$$h_\eta^\ell : \mathcal{A}_\ell \to \Delta^{N-1},$$

which predicts the policy index i from the latent action $\mathbf{a}_\ell$. The latent-action policy-discrimination loss is

$$\mathcal{L}_{\text{pol}}^\ell = \mathop{\mathbb{E}}_{(s_t,s_{t+1},i)\sim\mathcal{D}} \left[-\ln h_\eta^\ell\left(i|\mathbf{a}_\ell\right) \right]. \tag{4}$$

Table 1: Final Action Decoding MSE with mean and standard error over 4 seeds. ↓ is better. Bolded values indicate the lowest MSE in the column, accounting for overlapping standard error bounds.

	Cheetah		Hopper		Walker	
	Clean	**Video**	**Clean**	**Video**	**Clean**	**Video**
No Regulrization	$\mathbf{0.41 \pm 0.01}$	0.76 ± 0.00	0.39 ± 0.02	0.71 ± 0.01	$\mathbf{0.53 \pm 0.01}$	0.73 ± 0.01
POLDISC-z (ours)	$\mathbf{0.42 \pm 0.01}$	$\mathbf{0.68 \pm 0.01}$	0.39 ± 0.01	0.64 ± 0.02	0.56 ± 0.01	$\mathbf{0.71 \pm 0.01}$
POLDISC-a (ours)	0.43 ± 0.01	$\mathbf{0.70 \pm 0.01}$	$\mathbf{0.36 \pm 0.01}$	$\mathbf{0.62 \pm 0.02}$	$\mathbf{0.53 \pm 0.01}$	$\mathbf{0.70 \pm 0.01}$
Noise (Garrido et al., 2026a)	0.59 ± 0.02	0.75 ± 0.01	0.58 ± 0.02	0.72 ± 0.01	0.55 ± 0.01	0.74 ± 0.01
Sparsity (Garrido et al., 2026a)	0.98 ± 0.01	0.79 ± 0.01	0.73 ± 0.03	0.90 ± 0.03	0.84 ± 0.02	0.79 ± 0.01

This objective encourages the inferred latent action to encode policy-relevant controllable variation. Unlike POLDISC-z , which supervises the representation pair directly, POLDISC-a places the policy-discrimination pressure on the latent action bottleneck itself.

The full objective for this variant is

$$\mathcal{L}^{\ell} = \mathcal{L}_{\text{dyn}} + \lambda_{\ell} \mathcal{L}^{\ell}_{\text{pol}}. \tag{5}$$

Please refer to Section 6 for a theoretical treatment of these discrmination objectives.

4 Experiments

We evaluate whether policy-discrimination supervision improves the quality of learned latent actions. For each environment, we construct an action-free multi-policy dataset from five policy checkpoints, each containing 1M transitions. The checkpoints span a range of behavior quality, from uniformely random action policies to expert policies. Data from each checkpoint is treated as a distinct policy class for the policy-discrimination objectives. We use two environments for rendering the policies: `clean` and `video`. The `clean` environment is the standard visual Mujoco rendering. The `video` environment, introduced in the Distracting Control Suite (Stone et al., 2021), injects a real-world video distraction behind the rendered agent. In this setting, the agent rendering represents the endogenous component and the video distraction is the exogenous component.

We train latent action models using the latent dynamics objective $\mathcal{L}_{\text{dyn}}$ together with the policy-discrimination objective $\mathcal{L}_{\text{pol}}$. After training, we evaluate on a held-out set of transitions with ground-truth actions. Since actions are used only for evaluation, this measures whether the learned latent actions recover the true underlying actions without direct action supervision during training. We report *action recovery*, which measures how well the inferred latent actions predict the true actions. For baselines, evaluate noise injection and sparsity regularization for latent action models. The noise baseline adds uses a gaussian prior with a target of unit variance. The sparsity baseline uses a principled ℓ_1 loss that introduces sparsity while preventing collapse. See Section 7 for more information.

