Why Latent Actions Fail, and How to Prevent It

Jung Min Lee[1] **Taehyun Cho**[1,2,3] **Li Zhao**[4] **Jungwoo Lee**[1,5]
[1]Seoul National University [2]Vector Institute
[3]Fields Institute [4]Microsoft Research Asia [5]Hodoo AI Labs
{jmleeluck,talium,junglee}@snu.ac.kr lizo@microsoft.com

Abstract

Latent action models (LAMs) aim to learn manipulation priors from unlabeled videos by compressing frame-to-frame changes. The frames of in-the-wild videos, however, contain not only the agent's own state but *exogenous state* such as background clutter. Since the exogenous state introduces changes unrelated to actions, it hinders reliable latent action learning. This paper investigates this problem analytically by extending a linear LAM framework to explicitly model exogenous state. Our analysis reveals two insights: (1) minimizing the standard reconstruction objective produces latent actions that encode exogenous information from future observation; and (2) learning in a representation space that focuses on endogenous components is a key to mitigating the interference of noise. We further show that previously proposed auxiliary objectives, such as action-supervision, provably encourage latent actions to be consistent across exogenous states. These findings are validated through experiments on both linear and nonlinear LAMs, providing a unified theoretical analysis of how exogenous state hinders latent action learning and why common remedies work.

1 Introduction

Latent action models (LAMs) have emerged as a promising approach for learning action-relevant features from videos without action labels. This is particularly advantageous to real-world robotics, where collecting action-labeled datasets at scale is challenging [1]. Obtaining action annotations typically requires extensive human-expert teleoperation, careful curation, and repeated calibration, which creates a major bottleneck for dataset scaling. LAMs mitigate this limitation by learning from unlabeled videos and producing pseudo-action labels that supervise downstream models, such as vision-language-action models (VLA) [2–5] and world models [6, 7].

For pseudo-action supervision to be effective, latent actions must align with ground-truth actions even in the absence of action labels. This is nontrivial in in-the-wild videos, where observations often contain manipulation-irrelevant variation such as background motion and camera-induced changes. Prior works theoretically show that latent actions can spuriously encode such *exogenous noise* [8, 9]. For instance, latent actions trained on egocentric data tend to encode camera motion rather than physical change, misleading downstream models. Therefore, to improve action alignment, reducing exogenous noise is necessary. Existing approaches reduce it by introducing additional training objectives, such as a small amount of action (or optical-flow) supervision, language descriptions, or multi-view datasets [3, 5, 10, 11]. These methods augment the standard LAM objective with tractable auxiliary objectives to mitigate exogenous noise. However, why exogenous noise degrades latent actions and how auxiliary objectives mitigate it have not been analyzed in a unified manner.

In this work, we provide a unified theoretical analysis of how exogenous variation degrades latent action learning and why auxiliary training objectives mitigate it, building on the linear LAM framework. We first extend linear LAM to Exogenous Block Markov Decision Process (Ex-BMDP) to explicitly model a source of exogenous noise. Under this formulation, we show that training a LAM with a

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

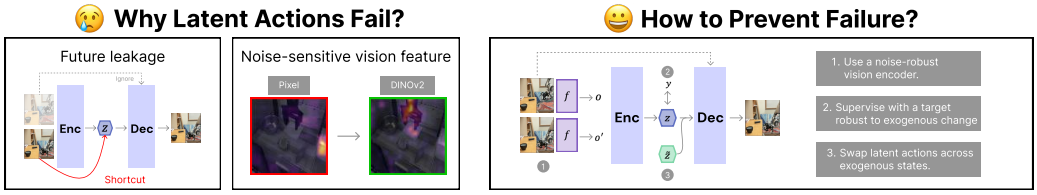


Figure 1: **Failure modes and remedies for latent action learning.** *(Left)* Future leakage and noise-sensitive visual features hinder the learning of action-centric latent representations. *(Right)* Our theoretical analysis explains why three remedies work: a noise-robust vision encoder, supervision with a robust target y, and latent action swapping across exogenous states.

reconstruction-based objective causes exogenous information from the future frame to leak into the latent action, reducing the capacity available for the action. We further analyze how the strength of exogenous noise behaves and explain why previously proposed auxiliary objectives improve action alignment. Our contributions are as follows.

1. We formulate and analyze latent action learning in Ex-BMDP, providing a framework that explicitly models the source of exogenous noise. Within this framework, we show that latent actions learned by reconstruction leak exogenous information from future frames, and that the energy of the resulting noise is governed by the sensitivity of the observation to exogenous change.
2. We provide a theoretical explanation for why previously proposed auxiliary objectives are effective under the Ex-BMDP framework, showing that they mitigate exogenous noise by encouraging consistency of latent actions across exogenous states.
3. We validate our analysis in controlled linear simulations and nonlinear LAMs. Additional experiments on real-world robot data and downstream control also support our analysis. These results support the applicability of our analysis beyond the linear setting.

2 Related Work

Latent Action Learning from Videos. Recent works have made substantial progress in learning latent actions from unlabeled videos. One line of research learns discrete latent actions using vector-quantized variational autoencoders (VQ-VAE) [12]. These discretized latent actions have been used to supervise VLA models [2, 3, 5] and world models [6] even when the ground-truth actions are not provided. In particular, latent actions have shown strong promise in robot manipulation, including cross-embodiment transfer [2] and the ability to scale VLA training with human videos in addition to robot datasets [3, 13]. While these approaches offer effective ways to incorporate latent actions into downstream embodied AI systems, they offer a limited understanding of when and why latent actions align with ground-truth actions.

Exogenous Noise in Latent Action Learning Exogenous noise causes frame transitions, but is independent of the agent's state and action. For instance, the motion of other agents or background movements are representative examples of it. Prior works show that such noise hinders reliable latent action learning [8–10]. Zhang et al. [8] show that LAMs may prioritize exogenous noise over action-induced changes when noise has larger variance. However, their formulation under controlled Markov process (CMP) does not explicitly represent exogenous noise and it is introduced only in the linear analysis as an additive residual in the observation transition. This leaves the open question about how exogenous noise should be formulated and analyzed beyond the additive linear setting.

To mitigate exogenous interference in practice, LAOM [10] and LAOF [11] supervise latent action learning with a small amount of action labels and optical flow, respectively. UniVLA [3] uses language descriptions to encode task-centric latent actions, while MVP-LAM [5] leverages multi-view datasets to encode consistent transitions across viewpoints, achieving higher mutual information with actions. Despite promising empirical results, the theoretical basis for why such auxiliary objectives alleviate exogenous interference remains an open question.

3 Setup

In this section, we first discuss the Exogenous Block Markov decision process (Ex-BMDP) proposed by Efroni et al. [14]. Ex-BMDP allows us to describe the general latent action learning in various video datasets. Then, we explain the standard latent action learning scenario in Ex-BMDP and extend the linear LAM setting to Ex-BMDP.

3.1 Exogenous Block Markov Decision Process

For a given set $\mathcal{U}$, we use $\Delta(\mathcal{U})$ to denote the set of all possible probability distributions over $\mathcal{U}$. We first consider the Block Markov Decision Process (BMDP) system. BMDP is specified by

$$\mathcal{M} = (\mathcal{O}, \mathcal{X}, \mathcal{A}, T, q, R, H, \mu) \tag{1}$$

where $\mathcal{O}$ is a set of observations, $\mathcal{X}$ is a set of latent states, $\mathcal{A}$ is a finite set of actions. We first start from initial distribution $\mu \in \Delta(\mathcal{X})$, and the agent interacts with the environment by repeatedly generating H-step trajectories $(x_1, o_1, a_1, r_1, \ldots, x_H, o_H, a_H, r_H)$ where the agent first receives observation $o_t \sim q(\cdot \mid x_t)$ from emission function $q : \mathcal{X} \to \Delta(\mathcal{O})$ and chooses actions $a_t \sim \pi(\cdot|o_t)$ using policy $\pi : \mathcal{O} \to \Delta(\mathcal{A})$. Then, the latent state evolves $x_{t+1} \sim T(\cdot \mid x_t, a_t)$ and the agent receives a reward $r_t = R(x_t, a_t)$.

Ex-BMDP is an extension of BMDP by assuming that $\mathcal{X} = \mathcal{S} \times \Xi$ where $\mathcal{S}$ is the set of (minimal) *endogenous states*, which is the (minimal) information that is necessary to control the agents and fully discards unnecessary information [10, 15], and Ξ is a set of *exogenous states*. For $x \in \mathcal{X}$, Ex-BMDP further assumes that the initial distribution and transition functions are decoupled, that is

$$\mu(x) = \mu_s(s)\mu_\xi(\xi), \qquad T(x'|x, a) = T_s(s'|s, a)T_\xi(\xi'|\xi) \tag{2}$$

where $\mu_s \in \Delta(\mathcal{S})$ and $T_s : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ are initial distribution and transition function over state space, while $\mu_\xi \in \Delta(\Xi)$ and $T_\xi : \Xi \to \Delta(\Xi)$ are those over the exogenous state space.