As we can see in Table 1, both the POLDISC-a and POLDISC-z methods achieve among the lowest action-decoding error of all the baselines. In the `clean` environments, the "No Regularization" baseline can perform similarly to our methods; this is likely because the only movement in the frames is due to the agent itself and so the latent action is not polluted with any irrelevant information. In contrast, in each of the `video` environments, the POLDISC-z and POLDISC-a methods achieve lower action decoding than any of the other methods.

5 Conclusion

In this work we introduce a weak supervision method for latent action learning. We demonstrate that better action representations can be captured when using this weak supervision baseline regularization methods. We present this work as an invitation for further investigation.

References

Joshua Albrecht, Abraham J. Fetterman, Bryden Fogelman, Ellie Kitanidis, Bartosz Wroblewski, Nicole Seo, Michael Rosenthal, Maksis Knutins, Zachary Polizzi, James B. Simon, and Kanjun Qiu. Avalon: A benchmark for RL generalization using procedurally generated worlds. In *Advances in Neural Information Processing Systems*, 2022.

Linxi Fan, Guanzhi Wang, Yunfan Jiang, Ajay Mandlekar, Yuncong Yang, Haoyi Zhu, Andrew Tang, De-An Huang, Yuke Zhu, and Anima Anandkumar. MineDojo: Building open-ended embodied agents with internet-scale knowledge. *Advances in Neural Information Processing Systems*, 35: 18343–18362, 2022.

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4RL: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*, 2020.

Florian Fuchs, Yunlong Song, Elia Kaufmann, Davide Scaramuzza, and Peter Dürr. Super-human performance in gran turismo sport using deep reinforcement learning. *IEEE Robotics and Automation Letters*, 6(3):4257–4264, 2021.

Shenyuan Gao, Siyuan Zhou, Yilun Du, Jun Zhang, and Chuang Gan. Adaworld: Learning adaptable world models with latent actions, 2025. URL https://arxiv.org/abs/2503.18938.

Quentin Garrido, Tushar Nagarajan, Basile Terver, Nicolas Ballas, Yann LeCun, and Michael Rabbat. Learning latent action world models in the wild. *arXiv preprint arXiv:2601.05230*, 2026a.

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

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*, pp. 1861–1870. PMLR, 2018.

Heinrich Küttler, Nantas Nardelli, Alexander H. Miller, Roberta Raileanu, Marco Selvatici, Edward Grefenstette, and Tim Rocktäschel. The NetHack learning environment. In *Advances in Neural Information Processing Systems*, volume 33, pp. 7671–7684, 2020.

Sébastien Lachapelle, Pau Rodriguez, Yash Sharma, Katie E Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ica. In *Conference on Causal Learning and Reasoning*, pp. 428–484. PMLR, 2022.

Alexander Levine, Peter Stone, and Amy Zhang. Offline action-free learning of ex-bmdps by comparing diverse datasets, 2025. URL https://arxiv.org/abs/2503.21018.

Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martín-Martín, Chen Wang, Gabrael Levine, Wensi Ai, Benjamin Martinez, et al. BEHAVIOR-1k: A human-centered, embodied AI benchmark with 1,000 everyday activities and realistic simulation. *arXiv preprint arXiv:2403.09227*, 2024.

Lin Li, Qihang Zhang, Yiming Luo, Shuai Yang, Ruilin Wang, Fei Han, Mingrui Yu, Zelin Gao, Nan Xue, Xing Zhu, et al. Causal world modeling for robot control. *arXiv preprint arXiv:2601.21998*, 2026.

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.

Ajay Mandlekar, Soroush Nasiriany, Bowen Wen, Iretiayo Akinola, Yashraj Narang, Linxi Fan, Yuke Zhu, and Dieter Fox. Mimicgen: A data generation system for scalable robot learning using human demonstrations, 2023. URL https://arxiv.org/abs/2310.17596.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

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. *arXiv preprint arXiv:2502.00379*, 2025.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744, 2022.

Jonas Pai, Liam Achenbach, Victoriano Montesinos, Benedek Forrai, Oier Mees, and Elvis Nava. mimic-video: Video-action models for generalizable robot control beyond vlas. *arXiv preprint arXiv:2512.15692*, 2025.

Rafael Rodriguez-Sanchez, Cameron Allen, and George Konidaris. From pixels to factors: Learning independently controllable state variables for reinforcement learning, 2025. URL https://arxiv.org/abs/2510.02484.

Dominik Schmidt and Minqi Jiang. Learning to act without actions. *arXiv preprint arXiv:2312.10812*, 2023.

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

Austin Stone, Oscar Ramirez, Kurt Konolige, and Rico Jonschkowski. The distracting control suite – a challenging benchmark for reinforcement learning from pixels, 2021. URL https://arxiv.org/abs/2101.02722.

Homer Rich Walke, Kevin Black, Abraham Lee, Moo Jin Kim, Max Du, Chongyi Zheng, Tony Zhao, Philippe Hansen-Estruch, Quan Vuong, Andre He, Vivek Myers, Kuan Fang, Chelsea Finn, and Sergey Levine. BridgeData V2: A dataset for robot learning at scale. In *Conference on Robot Learning*, pp. 1723–1736. PMLR, 2023.

Danru Xu, Dingling Yao, Sébastien Lachapelle, Perouz Taslakian, Julius Von Kügelgen, Francesco Locatello, and Sara Magliacane. A sparsity principle for partially observable causal representation learning. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 55389–55433, 2024.

Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, et al. World action models are zero-shot policies. *arXiv preprint arXiv:2602.15922*, 2026.

Supplementary Materials

The following content was not necessarily subject to peer review.

6 Theory

Theorem 1 (Endogenous Independence)**.** *We aim to prove that the exogenous variable $\mathbf{e}_t$ does not depend on the estimated latent action $\hat{\mathbf{a}}_t$, i.e.,*

$$\frac{\partial \mathbf{e}_t}{\partial \hat{\mathbf{a}}_t} = 0.$$

To establish this, consider any set $Z \subseteq \mathcal{A} \times \mathcal{E}$ satisfying the following assumptions:

Assumption 1 (Non-Separability)**.** *Z cannot be expressed as $A \times \mathcal{E}$ for any $A \subset \mathcal{A}$.*

Assumption 2 (Domain-Induced Variability)**.** *There exist $u_1, u_2 \in \mathcal{U}$ such that*

$$\int_{(\mathbf{a},\mathbf{e})\in Z} p\left(\mathbf{a}, \mathbf{e} \mid \mathbf{u}_1\right) d\mathbf{a}d\mathbf{e} \neq \int_{(\mathbf{a},\mathbf{e})\in Z} p\left(\mathbf{a}, \mathbf{e} \mid \mathbf{u}_2\right) d\mathbf{a}d\mathbf{e} \tag{6}$$

Assumption 3 (Partial Domain Invariance)**.** *Either the latent action $\mathbf{a}$ or the exogenous variable $\mathbf{e}$ is independent of the domain variable $\mathbf{u}$. That is, either*

$$p(\mathbf{a} \mid \mathbf{u}_1) = p(\mathbf{a} \mid \mathbf{u}_2) \quad \textit{or} \quad p(\mathbf{e} \mid \mathbf{u}_1) = p(\mathbf{e} \mid \mathbf{u}_2).$$

Assumption 4 (Recoverability of Observations)**.** *We assume that we can learn a conditional generative model that, given a domain variable $\mathbf{u}$, generates an estimated observation $\hat{\mathbf{o}}$ whose distribution matches the true conditional distribution of observations. That is,*

$$p(\hat{\mathbf{o}} \mid \mathbf{u}) = p(\mathbf{o} \mid \mathbf{u}),$$

where $\mathbf{u} \in \mathcal{U}$.

Proof. **Step 1.** Consider the generative process $\mathbf{o}_{t+1} = f(\mathbf{o}_t, \mathbf{a}_t, \mathbf{e}_t)$, and the learned estimator $\hat{f}$ such that $\hat{\mathbf{o}}_{t+1} = \hat{f}(\mathbf{o}_t, \hat{\mathbf{a}}_t, \hat{\mathbf{e}}_t)$.