3.2 Standard LAM training

Ex-BMDP in latent action learning. The standard latent action learning scenario can be expressed under Ex-BMDP. Assume that the image is generated by a deterministic rendering function $I = g(s, \xi)$ where $g : \mathcal{S} \times \Xi \to \mathbb{R}^{C\times H\times W}$. To model the common LAM setting with a vision encoder, we further introduce an encoder $f(\cdot) : \mathbb{R}^{C\times H\times W} \to \mathcal{O}$, and define the observation as

$$o = f(I) = f(g(s, \xi)) \equiv h(s, \xi). \tag{3}$$

Here, $h : \mathcal{S} \times \Xi \to \mathcal{O}$ denotes the deterministic emission map from (s, ξ) to $\mathcal{O}$. This formulation also includes the pixel-space setting as a special case, where f is the identity map and $\mathcal{O} = \mathbb{R}^{C\times H\times W}$. We additionally assume the data-collecting policy is an *endogenous policy*, i.e., $\pi(a|o) = \pi(a|s)$ [9, 15]. This ensures that the action a does not depend on exogenous variation given the endogenous state.

Latent action learning. Given a dataset $\mathcal{D}$ of consecutive observations, a latent action model learns a latent action from a pair of observations and predicts the next observation from the current observation and the inferred latent action. Concretely, the inverse dynamics model ψ_{IDM} infers a latent action z from (o, o'), and the forward dynamics model ψ_{FDM} reconstructs o' from (o, z)

$$z = \psi_{\mathrm{IDM}}(o, o'), \quad \hat{o}' = \psi_{\mathrm{FDM}}(o, z), \tag{4}$$

with the standard LAM objective

$$\mathcal{L}_{\mathrm{LAM}} = \min_{\psi_{\mathrm{IDM}}, \psi_{\mathrm{FDM}}} \mathbb{E}_{\mathcal{D}} \left[\|o' - \hat{o}'\|_2^2 \right]. \tag{5}$$

Both models are trained in an end-to-end manner with Equation (5).

Goal. Our goal is to learn latent actions from reward- and action-free data that are highly informative about ground-truth actions. Following Zhang et al. [8] and Lee et al. [5], we formalize this desideratum as maximizing the mutual information between the latent action z and the action a:

$$\max_{z \in \mathbb{R}^{d_z}} \mathcal{I}(z; a). \tag{6}$$

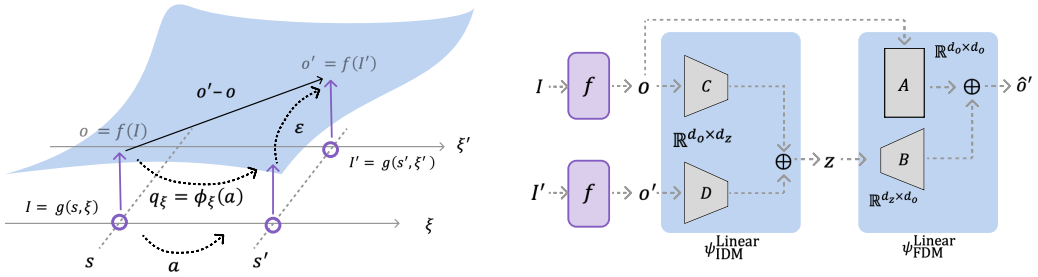


Figure 2: **Overview of Linear LAM.** *(Left)*: Observation transitions contain a state-driven component $q_\xi = \phi_\xi(a)$ and a ξ-driven exogenous component ε. *(Right)*: Architecture of linear LAM with a pretrained vision encoder f.

3.3 Linear LAM setting

The linear LAM framework provides theoretical simplicity to analyze latent actions [8]. Consider observations $o \in \mathbb{R}^{d_o}$ and the action $a \in \mathbb{R}^{d_a}$ which has lower dimensionality ($d_a \ll d_o$). For theoretical simplicity, linear LAM assumes that the forward dynamics are additive in observation space, consisting of the observation o, a controllable change $q_\xi = \phi_\xi(a)$ where $\phi_\xi : \mathbb{R}^{d_a} \to \mathbb{R}^{d_o}$ is the action-effect function, and exogenous noise ε.

$$o' = o + q_\xi + \varepsilon \tag{7}$$

Also, linear LAM assumes that ψ_{IDM} and ψ_{FDM} are implemented as linear maps

$$z = \psi^{\text{Linear}}_{\text{IDM}}(o, o') = Co + Do', \tag{8}$$

$$\hat{o}' = \psi^{\text{Linear}}_{\text{FDM}}(o, z) = Ao + Bz, \tag{9}$$

where $z \in \mathbb{R}^{d_z}$ with $d_z \ll d_o$. All parameters $A \in \mathbb{R}^{d_o \times d_o}$, $B \in \mathbb{R}^{d_o \times d_z}$, and $C, D \in \mathbb{R}^{d_z \times d_o}$ are learnable. We let $\theta = (A, B, C, D)$ for brevity. Figure 2 (Right) overviews this setup.

4 Theoretical Analysis and Experiments

We now study how the exogenous state affects latent action learning. We focus on the regime where the energy of ε is comparable to or exceeds that of q_ξ, the typical setting for latent action learning from in-the-wild videos. We first show that the reconstruction-based objective causes exogenous information from the future frame to leak into latent actions (Section 4.2). We then identify the visual representation as the dominant factor controlling the energy of exogenous noise (Section 4.3). Finally, we analyze two families of auxiliary objectives proposed in prior work and show that both enforce consistency of latent actions across exogenous states (Section 4.4 and 4.5).

Linear LAM experiment. Before our theoretical analysis, we first describe the linear LAM experimental setup used to empirically validate the results throughout this section. We first construct synthetic trajectories. The state evolves as $s_{t+1} = s_t + a_t$, where $a_t \sim \mathcal{N}(0, I)$. For the exogenous state, we consider n_ξ number of exogenous states and they switch with probability $p_{\text{switch}} \in [0, 1]$ by uniformly choosing from the other $n_\xi - 1$ exogenous states. Observations are generated by a ξ-dependent linear map $o = h(s, \xi) = H_\xi s$ where $H_\xi = H_0 + \alpha R_\xi$ with random matrices R_ξ and positive constant $\alpha \in \mathbb{R}_{\geq 0}$. To make R_ξ, we first initialize n_ξ different random matrices and choose one of them for each ξ. Note that H_0 is fixed across the exogenous state ξ.

To evaluate the action-alignment of latent actions, we train a linear probe [16] with a simple mean squared error (MSE) objective and measure the normalized mean squared error (NMSE) of the linear probe. This measures *how much action-relevant information is included in latent actions* and it is widely used to evaluate the quality of latent action [5, 10]. This evaluation metric serves as a proxy for Equation (6), since for jointly Gaussian z and a, $-\log \mathbb{E}[\|\hat{a}(z) - a\|^2]$ lower-bounds $\mathcal{I}(z; a)$ up

to a constant, where $\hat{a}(z)$ is the least-squares estimate of a given z. Here, we freeze the LAM and evaluate the quality of latent actions using a linear map $\hat{a}(z) = Mz + b$ where $M \in \mathbb{R}^{d_a \times d_z}, b \in \mathbb{R}^{d_a}$ are trainable. For each configuration, we report the mean and standard deviation across 4 seeds. We provide additional details of linear LAM experiments in Appendix C.

4.1 Additive dynamics in Ex-BMDP

For theoretical tractability, we introduce the following assumptions on the Ex-BMDP, under which additive dynamics in Equation 7 are naturally derived.

Assumption 4.1. *Suppose we have the dataset $\mathcal{D}$ generated by an Ex-BMDP with deterministic emission $h : \mathcal{S} \times \Xi \to \mathcal{O}$ and deterministic endogenous transition $s' = T_s(s, a)$. We assume that for all $s \in \mathcal{S}$, $\xi \in \Xi$, and $a \in \mathcal{A}$, there exists a function $\phi_\xi : \mathcal{A} \to \mathcal{O}$ such that*

$$\phi_\xi(a) = h(s', \xi) - h(s, \xi), \quad \textit{where } s' = T_s(s, a). \tag{10}$$

The exogenous noise ε arises from variation of the exogenous state at a fixed endogenous state:

$$\varepsilon = h(s', \xi') - h(s', \xi). \tag{11}$$

Justification of Assumption 4.1. The key idea is to decompose the observation change into two components according to its source: changes caused by the agent's action are captured through the endogenous transition under a fixed exogenous state (Equation (10)), while changes unrelated to the action are attributed to exogenous state variation under a fixed endogenous state (Equation (11)). Since we prefer additive dynamics for theoretical simplicity, we express both components as simple differences in observation space. Although this decomposition is derived under the linear setting, the underlying principle—that action-relevant and action-irrelevant changes are observed through distinct state components—is not specific to linearity and is expected to hold more broadly.