According to Assumption 3, either the estimated latent action $\hat{\mathbf{a}}_t$ or the estimated exogenous variable $\hat{\mathbf{e}}_t$ is independent of the domain variable $\mathbf{u}$. Let $Z_{\text{inv}} \subseteq \mathcal{Z}_{\text{inv}}$ denote a measurable subset of the domain-invariant latent variable (i.e., either $\hat{\mathbf{a}}_t$ or $\hat{\mathbf{e}}_t$). Then, the independence implies that:

$$P\left[\hat{f}_{\text{inv}}^{-1}(\hat{\mathbf{o}}_{t+1}) \in Z_{\text{inv}} \,\middle|\, \mathbf{u}_1\right] = P\left[\hat{f}_{\text{inv}}^{-1}(\hat{\mathbf{o}}_{t+1}) \in Z_{\text{inv}} \,\middle|\, \mathbf{u}_2\right], \quad \forall\, \mathbf{u}_1, \mathbf{u}_2 \in \mathcal{U},$$

where $\hat{f}_{\text{inv}}^{-1}$ denotes the (estimated) inverse mapping from observation to the domain-invariant component of the latent space.

As this implies a symmetry between $\mathbf{a}$ and $\mathbf{e}$ with respect to their independence from the domain variable $\mathbf{u}$, we assume for simplicity that $\mathcal{Z}_{\text{inv}} = \mathcal{A}$, meaning that the action variable is independent of the domain.

Hence, we have the following:

$$P\left[\hat{f}_{a}^{-1}(\hat{\mathbf{o}}_{t+1}) \in A \,\middle|\, \mathbf{u}_1\right] = P\left[\hat{f}_{a}^{-1}(\hat{\mathbf{o}}_{t+1}) \in A \,\middle|\, \mathbf{u}_2\right], \quad \forall\, \mathbf{u}_1, \mathbf{u}_2 \in \mathcal{U}, \tag{7}$$

for any set $A \in \mathcal{A}$.

Since Assumption 4 ensures that the generative model matches the distribution of observations in each domain, we can replace $\hat{\mathbf{o}}_{t+1}$ with the true observation $\mathbf{o}_{t+1}$ in Eq. 7, yielding:

$$P\left[\hat{f}_a^{-1}(\mathbf{o}_{t+1}) \in A \,\middle|\, \mathbf{u}_1\right] = P\left[\hat{f}_a^{-1}(\mathbf{o}_{t+1}) \in A \,\middle|\, \mathbf{u}_2\right], \quad \forall\, \mathbf{u}_1, \mathbf{u}_2 \in \mathcal{U},$$

We then define a function $h = \hat{f}^{-1} \circ f$, assuming both f and $\hat{f}$ are injective and smooth. We have $h : \mathcal{Z} \to \mathcal{Z}$ is also assumed to be injective and smooth, which characterizes the mapping between the estimated and ground-truth latent variables, including actions and exogenous factors. That is:

$$(\hat{\mathbf{a}}_t, \hat{\mathbf{e}}_t) = h(\mathbf{a}_t, \mathbf{e}_t) = \big(h_a(\mathbf{a}_t, \mathbf{e}_t),\ h_e(\mathbf{a}_t, \mathbf{e}_t)\big),$$

which captures how the estimated latent variables are transformed into their ground-truth counterparts in a composed way. We then have:

$$P\left[h_a(\mathbf{a}_t, \mathbf{e}_t) \in A \,|\, \mathbf{u}_1\right] = P\left[h_a(\mathbf{a}_t, \mathbf{e}_t) \in A \,|\, \mathbf{u}_2\right], \quad \forall\, \mathbf{u}_1, \mathbf{u}_2 \in \mathcal{U},$$

which indicates that:

$$\int_{(\mathbf{a}_t, \mathbf{e}_t) \in h_a^{-1}(A)} p\left(\mathbf{a}_t, \mathbf{e}_t \mid \mathbf{u}_1\right) \mathbf{a}_t d\mathbf{e}_t = \int_{(\mathbf{a}_t, \mathbf{e}_t) \in h_a^{-1}(A)} p\left(\mathbf{a}_t, \mathbf{e}_t \mid \mathbf{u}_2\right) \mathbf{a}_t d\mathbf{e}_t \tag{8}$$

where $h_a^{-1}(A)$ is the pre-image of set A. Since at each domain u and at each time step t, $\mathbf{a}$ and $\mathbf{e}$ are indepdent, we then have