With Assumption 4.1, we have

$$\begin{aligned}
\phi_\xi(a) + \varepsilon &= \cancel{h(s', \xi)} - h(s, \xi) + h(s', \xi') - \cancel{h(s', \xi)} \\
&= h(s', \xi') - h(s, \xi) \\
&= o' - o
\end{aligned}$$

Therefore, linear LAM is induced by Ex-BMDP under Assumption 4.1. Figure 2 (Left) summarizes this decomposition.

4.2 Analysis 1: Future exogenous state leaks into latent action

A central question is what information the latent action z encodes while minimizing $\mathcal{L}_{\mathrm{LAM}}$. Previous work shows that z captures both controllable changes and exogenous noise [8]. Ideally, z should encode only the action-induced change between o and o', and any factor irrelevant to the action reduces z's capacity for action-relevant information. In this section, we revisit this issue under the Ex-BMDP framework. The key insight is that ψ_{FDM} must reconstruct o' from (o, z) while ξ' is observable only through o', so any ξ'-dependent information must be carried by z. We formalize this leakage below.

Proposition 4.2. *Suppose exogenous noise ε is non-degenerate in the observation space, i.e., its covariance with respect to ξ' is positive definite. Then, no global minimizer of $\mathcal{L}_{\mathrm{LAM}}$ in linear LAM admits a latent action that is ξ'-independent.*

See Appendix A.1 for the proof. The non-degeneracy condition is natural in our setting, as large-scale in-the-wild videos typically contain diverse exogenous factors. Under this condition, Proposition 4.2 formalizes the *future leakage* empirically observed in recent LAMs [11, 17, 18], where z absorbs information from the future frame instead of encoding the transition dynamics. A ξ'-dependent z drives this leakage by allowing ψ_{FDM} to take a shortcut, copying the future frame rather than predicting it from the action. This, in turn, leaves z uninformative about what has changed due to the action. Appendix B.1 further derives a lower bound of leakage under weight-decay scenario.

Experimental Settings. Motivated by Garrido et al. [19], we measure future leakage by changing the exogenous state of the next observation when encoding latent actions. Ideally, $\psi_{\mathrm{FDM}}(o, z)$

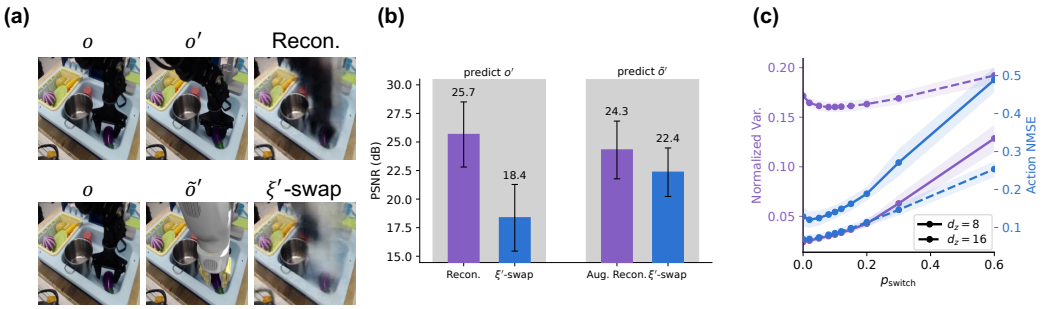


Figure 3: **Future exogenous state leaks into latent actions.** **(a)** Predicted next observation under the `Recon.` and ξ'-`swap`. **(b)** PSNR of three settings. ξ'-`swap` drops sharply when targeting o' but matches `Aug. Recon.` when targeting $\tilde{o}'$. **(c)** Normalized variance $\mathrm{Var}_{\xi'}(z|s,\xi,a)/\|z\|^2$ and action NMSE as functions of p_{switch}. Both metrics rise together as p_{switch} grows, indicating that z encodes ξ' at the cost of action capacity.

should predict the next observation of o. However, if LAM encodes the future observation, it would reconstruct that observation, even when it differs from the actual next observation of o. To validate this in practice, we use LAPA [2] evaluated on Bridge V2 [20] with paired transitions from OXE-AugE [21], which augments the original trajectories in Open X-Embodiment [22] with simulated robot embodiments while preserving the underlying action. Here, $o = h(s,\xi)$ and $\tilde{o} = h(s,\tilde{\xi})$ share the same state s but differ in the exogenous state $\xi \neq \tilde{\xi}$ (different embodiments). We compare three cases: (i) `Recon.`: $\psi_{\text{FDM}}(o,z)$ with $z = \psi_{\text{IDM}}(o,o')$, (ii) `Aug. Recon.`: $\psi_{\text{FDM}}(\tilde{o},\tilde{z})$ with $\tilde{z} = \psi_{\text{IDM}}(\tilde{o},\tilde{o}')$, and (iii) ξ'-`swap`: $\psi_{\text{FDM}}(o, z_{\text{cross}})$ with $z_{\text{cross}} = \psi_{\text{IDM}}(o,\tilde{o}')$. If z_{cross} did not encode $\tilde{\xi}'$ and ψ_{FDM} did not rely on the shortcut through z, $\psi_{\text{FDM}}(o, z_{\text{cross}})$ would predict o', not $\tilde{o}'$.

Results. As Figure 3(a) shows, ξ'-`swap` outputs a frame resembling $\tilde{o}'$ rather than o', despite being conditioned on o. Quantitatively, as shown in Figure 3(b), ξ'-`swap` drops 7.3 dB from `Recon.` in predicting o' but only 1.9 dB from `Aug. Recon.` in predicting $\tilde{o}'$. The 1.9 dB gap is comparable to the 1.4 dB drop from `Recon.` to `Aug. Recon.`, showing that the degradation under ξ'-`swap` is not attributable to distribution shift from the augmentation. This confirms that LAM encodes the next exogenous state as part of z, which ψ_{FDM} then uses as a visual shortcut by copying the future frame into its prediction. This in turn incentivizes z to encode even more future information during training. In Figure 3(c), we train a linear LAM with varying p_{switch} and report the normalized variance $\mathrm{Var}_{\xi'}(z|s,\xi,a)/\|z\|^2$ alongside action NMSE. Both metrics grow with p_{switch}, confirming that latent actions encode ξ' at the expense of action-relevant capacity. This trend is consistent with theoretical result in Appendix B.1.

4.3 Analysis 2: Exogenous variation and sensitivity of vision encoder

Given prior theoretical results that linear LAM behaves like principal component analysis (PCA) [8], we now analyze *which factors determine the energy of exogenous noise*.

Remark 4.3. *For any fixed $s' \in \mathcal{S}$, the conditional second moment of ε decomposes as*

$$\mathbb{E}\big[\|\varepsilon\|_2^2 \mid s'\big] = \underbrace{\mathbb{P}[\xi' \neq \xi \mid s']}_{\text{frequency of } \xi\text{-changes}} \underbrace{\mathbb{E}\big[\|h(s',\xi') - h(s',\xi)\|_2^2 \mid s',\, \xi' \neq \xi\big]}_{\text{sensitivity of the vision encoder to } \xi\text{-change}}. \tag{12}$$

Equation (12) decomposes exogenous energy into two factors: (i) ξ-switching frequency and (ii) representation sensitivity to exogenous state at a fixed state. Since the ξ-switching frequency $\mathbb{P}[\xi' \neq \xi|s']$ is defined by the environment itself, we focus on the representation sensitivity to exogenous factors. Define the ξ-sensitivity δ_h of emission function h as follows.

$$\delta_h(s) = \sup_{\xi,\tilde{\xi}\in\Xi} \|h(s,\tilde{\xi}) - h(s,\xi)\|^2 \tag{13}$$

Then, the second moment of ε satisfies

$$\mathbb{E}[\|\varepsilon\|^2|s'] \leq \mathbb{P}[\xi' \neq \xi|s']\delta_h(s') \tag{14}$$

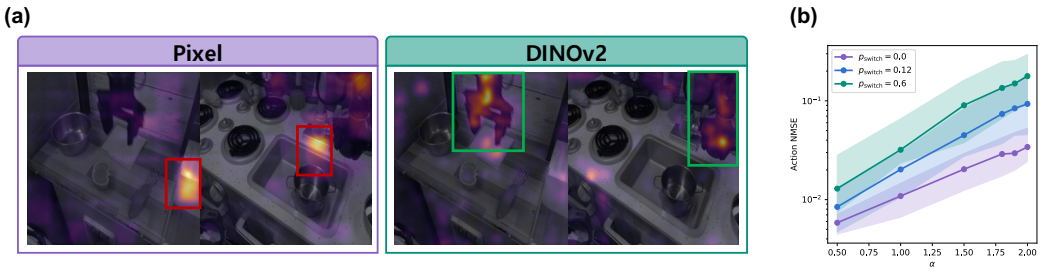


Figure 4: **Sensitivity of vision encoder to exogenous variation affects action alignment. (a)** Attention maps on Bridge V2 for LAMs trained with pixel (left) versus DINOv2 observations (right). DINOv2 attends to manipulation-relevant regions (green boxes), while pixel-based LAMs often attend to background or manipulation-irrelevant factors (red boxes). **(b)** In the linear LAM, action NMSE increases with α and with higher p_{switch}.

As Equation (14) indicates, reducing the ξ-sensitivity of the vision encoder reduces the energy of exogenous noise. We empirically verify this in Figure 4. As Figure 4(a) shows, DINOv2 [23] often provides an object-centric representation that is more robust to ξ, and therefore LAM trained on DINOv2 focuses on manipulation-relevant parts. By contrast, LAM trained on raw images often fails to focus on manipulation-relevant features. Accordingly, in linear LAM, Figure 4(b) shows that action NMSE increases with the ξ-sensitivity of the vision encoder. Here, we use α as a proxy for ξ-sensitivity. These results highlight that the choice of vision encoder f is important for latent action learning, motivating the use of representations with low ξ-sensitivity rather than raw pixels [3, 18, 24–26].

4.4 Analysis 3: Exogenous-swap reconstruction learns CCA

For theoretical simplicity, we assume access to synchronized transition tuples (s, s') observed under two exogenous states ξ and $\tilde{\xi}$.

Definition 4.4. *Let $o = h(s, \xi)$ and $\tilde{o} = h(s, \tilde{\xi})$. We define the exogenous-swap reconstruction objective as follows:*

$$\mathcal{L}_{\text{X-exo}}(\theta) = \mathbb{E}[\|o' - \psi_{\text{FDM}}(o, \tilde{z})\|^2], \quad \tilde{z} = \psi_{\text{IDM}}(\tilde{o}, \tilde{o}'). \tag{15}$$

Under the linear LAM assumptions, the following proposition identifies the minimizer of this objective and clarifies its connection to canonical correlation analysis (CCA).

Proposition 4.5. *In the linear LAM setting, define $u = q_\xi + \varepsilon$ and $\tilde{u} = q_{\tilde{\xi}} + \tilde{\varepsilon}$ and suppose the observation is uncorrelated with both u and $\tilde{u}$. Then, the exogenous-swap reconstruction objective reduces to*

$$\mathcal{L}_{\text{X-exo}}(\theta) = \mathbb{E}\left[\|u - BD\tilde{u}\|^2\right]. \tag{16}$$

When u and $\tilde{u}$ are whitened, i.e., $\Sigma_{uu} = \Sigma_{\tilde{u}\tilde{u}} = I$, the optimal low-rank reconstruction is determined by the singular directions of $\Sigma_{u\tilde{u}}$, corresponding to the canonical directions of CCA.

See proof in Appendix A.2 and further analysis in Appendix B.2. Proposition 4.5 shows that, under whitening, minimizing the exogenous-swap reconstruction objective is equivalent to CCA, which keeps transition directions that carry large energy under both ξ and $\tilde{\xi}$ simultaneously. Transition components that have high energy under only a specific exogenous state are suppressed, while components that persist across exogenous states—which are more likely to be action-driven—are preserved. This result is consistent with MVP-LAM [5], which treats viewpoint as the exogenous state and uses an exogenous-swap reconstruction objective on multi-view data to obtain latent actions well-aligned with ground-truth actions. To investigate the effectiveness of $\mathcal{L}_{\text{X-exo}}$, we train linear LAM with $\mathcal{L}_{\text{LAM}} + \lambda\mathcal{L}_{\text{X-exo}}$ and compare it to linear LAM trained only with $\mathcal{L}_{\text{LAM}}$. Figure 5 (a) (left) shows that $\mathcal{L}_{\text{X-exo}}$ consistently improves linear probe action prediction.

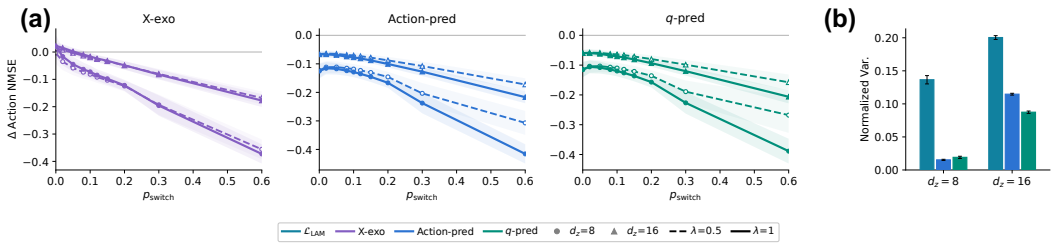


Figure 5: **Effect of $\mathcal{L}_{\text{X-exo}}$ and $\mathcal{L}_{\xi\text{-robust}}$. (a)** Δ action-validation NMSE relative to $\mathcal{L}_{\text{LAM}}$ (negative values indicate improvement) for three auxiliary training objectives: (left) X-exo, (mid) Action-pred, and (right) q-pred. **(b)** Normalized variance of latent action across exogenous states for different latent action dimensions d_z (lower is better), measuring how consistent the latent actions are. Shaded regions in (a) and error bars in (b) show the standard deviation across 4 random seeds.

4.5 Analysis 4: Supervision with targets robust to exogenous variation improves consistency

Prior work has shown that action prediction improves the quality of latent actions and, in turn, downstream performance [8, 10]. Since ground-truth action labels are rarely available at scale, a practical alternative is to predict optical flow obtained from pretrained estimators such as DPFlow [27], which requires no additional annotation. In this section, we unify these prediction targets under a single formulation.

Definition 4.6. *Let $y : \mathbb{R}^{d_o} \times \mathbb{R}^{d_o} \to \mathbb{R}^{d_y}$ be a target label of each observation pair. We say that y is η-robust to exogenous variation if, for any $(s, s') \in \mathcal{S}^2$ and any $\xi, \xi', \tilde{\xi}, \tilde{\xi}' \in \Xi$,*

$$\|y(o, o') - y(\tilde{o}, \tilde{o}')\|^2 \leq \eta, \tag{17}$$

where $o = h(s, \xi)$, $o' = h(s', \xi')$, $\tilde{o} = h(s, \tilde{\xi})$, and $\tilde{o}' = h(s', \tilde{\xi}')$. Typical examples include action labels and optical flow. Given such a target, we define the robust-target prediction objective

$$\mathcal{L}_{\xi\text{-robust}}(\theta_{\text{IDM}}, W; y) = \mathbb{E}\left[\|y(o, o') - W\psi_{\text{IDM}}(o, o')\|^2\right]. \tag{18}$$

where θ_{IDM} denotes the parameters of ψ_{IDM} ($\theta_{\text{IDM}} = (C, D)$ if linear LAM), and $W \in \mathbb{R}^{d_y \times d_z}$.

The following proposition shows that using a supervision target y with a small η improves the consistency of latent actions across exogenous factors.

Proposition 4.7. *In the linear LAM setting, the robust-target prediction objective with an η-robust prediction target y satisfies the following inequality:*

$$\mathbb{E}[\|W(\psi_{\text{IDM}}(o, o') - \psi_{\text{IDM}}(\tilde{o}, \tilde{o}'))\|^2] \leq 6\mathcal{L}_{\xi\text{-robust}}(\theta_{\text{IDM}}, W; y) + 3\eta \tag{19}$$

where $o = h(s, \xi)$ and $\tilde{o} = h(s, \tilde{\xi})$ for two distinct exogenous states $\xi \neq \tilde{\xi}$. In particular, if W has full column rank (which requires $d_y \geq d_z$), then the latent action $z = \psi_{\text{IDM}}(o, o')$ becomes consistent across exogenous states as $\mathcal{L}_{\xi\text{-robust}}$ and η decrease.

See proof in Appendix A.3. Proposition 4.7 shows that predicting an η-robust target yields an explicit bound on latent action mismatch across the exogenous state. Thus, minimizing $\mathcal{L}_{\xi\text{-robust}}$ with a smaller η encourages ψ_{IDM} to produce latent actions that are consistent across exogenous states.

To show how $\mathcal{L}_{\xi\text{-robust}}$ reduces the effect of exogenous states, we train a linear LAM with $\mathcal{L}_{\text{LAM}} + \lambda\mathcal{L}_{\xi\text{-robust}}$. For an η-robust target, we used the ground-truth action a and controllable changes q_ξ. Figure 5 (a) (middle and right) shows that using action a and controllable change q_ξ as prediction targets y (denoted Action-pred and q-pred, respectively) consistently improves action alignment. In addition, Figure 5 (b) shows the normalized variance $\text{Var}_{\xi,\xi'}(z|s, a)/\|z\|^2$ for the different latent action dimenstions d_z. Adding $\mathcal{L}_{\xi\text{-robust}}$ reduces this variance, indicating that the learned latent actions are more consistent across exogenous states.