$$\int_{(\mathbf{a}_t, \mathbf{e}_t) \in h_a^{-1}(A)} p\left(\mathbf{e}_t \mid \mathbf{u}_1\right) p(\mathbf{a}_t) \mathbf{a}_t d\mathbf{e}_t = \int_{(\mathbf{a}_t, \mathbf{e}_t) \in h_a^{-1}(A)} p\left(\mathbf{e}_t \mid \mathbf{u}_2\right) p(\mathbf{a}_t) d\mathbf{a}_t d\mathbf{e}_t \tag{9}$$

Hence, in order to establish the independence between $\mathbf{e}$ and $\mathbf{a}$, we need to formally show that $\forall \mathbf{a} \in \mathcal{A}$, the preimage under h_a satisfies

$$h_a^{-1}(\mathbf{a}) = A \times \mathcal{E}, \quad \text{where } A \neq \emptyset \text{ and } A \subseteq \mathcal{A}.$$

This yields a mutually implicative condition: $\forall \mathbf{a} \in \mathcal{A}, r \in \mathbb{R}^+$, it follows that $h_a^{-1}\left(\mathcal{A}_r\left(\mathbf{a}\right)\right) = A^+ \times \mathcal{E}$, where $\mathcal{A}_r\left(\mathbf{a}\right) := \Big\{\mathbf{a}' \in \mathcal{A} : \|\mathbf{a}' - \mathbf{a}\|^2 < r\}$, $A^+ \neq \emptyset$, and $A^+ \subseteq \mathcal{A}$. Since h_a is continuous and $\mathcal{A}_r(\mathbf{a})$ is open, the preimage $h_a^{-1}\left(\mathcal{A}_r(\mathbf{a})\right)$ is also open. Moreover, under Assumption 3, which ensures that the generative model can recover the observational distribution, and given that h is assumed to be continuous and smooth, we conclude that h is bijective. Therefore, the preimage $h_a^{-1}\left(\mathcal{A}_r(\mathbf{a})\right)$ is non-empty.

Step 2 We prove the above condition by contradiction. Suppose that the preimage $h_a^{-1}\left(\mathcal{A}_r(\mathbf{a})\right)$ depends on both latent actions and exogenous variables. Then, we can partition this set into two parts: a set E_a that corresponds to variations in the exogenous variable, and the remaining part $h_a^{-1}\left(\mathcal{A}_r(\mathbf{a})\right) \setminus E_a$, which lies entirely within the action subspace. We have the conditions that: $\exists A* \in \mathcal{A}_r(\mathbf{a})$, $E_a = \{(\mathbf{a}, \mathbf{e}) \in \mathcal{Z} : (\mathbf{a}, \mathbf{e}) \in h_a^{-1}\left(\mathcal{A}_r(\mathbf{a})\right), h_a^{-1}\left(\mathcal{A}_r(\mathbf{a})\right) \setminus E_a \times \mathcal{E} \subsetneq h_a^{-1}\left(\mathcal{A}_r(\mathbf{a})\right)\} \neq \emptyset$. This implies that E_a is a subset where the preimage of the action does not span the full noise space $\mathcal{E}$, indicating that certain values of $\mathbf{e}$ are not accessible for a given $\mathbf{a}$. This, in turn, suggests a dependence between h_a and the exogenous variable $\mathbf{e}$.

Given Eq. 9, we have

$$\int_{(\mathbf{a}_t, \mathbf{e}_t) \in h_a^{-1}(A*)} p(\mathbf{a}_t)\big(p\left(\mathbf{e}_t \mid \mathbf{u}_1\right) - p\left(\mathbf{e}_t \mid \mathbf{u}_2\right)\big) d\mathbf{a}_t d\mathbf{e}_t = 0 \tag{10}$$

which can be further expressed as:

$$
\begin{aligned}
&\int_{(\mathbf{a}_t,\mathbf{e}_t)\in h_a^{-1}(A*)} p(\mathbf{a}_t)\big(p\left(\mathbf{e}_t \mid \mathbf{u}_1\right) - p\left(\mathbf{e}_t \mid \mathbf{u}_2\right)\big) d\mathbf{a}_t d\mathbf{e}_t \\
&= \underbrace{\int_{(\mathbf{a}_t,\mathbf{e}_t)\in h_a^{-1}(A*)\setminus E_a} p(\mathbf{a}_t)\big(p\left(\mathbf{e}_t \mid \mathbf{u}_1\right) - p\left(\mathbf{e}_t \mid \mathbf{u}_2\right)\big) d\mathbf{a}_t d\mathbf{e}_t}_{T_1} \\
&+ \underbrace{\int_{(\mathbf{a}_t,\mathbf{e}_t)\in E_a} p(\mathbf{a}_t)\big(p\left(\mathbf{e}_t \mid \mathbf{u}_1\right) - p\left(\mathbf{e}_t \mid \mathbf{u}_2\right)\big) d\mathbf{a}_t d\mathbf{e}_t}_{T_2} .
\end{aligned}
\tag{11}
$$

For T_1, since by definition, $h_a^{-1}(A*)\setminus E_a \in B_a \times \mathcal{E}$, where $B_a \in \mathcal{A}$, hence, $\int_{(\mathbf{a}_t,\mathbf{e}_t)\in h_a^{-1}(A*)\setminus E_a} p(\mathbf{a}_t)\big(p\left(\mathbf{e}_t \mid \mathbf{u}_1\right) - p\left(\mathbf{e}_t \mid \mathbf{u}_2\right)\big) d\mathbf{a}_t d\mathbf{e}_t = \int_{B_a} p(\mathbf{a}_t) \int_{\mathbf{e}\in\mathcal{E}} \big(p\left(\mathbf{e}_t \mid \mathbf{u}_1\right) - p\left(\mathbf{e}_t \mid \mathbf{u}_2\right)\big) d\mathbf{a}_t d\mathbf{e}_t = 0$. For T_2, since h_a^{-1} is open and non-empty, and because of the continuity and smoothness of h_a, for every $(\mathbf{a},\mathbf{e}) \in E_a$, there exists $r(\mathbf{a}) \in \mathbb{R}^+$ such that $\mathcal{A}_r(\mathbf{a}) \subseteq E_a$. Since $p(\mathbf{a}) > 0$ over $\mathcal{A}$, we have $p[(\mathbf{a}_t,\mathbf{e}_t) \in E_a \mid u] \geq p[(\mathbf{a}_t,\mathbf{e}_t) \in \mathcal{A}_r(\mathbf{a}) \times \mathcal{E} \mid u] > 0, \forall u \in \mathcal{U}$. Hence, based on the Domain-Induced Variability (Assumption 2), we have $T_2 \neq 0$. Therefore, we have $T_1 + T_2 \neq 0$, which contradicts Eq. 9.

We have proven that $\forall \mathbf{a} \in \mathcal{A}, r \in \mathbb{R}^+$, it follows that $h_a^{-1}\left(\mathcal{A}_r\left(\mathbf{a}\right)\right) = A^+ \times \mathcal{E}$, where $\mathcal{A}_r\left(\mathbf{a}\right) := \Big\{\mathbf{a}' \in \mathcal{A} : \|\mathbf{a}' - \mathbf{a}\|^2 < r\}, A^+ \neq \emptyset$, and $A^+ \subseteq \mathcal{A}$ is true. Hence, $h_a(\odot)$ does not depend on the exogenous noise variables $\mathbf{e}$, that is

$$
\frac{\partial \mathbf{e}_t}{\partial \hat{\mathbf{a}}_t} = 0.
$$

□

Theorem 2 (Permutation-Invaraint Identifiability of Latent Actions)**.** *The latent actions* $\mathbf{a}$ *are identified element-wise up to a permutation by the encoder function* $\hat{f}^{-1}$*, through an invertible mapping* h_a*. Specifically, we have*

$$
\hat{\mathbf{a}} = h_a(\pi(\mathbf{a})),
$$

where π *is an element-wise permutation function. This follows from Theorem 1, together with the following assumption on sparsity.*

Assumption 5 (Jacobian Sparsity)**.** *For all* $k \in \{1, \ldots, n\}$*, there exists a subset of indices* $\mathcal{C}_k$ *such that*

$$
\bigcap_{i\in\mathcal{C}_k} \left(\mathcal{F}_a\right)_{i,:} = \{k\},
$$

i.e., the k*-th latent variable is the only one influencing all observed dimensions indexed by* $\mathcal{C}_k$*, where* $\mathcal{F}_a$ *is the support of function* f *w.r.t. latent action* $\mathbf{a}$*.*

Proof. First, we use the chain rule to the composition of functions: $\hat{f} \odot h = f$, so that the derivates of f^{-1} w.r.t. $(\mathbf{a},\mathbf{e})$ as: $\mathbf{J}_{\hat{f}} = \mathbf{J}_f \mathbf{J}_{h^{-1}}$. The Jacobian matrix of h is:

$$
\mathbf{J}_h = \left[\begin{array}{c|c} \frac{\partial \hat{\mathbf{a}}}{\partial \mathbf{a}} & \frac{\partial \hat{\mathbf{a}}}{\partial \mathbf{e}} \\ \hline \frac{\partial \hat{\mathbf{e}}}{\partial \mathbf{a}} & \frac{\partial \hat{\mathbf{e}}}{\partial \mathbf{e}} \end{array}\right]
\tag{12}
$$

Since h^{-1} is invertible, the determinant of $\mathbf{J}_{h^{-1}}$ is non-zero. Also, since we prove that $\frac{\partial \hat{\mathbf{a}}}{\partial \mathbf{e}} = 0$. So in order to make $\mathbf{J}_{h^{-1}}$ is non-zero, then $\frac{\partial \hat{\mathbf{a}}}{\partial \mathbf{a}} \neq 0$ and $\frac{\partial \hat{\mathbf{e}}}{\partial \mathbf{e}} \neq 0$. Together with the condition that h is invertible, then h_a is invertible, meaning that there must be an invertible function that $\hat{\mathbf{a}} = h_a(\mathbf{a})$.

Then, based on Theorem 5 (mechanism sparsity) in Lachapelle et al. (2022), we can achieve element-wise identifiability of $\mathbf{a}$, where $\hat{\mathbf{a}} = h_a(\pi(\mathbf{a}))$, and π is a permutation. □

7 Baselines

Sparsity. The sparsity baselines aims to minimize the ℓ_1 norm of the latent actions without collapsing the ℓ_2 norm. Garrido et al. (2026a) implement introduce a sparsity regularization as follows, where $D \in \mathbb{R}$ is the dimension of the latent action and N is the size of the batch, and A represents the current batch in $\mathbb{R}^{N \times D}$:

$$\mathcal{L}_{\text{sparsity}} = VCM() + \frac{1}{N} \sum_i E(Z_i), \tag{13}$$

with

$$E(z) = \lambda_{l2} \max\left(\sqrt{D} - \|z\|_2^2, 0\right) + \lambda_{l1} \|z\|_1 \tag{14}$$

and

$$\begin{aligned} VCM(Z) = \lambda_V \frac{1}{D} \sum_d \max\left(1 - \sqrt{\mathrm{Var}(Z_{\cdot,d})}, 0\right) \\ + \lambda_C \frac{1}{D(D-1)} \sum_{i \neq j} \mathrm{Cov}(Z)_{i,j}^2 \\ + \lambda_M \frac{1}{ND} \sum_{i,j} Z_{i,j}. \end{aligned} \tag{15}$$

Each component of the VCM loss is weighted λ_V, λ_C, and λ_M. In practice we use the values prescibed by Gao et al. (2025) ($\lambda_{\ell_1} = 1$, $\lambda_{\ell_2} = 1$, $\lambda_V = 0.1$, $\lambda_C = 0.001$, and $\lambda_M = 0.1$).

Noise. We use KL divergence term to a gaussian prior to inject noise into the latent in a principled manner. The loss is as formulated as follows:

$$\mathcal{L}_{\text{noise}} = -\beta \, D_{KL} \left(q(a_l \mid z_t, z_{t+1}) \, \| \, \mathcal{N}(0, 1)\right) \tag{16}$$