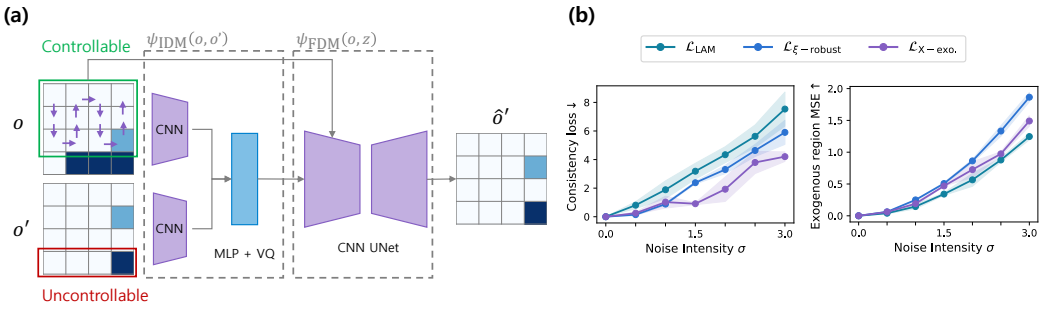


Figure 6: **Practical LAM. (a)** Architecture of practical LAM. ψ_{IDM} is implemented with CNN and MLP layers with vector quantization (VQ), while ψ_{FDM} uses a UNet architecture with CNN layers. **(b)** Result of practical LAM. *(Left)*: Consistency loss; lower is better. *(Right)*: Exogenous region MSE; higher is better. Shaded regions show standard deviation across 3 random seeds.

5 Practical LAM

Our theoretical analyses rely on Assumption 4.1, which is introduced for theoretical tractability. To investigate whether our theoretical insights extend beyond the linear setting, we conduct experiments with *non-additive transition dynamics* and *nonlinear LAMs*.

Dataset. Motivated by Zhang et al. [8], we construct a 4×4 grid-world dataset to evaluate practical LAM in a controlled nonlinear setting. The top three rows contain a controllable square that moves according to one of four actions (left, right, up, down), while the bottom row contains pixel values drawn independently as $\text{Bernoulli}(0.5) \cdot \sigma$ at each transition to mimic exogenous states, where σ controls the noise intensity. We collect transitions by following a deterministic policy that traces a snaking path through the controllable region (purple arrow in Figure 6(a)), so that the action is fully determined by the current observation. This induces a strong correlation between o and a. Furthermore, the pixel-level multiplicative exogenous noise breaks Assumption 4.1 on additive transition dynamics. We sweep σ to cover different noise regimes.

Model. Following common practice in latent action learning [2, 28], we implement ψ_{IDM} and ψ_{FDM} as a VQ-VAE, where ψ_{IDM} consists of CNN and MLP layers followed by vector quantization, and ψ_{FDM} is a CNN-based UNet (Figure 6(a)). We provide implementation and training details of ψ_{IDM} and ψ_{FDM} in Appendix C. We compare a LAM trained with $\mathcal{L}_{\text{LAM}}$ against LAMs trained with $\mathcal{L}_{\text{LAM}} + \lambda\mathcal{L}_{\text{X-exo}}$ and $\mathcal{L}_{\text{LAM}} + \lambda\mathcal{L}_{\xi\text{-robust}}$. For $\mathcal{L}_{\text{X-exo}}$, we sample paired observations under two distinct exogenous states at the same noise intensity. For $\mathcal{L}_{\xi\text{-robust}}$, we use the ground-truth action a as the η-robust target.

Evaluation. To investigate whether our theoretical analysis carries over to the general latent action learning, we measure how much latent actions are corrupted by exogenous states using two metrics: (i) the pixel MSE of the uncontrollable region (*Exogenous region MSE*), and (ii) the latent action MSE across exogenous states (*Consistency loss*). A higher Exogenous region MSE means the latent action does not predict the exogenous region (and thus does not encode it), while a lower Consistency loss means latent actions are consistent across exogenous states. We report the mean and standard deviation across 3 random seeds for each configuration.

Result. Figure 6 (b) reports the Consistency loss (left) and Exogenous region MSE (right) as we sweep the noise intensity σ. At $\sigma = 0$, where no exogenous noise is present, all three LAMs perform comparably. As σ increases, the vanilla LAM trained with $\mathcal{L}_{\text{LAM}}$ progressively encodes exogenous information into the latent action, yielding a low Exogenous region MSE and a high Consistency loss.

In contrast, augmenting $\mathcal{L}_{\text{LAM}}$ with either $\mathcal{L}_{\xi\text{-robust}}$ or $\mathcal{L}_{\text{X-exo}}$ keeps the Exogenous region MSE high and the Consistency loss low across all noise levels. This indicates that auxiliary objectives in the practical LAM induce latent actions that remain consistent across exogenous variations and do not predict the exogenous region. These results show that the auxiliary objectives improve the consistency of latent actions even when Assumption 4.1 does not hold.

Additional experiments. To test whether the remedies also improve action prediction and control, we conduct additional experiments with nonlinear LAMs (Appendix D). On the Distracting Control Suite [29], the remedies improve downstream control performance (Appendix D.1). On the real-world Bridge V2 dataset [20], they reduce future leakage and improve action alignment (Appendix D.2). These results are consistent with our analysis of linear LAMs and show that the same remedies are also effective for practical nonlinear LAMs.

6 Conclusion

Limitations and Future work. A limitation of our work is that our theoretical results rely on a linear LAM framework that has a gap to practical LAM. An important direction for future work is to relax these assumptions. Another promising direction is to extend our analysis to downstream policy learning, particularly VLA that leverage latent actions pretrained on in-the-wild videos.

In this paper, we presented a unified theoretical analysis of LAMs under the Ex-BMDP setting. Our analysis covers two failure modes of the standard latent action learning. First, the reconstruction objective induces latent actions that encode exogenous state from future observations, which is empirically known as future leakage. Second, the effect of exogenous noise grows when the observation space is sensitive to the exogenous state. To address these failure modes, we analyzed two families of auxiliary objectives—exogenous-swap reconstruction and robust-target prediction— and proved that both enforce consistency of latent actions across exogenous states, thereby improving action alignment. Experiments with linear and nonlinear LAMs support our analysis and indicate that these insights extend beyond the linear setting.

Acknowledgments

We thank Yerin Kim for supporting the visualizations in this work. This work is in part supported by the National Research Foundation of Korea (NRF, RS-2024-00451435(20%), RS-2024-00413957(20%)), Institute of Information & communications Technology Planning & Evaluation (IITP, RS-2025-02305453(15%), RS-2025-02273157(15%), RS-2025-25442149(15%), RS-2021-II211343(15%)) grant funded by the Ministry of Science and ICT (MSIT), Institute of New Media and Communications(INMAC), and the BK21 FOUR program of the Education, Artificial Intelligence Graduate School Program (Seoul National University), and Research Program for Future ICT Pioneers, Seoul National University in 2026.

References

[1] Robert McCarthy, Daniel C. H. Tan, Dominik Schmidt, Fernando Acero, Nathan Herr, Yilun Du, Thomas G. Thuruthel, and Zhibin Li. Towards generalist robot learning from internet video: A survey, 2024. URL `https://arxiv.org/abs/2404.19664`.

[2] Seonghyeon Ye, Joel Jang, Byeongguk Jeon, Se June 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. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=VYOe2eBQeh`.

[3] Qingwen Bu, Yanting Yang, Jisong Cai, Shenyuan Gao, Guanghui Ren, Maoqing Yao, Ping Luo, and Hongyang Li. Univla: Learning to act anywhere with task-centric latent actions. *arXiv preprint arXiv:2505.06111*, 2025.

[4] Xiaoyu Chen, Hangxing Wei, Pushi Zhang, Chuheng Zhang, Kaixin Wang, Yanjiang Guo, Rushuai Yang, Yucen Wang, Xinquan Xiao, Li Zhao, Jianyu Chen, and Jiang Bian. villa-x: Enhancing latent action modeling in vision-language-action models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL `https://openreview.net/forum?id=y5CaJb17Fn`.

[5] Jung Min Lee, Dohyeok Lee, Seokhun Ju, Taehyun Cho, Jin Woo Koo, Li Zhao, Sangwoo Hong, and Jungwoo Lee. MVP-LAM: Learning action-centric latent action via cross-viewpoint

reconstruction. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=r0SYyjDhxY.

[6] Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Bechtle, Feryal Behbahani, Stephanie Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando de Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative interactive environments, 2024. URL https://arxiv.org/abs/2402.15391.

[7] Shenyuan Gao, Siyuan Zhou, Yilun Du, Jun Zhang, and Chuang Gan. Adaworld: Learning adaptable world models with latent actions. In *International Conference on Machine Learning (ICML)*, 2025.

[8] Chuheng Zhang, Tim Pearce, Pushi Zhang, Kaixin Wang, Xiaoyu Chen, Wei Shen, Li Zhao, and Jiang Bian. What do latent action models actually learn? In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=DQMjemrVhe.

[9] 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.

[10] Alexander Nikulin, Ilya Zisman, Denis Tarasov, Lyubaykin Nikita, Andrei Polubarov, Igor Kiselev, and Vladislav Kurenkov. Latent action learning requires supervision in the presence of distractors. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=2gcEQCT7QW.

[11] Xizhou Bu, Jiexi Lyu, Fulei Sun, Ruichen Yang, Zhiqiang Ma, and Wei Li. Laof: Robust latent action learning with optical flow constraints. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27334–27344, June 2026.

[12] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6309–6318, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

[13] Xiaoyu Chen, Junliang Guo, Tianyu He, Chuheng Zhang, Pushi Zhang, Derek Cathera Yang, Li Zhao, and Jiang Bian. Igor: Image-goal representations are the atomic control units for foundation model in embodied ai. *arXiv preprint arXiv:2411.00785*, 2024. URL https://arxiv.org/abs/2411.00785.

[14] 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.

[15] Alex Lamb, Riashat Islam, Yonathan Efroni, Aniket Didolkar, Dipendra Misra, Dylan Foster, Olalekan Ogunmolu, Rajan Chari, Akshay Krishnamurthy, and John Langford. Guaranteed discovery of control-endogenous latent states with multi-step inverse models. *Transactions on Machine Learning Research*, February 2023. URL https://www.microsoft.com/en-us/research/publication/guaranteed-discovery-of-control-endogenous-latent-states-with-multi-step-inverse-models/.

[16] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes, 2017. URL https://openreview.net/forum?id=ryF7rTqgl.

[17] Jingwen Sun, Wenyao Zhang, Zekun Qi, Shaojie Ren, Zezhi Liu, Hanxin Zhu, Guangzhong Sun, Xin Jin, and Zhibo Chen. Vla-jepa: Enhancing vision-language-action model with latent world model, 2026. URL https://arxiv.org/abs/2602.10098.

[18] Jiange Yang, Yansong Shi, Haoyi Zhu, Mingyu Liu, Kaijing Ma, Yating Wang, Gangshan Wu, Tong He, and Limin Wang. Como: Learning continuous latent motion from internet videos for scalable robot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 42352–42363, June 2026.

[19] 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*, 2026.

[20] Homer 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 (CoRL)*, 2023.

[21] Guanhua Ji, Harsha Polavaram, Lawrence Yunliang Chen, Sandeep Bajamahal, Zehan Ma, Simeon Adebola, Chenfeng Xu, and Ken Goldberg. OXE-auge: A large-scale robot augmentation of OXE for scaling cross-embodiment policy learning. In *Forty-third International Conference on Machine Learning*, 2026. URL `https://openreview.net/forum?id=LcswwEzzX7`.

[22] Open X-Embodiment Collaboration, Abby O'Neill, Abdul Rehman, Abhinav Gupta, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, Albert Tung, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anchit Gupta, Andrew Wang, Andrey Kolobov, Anikait Singh, Animesh Garg, Aniruddha Kembhavi, Annie Xie, Anthony Brohan, Antonin Raffin, Archit Sharma, Arefeh Yavary, Arhan Jain, Ashwin Balakrishna, Ayzaan Wahid, Ben Burgess-Limerick, Beomjoon Kim, Bernhard Schölkopf, Blake Wulfe, Brian Ichter, Cewu Lu, Charles Xu, Charlotte Le, Chelsea Finn, Chen Wang, Chenfeng Xu, Cheng Chi, Chenguang Huang, Christine Chan, Christopher Agia, Chuer Pan, Chuyuan Fu, Coline Devin, Danfei Xu, Daniel Morton, Danny Driess, Daphne Chen, Deepak Pathak, Dhruv Shah, Dieter Büchler, Dinesh Jayaraman, Dmitry Kalashnikov, Dorsa Sadigh, Edward Johns, Ethan Foster, Fangchen Liu, Federico Ceola, Fei Xia, Feiyu Zhao, Felipe Vieira Frujeri, Freek Stulp, Gaoyue Zhou, Gaurav S. Sukhatme, Gautam Salhotra, Ge Yan, Gilbert Feng, Giulio Schiavi, Glen Berseth, Gregory Kahn, Guangwen Yang, Guanzhi Wang, Hao Su, Hao-Shu Fang, Haochen Shi, Henghui Bao, Heni Ben Amor, Henrik I Christensen, Hiroki Furuta, Homanga Bharadhwaj, Homer Walke, Hongjie Fang, Huy Ha, Igor Mordatch, Ilija Radosavovic, Isabel Leal, Jacky Liang, Jad Abou-Chakra, Jaehyung Kim, Jaimyn Drake, Jan Peters, Jan Schneider, Jasmine Hsu, Jay Vakil, Jeannette Bohg, Jeffrey Bingham, Jeffrey Wu, Jensen Gao, Jiaheng Hu, Jiajun Wu, Jialin Wu, Jiankai Sun, Jianlan Luo, Jiayuan Gu, Jie Tan, Jihoon Oh, Jimmy Wu, Jingpei Lu, Jingyun Yang, Jitendra Malik, João Silvério, Joey Hejna, Jonathan Booher, Jonathan Tompson, Jonathan Yang, Jordi Salvador, Joseph J. Lim, Junhyek Han, Kaiyuan Wang, Kanishka Rao, Karl Pertsch, Karol Hausman, Keegan Go, Keerthana Gopalakrishnan, Ken Goldberg, Kendra Byrne, Kenneth Oslund, Kento Kawaharazuka, Kevin Black, Kevin Lin, Kevin Zhang, Kiana Ehsani, Kiran Lekkala, Kirsty Ellis, Krishan Rana, Krishnan Srinivasan, Kuan Fang, Kunal Pratap Singh, Kuo-Hao Zeng, Kyle Hatch, Kyle Hsu, Laurent Itti, Lawrence Yunliang Chen, Lerrel Pinto, Li Fei-Fei, Liam Tan, Linxi "Jim" Fan, Lionel Ott, Lisa Lee, Luca Weihs, Magnum Chen, Marion Lepert, Marius Memmel, Masayoshi Tomizuka, Masha Itkina, Mateo Guaman Castro, Max Spero, Maximilian Du, Michael Ahn, Michael C. Yip, Mingtong Zhang, Mingyu Ding, Minho Heo, Mohan Kumar Srirama, Mohit Sharma, Moo Jin Kim, Muhammad Zubair Irshad, Naoaki Kanazawa, Nicklas Hansen, Nicolas Heess, Nikhil J Joshi, Niko Suenderhauf, Ning Liu, Norman Di Palo, Nur Muhammad Mahi Shafiullah, Oier Mees, Oliver Kroemer, Osbert Bastani, Pannag R Sanketi, Patrick "Tree" Miller, Patrick Yin, Paul Wohlhart, Peng Xu, Peter David Fagan, Peter Mitrano, Pierre Sermanet, Pieter Abbeel, Priya Sundaresan, Qiuyu Chen, Quan Vuong, Rafael Rafailov, Ran Tian, Ria Doshi, Roberto Mart'in-Mart'in, Rohan Baijal, Rosario Scalise, Rose Hendrix, Roy Lin, Runjia Qian, Ruohan Zhang, Russell Mendonca, Rutav Shah, Ryan Hoque, Ryan Julian, Samuel Bustamante, Sean Kirmani, Sergey Levine, Shan Lin, Sherry Moore, Shikhar Bahl, Shivin Dass, Shubham Sonawani, Shubham Tulsiani, Shuran Song, Sichun Xu, Siddhant Haldar, Siddharth Karamcheti, Simeon Adebola, Simon Guist, Soroush Nasiriany, Stefan Schaal, Stefan Welker, Stephen Tian, Subramanian Ramamoorthy, Sudeep Dasari, Suneel Belkhale, Sungjae Park, Suraj Nair, Suvir Mirchandani, Takayuki Osa, Tanmay Gupta, Tatsuya Harada,

Tatsuya Matsushima, Ted Xiao, Thomas Kollar, Tianhe Yu, Tianli Ding, Todor Davchev, Tony Z. Zhao, Travis Armstrong, Trevor Darrell, Trinity Chung, Vidhi Jain, Vikash Kumar, Vincent Vanhoucke, Vitor Guizilini, Wei Zhan, Wenxuan Zhou, Wolfram Burgard, Xi Chen, Xiangyu Chen, Xiaolong Wang, Xinghao Zhu, Xinyang Geng, Xiyuan Liu, Xu Liangwei, Xuanlin Li, Yansong Pang, Yao Lu, Yecheng Jason Ma, Yejin Kim, Yevgen Chebotar, Yifan Zhou, Yifeng Zhu, Yilin Wu, Ying Xu, Yixuan Wang, Yonatan Bisk, Yongqiang Dou, Yoonyoung Cho, Youngwoon Lee, Yuchen Cui, Yue Cao, Yueh-Hua Wu, Yujin Tang, Yuke Zhu, Yunchu Zhang, Yunfan Jiang, Yunshuang Li, Yunzhu Li, Yusuke Iwasawa, Yutaka Matsuo, Zehan Ma, Zhuo Xu, Zichen Jeff Cui, Zichen Zhang, Zipeng Fu, and Zipeng Lin. Open X-Embodiment: Robotic learning datasets and RT-X models. `https://arxiv.org/abs/2310.08864`, 2023.

[23] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL `https://openreview.net/forum?id=a68SUt6zFt`. Featured Certification.

[24] Taekyung Kim, Jeongeun Park, Sangdoo Yun, Dongyoon Han, and Byeongho Heo. Self-supervised visual state representation learning for robotics from dynamic scenes. In *7th Robot Learning Workshop: Towards Robots with Human-Level Abilities*, 2025. URL `https://openreview.net/forum?id=bEM2WGagcJ`.

[25] Mingyu Liu, Jiuhe Shu, Hui Chen, Zeju Li, Canyu Zhao, Jiange Yang, Shenyuan Gao, Hao Chen, and Chunhua Shen. Stamo: Unsupervised learning of generalizable robot motion from compact state representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 35014–35024, June 2026.

[26] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009, June 2022.

[27] Henrique Morimitsu, Xiaobin Zhu, Roberto M. Cesar, Xiangyang Ji, and Xu-Cheng Yin. Dpflow: Adaptive optical flow estimation with a dual-pyramid framework. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17810–17820, 2025. doi: 10.1109/CVPR52734.2025.01659.

[28] Dominik Schmidt and Minqi Jiang. Learning to act without actions. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

[29] 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.

[30] Saran Tunyasuvunakool, Alistair Muldal, Yotam Doron, Siqi Liu, Steven Bohez, Josh Merel, Tom Erez, Timothy Lillicrap, Nicolas Heess, and Yuval Tassa. dm_control: Software and tasks for continuous control. *Software Impacts*, 6:100022, 2020. ISSN 2665-9638. URL `https://www.sciencedirect.com/science/article/pii/S2665963820300099`.

[31] Zichen Jeff Cui, Hengkai Pan, Aadhithya Iyer, Siddhant Haldar, and Lerrel Pinto. Dynamo: In-domain dynamics pretraining for visuo-motor control. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 33933–33961. Curran Associates, Inc., 2024. doi: 10.52202/079017-1069. URL `https://proceedings.neurips.cc/paper_files/paper/2024/file/3b8db54b629e00537b59cbc6612026d7-Paper-Conference.pdf`.

A Main Proofs

A.1 Proof of Proposition 4.2

Proof. Suppose $\theta^\star = (A^\star, B^\star, C^\star, D^\star)$ is a global minimizer of $\mathcal{L}_{\mathrm{LAM}}$ and that $z^\star = C^\star o + D^\star o'$ is ξ'-independent. Let Π denote the orthogonal projection onto $\mathrm{range}(B^\star)$, and let $B^{\star\dagger}$ denote the pseudo-inverse of $B^\star$, so that $B^\star B^{\star\dagger} = \Pi$.

Decomposing the residual $o' - A^\star o - B^\star z^\star$ along $\mathrm{range}(B^\star)$ and its orthogonal complement,

$$\mathcal{L}_{\mathrm{LAM}}(\theta^\star) = \mathbb{E}\Big[\|(I-\Pi)(o' - A^\star o)\|^2\Big] + \mathbb{E}\Big[\|\Pi o' - \Pi A^\star o - B^\star z^\star\|^2\Big], \tag{20}$$

where the cross term vanishes since $B^\star z^\star \in \mathrm{range}(B^\star)$. Since $o = h(s,\xi)$ and $z^\star$ are both ξ'-independent, $\Pi A^\star o + B^\star z^\star$ is a function of (s, ξ, a, s') alone—it does not depend on ξ'. Hence, using $\mathbb{E}\|X - c\|^2 \geq \mathrm{Var}(X)$ for any constant c,

$$\mathcal{L}_{\mathrm{LAM}}(\theta^\star) \geq \mathbb{E}\Big[\|(I-\Pi)(o' - A^\star o)\|^2\Big] + \mathbb{E}_{s,\xi,a,s'}[\mathrm{Var}_{\xi'}[\Pi o' \mid s, \xi, a, s']]\,. \tag{21}$$

Now consider $\tilde{\theta} = (A^\star, B^\star, \tilde{C}, \tilde{D})$ with $\tilde{C} = -B^{\star\dagger}A^\star$ and $\tilde{D} = B^{\star\dagger}$, so that $B^\star(\tilde{C}o + \tilde{D}o') = \Pi(o' - A^\star o)$. Substituting into the decomposition,

$$\mathcal{L}_{\mathrm{LAM}}(\tilde{\theta}) = \mathbb{E}\Big[\|(I-\Pi)(o' - A^\star o)\|^2\Big]\,. \tag{22}$$

Subtracting (22) from (21),

$$\mathcal{L}_{\mathrm{LAM}}(\theta^\star) - \mathcal{L}_{\mathrm{LAM}}(\tilde{\theta}) \geq \mathbb{E}_{s,\xi,a,s'}[\mathrm{Var}_{\xi'}[\Pi o' \mid s, \xi, a, s']] > 0, \tag{23}$$

where the strict inequality holds since the ξ'-dependent component of $h(s', \xi')$ has nonzero projection onto $\mathrm{range}(B^\star)$ (Section 4). This contradicts the global optimality of $\theta^\star$.

□

A.2 Proof of Proposition 4.5

Proof.

$$\begin{aligned} o' - \psi^{\mathrm{Linear}}_{\mathrm{FDM}}(o, \tilde{z}) &= o' - Ao - B(C\tilde{o} + D\tilde{o}') \\ &= o + u - Ao - BC\tilde{o} - BD(\tilde{o} + \tilde{u}) \\ &= (I - A)o - B(C + D)\tilde{o} + u - BD\tilde{u} \end{aligned}$$

Since $\mathbb{E}[ou^T] = \mathbb{E}[o\tilde{u}^T] = \mathbb{E}[\tilde{o}u^T] = \mathbb{E}[\tilde{o}\tilde{u}^T] = 0$, we have

$$\mathcal{L}_{\text{X-exo}} = \mathbb{E}[\|(I-A)o - B(C+D)\tilde{o}\|^2] + \mathbb{E}[\|u - BD\tilde{u}\|^2]$$

The choices $A = I$ and $C = -D$ minimize $\mathcal{L}_{\text{X-exo}}$. Then,

$$\mathcal{L}_{\text{X-exo}} = \mathbb{E}[\|u - BD\tilde{u}\|^2]$$

Let $P = BD$ and $\Sigma_{uu} = \Sigma_{\tilde{u}\tilde{u}} = I$. Then,

$$\begin{aligned} \mathbb{E}[\|u - P\tilde{u}\|^2] &= \mathrm{Tr}(\Sigma_{uu}) - 2\mathrm{Tr}(P\Sigma_{\tilde{u}u}) + \mathrm{Tr}(P\Sigma_{\tilde{u}\tilde{u}}P^T) \\ &= d_o - 2\mathrm{Tr}(P\Sigma_{\tilde{u}u}) + \mathrm{Tr}(PP^T) \\ &= \|P - \Sigma_{u\tilde{u}}\|_F^2 + \underbrace{d_o - \|\Sigma_{u\tilde{u}}\|_F^2}_{\text{constant}} \end{aligned}$$

Let $\Sigma_{u\tilde{u}} = \sum_{k=1}^{d_o} \sigma_k u_k v_k^T$ be the SVD. Recall that CCA finds pairs of directions (a_k, b_k) along which two random vectors are maximally correlated, subject to each pair being uncorrelated with all previous ones:

$$(a_k^\star, b_k^\star) = \underset{\|a_k\|=\|b_k\|=1}{\arg\max} \ \mathrm{corr}(a_k^T u,\ b_k^T \tilde{u}) \quad \text{s.t.} \quad a_k^T a_j^\star = b_k^T b_j^\star = 0, \quad j = 1, \ldots, k-1. \tag{24}$$

Since $\Sigma_{uu} = \Sigma_{\tilde{u}\tilde{u}} = I$, $\mathrm{corr}(a_k^T u, b_k^T \tilde{u}) = a_k^T \Sigma_{u\tilde{u}} b_k$ whose solution is $(a_k, b_k) = (u_k, v_k)$ with correlation σ_k. By the Eckart–Young theorem, minimizing $\|P - \Sigma_{u\tilde{u}}\|_F^2$ under $\mathrm{rank}(P) \le d_z$ also yields

$$P^\star = \sum_{k=1}^{d_z} \sigma_k u_k v_k^T, \tag{25}$$

which selects the same top-d_z directions $(u_k, v_k)_{k=1}^{d_z}$. Therefore, under whitening, minimizing $\mathcal{L}_{\text{X-exo}}$ recovers the top-d_z canonical directions of CCA between u and $\tilde{u}$. □

A.3 Proof of Proposition 4.7

Proof. For notational simplicity, let $\tilde{y} = y(\tilde{o}, \tilde{o}')$ and $\tilde{z} = \psi_{\text{IDM}}(\tilde{o}, \tilde{o}')$. Recall the elementary inequality

$$\|a + b + c\|_2^2 \le 3\big(\|a\|_2^2 + \|b\|_2^2 + \|c\|_2^2\big).$$

Since

$$W(z - \tilde{z}) = (\tilde{y} - W\tilde{z}) - (y - Wz) + (y - \tilde{y}),$$

it follows that

$$\mathbb{E}\big[\|W(z - \tilde{z})\|_2^2\big] \le 3\,\mathbb{E}\big[\|\tilde{y} - W\tilde{z}\|_2^2\big] + 3\,\mathbb{E}\big[\|y - Wz\|_2^2\big] + 3\,\mathbb{E}\big[\|y - \tilde{y}\|_2^2\big]. \tag{26}$$

By construction,

$$\mathcal{L}_{\xi\text{-robust}}(\theta_{\text{IDM}}, W; y) = \mathbb{E}\big[\|y - Wz\|_2^2\big] = \mathbb{E}\big[\|\tilde{y} - W\tilde{z}\|_2^2\big],$$

and since y is η-robust, $\mathbb{E}[\|y - \tilde{y}\|_2^2] \le \eta$. Combining these gives

$$\mathbb{E}\big[\|W(z - \tilde{z})\|_2^2\big] \le 6\,\mathcal{L}_{\xi\text{-robust}}(\theta_{\text{IDM}}, W; y) + 3\eta. \tag{27}$$

Now suppose W has full column rank, and let $\sigma_W > 0$ denote its smallest singular value. Then for any vector v, $\|Wv\|_2^2 \ge \sigma_W^2 \|v\|_2^2$, which together with (27) yields

$$\mathbb{E}\big[\|z - \tilde{z}\|_2^2\big] \le \frac{6\,\mathcal{L}_{\xi\text{-robust}}(\theta_{\text{IDM}}, W; y) + 3\eta}{\sigma_W^2}. \tag{28}$$

Hence, minimizing $\mathcal{L}_{\xi\text{-robust}}$ forces the latent action z to remain consistent across exogenous states, and the bound tightens as $\eta \to 0$. □

B Extended Theoretical Analysis and Discussion

B.1 Lower bound of leakage

We show that the lower bound of the future leakage $\mathrm{Var}_{\xi'}[z \mid s, \xi, a]$ is related to how large the exogenous noise is in the decoder's space. Let $z_0 = \tilde{C}o + \tilde{D}o'$. Since $B^\star z_0 = \Pi\,(o' - A^\star o)$, and only o' depends on ξ', the decoder leakage satisfies

$$\mathrm{Var}_{\xi'}[B^\star z_0 \mid s, \xi, a, s'] = \mathrm{Var}_{\xi'}[\Pi o' \mid s, \xi, a, s'] = \mathrm{Var}_{\xi'}[\Pi\varepsilon \mid s, \xi, a, s']\,.$$

So the decoder leakage, how much of the ξ' the latent action delivers to the decoder, is exactly the exogenous noise expressed in B's space. A lower bound on the latent action leakage $\mathrm{Var}_{\xi'}[z \mid s, \xi, a]$ is not well defined on its own, since the rescaling $B \to \lambda B$, $z \to z/\lambda$ leaves the decoder leakage unchanged while shrinking the latent one arbitrarily. We therefore add weight decay,

$$\mathcal{L}_{\text{LAM}}^{\lambda}(\theta) = \mathcal{L}_{\text{LAM}}(\theta) + \lambda\|\theta\|^2, \qquad \lambda \in (0, 1].$$

Let $\theta_\lambda^\star$ be its global optimizer. Comparing it with the closed-form decoder $\tilde{\theta}_\lambda = \big(A_\lambda^\star, B_\lambda^\star, -B_\lambda^{\star\dagger} A_\lambda^\star, B_\lambda^{\star\dagger}\big)$ gives $\mathcal{L}_{\text{LAM}}^{\lambda}(\theta_\lambda^\star) \le \mathcal{L}_{\text{LAM}}^{\lambda}(\tilde{\theta}_\lambda)$, hence

$$e_\lambda := \mathbb{E}\Big[\|\Pi_\lambda(o' - A_\lambda o) - B_\lambda z_\lambda\|^2\Big] \le \lambda\left(\|\tilde{\theta}_\lambda\|^2 - \|\theta_\lambda^\star\|^2\right), \qquad \Pi_\lambda = B_\lambda^\star B_\lambda^{\star\dagger}.$$

Applying Young's inequality $\|x + y\|^2 \le (1 + \eta)\|x\|^2 + (1 + 1/\eta)\|y\|^2$, which holds for every $\eta > 0$ and $\mathrm{Var}[X] \le \mathbb{E}[\|X\|^2]$,

$$\mathbb{E}[\mathrm{Var}_{\xi'}[\Pi_\lambda o' | s, \xi, a, s']] \le (1 + \eta)\mathbb{E}[\mathrm{Var}_{\xi'}[B_\lambda z_\lambda | s, \xi, a, s']] + \left(1 + \frac{1}{\eta}\right) e_\lambda, \quad \forall \eta > 0$$

Using $e_\lambda \le \lambda(\|\tilde{\theta}_\lambda\|^2 - \|\theta^\star_\lambda\|^2)$, we have

$$\mathbb{E}[\mathrm{Var}_{\xi'}[B_\lambda z_\lambda | s, \xi, a, s']] \ge \frac{1}{1+\eta}\mathbb{E}[\mathrm{Var}_{\xi'}[\Pi_\lambda \varepsilon | s, \xi, a, s']] - \frac{\lambda}{\eta}(\|\tilde{\theta}_\lambda\|^2 - \|\theta^\star_\lambda\|^2)$$

and using $\mathrm{Var}[Bz] \le \|B\|^2_{\mathrm{op}}\mathrm{Var}[z]$ yields the leakage lower bound

$$\mathbb{E}[\mathrm{Var}_{\xi'}[z_\lambda \mid s, \xi, a, s']] \ge \frac{1}{\|B\|^2_{\mathrm{op}}} \sup_{\eta>0}\left[\frac{1}{1+\eta}\mathbb{E}[\mathrm{Var}_{\xi'}[\Pi_\lambda \varepsilon \mid s, \xi, a, s']] - \frac{\lambda}{\eta}\left(\|\tilde{\theta}_\lambda\|^2 - \|\theta^\star_\lambda\|^2\right)\right] \ge 0.$$

Taking the supremum is legitimate because the bound holds for every $\eta > 0$, and it is non-negative since the right-hand side converges to zero as $\eta \to \infty$. In our linear LAM experiments, the exogenous noise is $\varepsilon = 0$ with probability $1 - p_{\mathrm{switch}}$ and $\varepsilon = \alpha\,(R_{\xi'} - R_\xi)\,s'$ with probability $p_{\mathrm{switch}}/(n_\xi - 1)$, so

$$\mathbb{E}[\mathrm{Var}_{\xi'}[\Pi_\lambda \varepsilon \mid s, \xi, a, s']] = \alpha^2\Big(p_{\mathrm{switch}}\, c_1 - p^2_{\mathrm{switch}}\, c_2\Big),$$

with $c_1 = \mathbb{E}\big[\mathbb{E}_{\xi'\neq\xi}[\|\Pi_\lambda(R_{\xi'} - R_\xi)\,s'\|^2]\big]$ and $c_2 = \mathbb{E}\big[\|\mathbb{E}_{\xi'\neq\xi}[\Pi_\lambda(R_{\xi'} - R_\xi)\,s']\|^2\big]$. Since $c_1 \ge c_2$ by Jensen and $p\,c_1 - p^2 c_2$ increases for $p \le 1/2$, this predicts the trend in Figure 3(c).

B.2 Proposition 4.5 without assumption

In Proposition 4.5, we introduce whitening and orthogonal assumption, i.e., $\Sigma_{uu} = \Sigma_{\tilde{u}\tilde{u}}$ and $\mathbb{E}[ou^T] = \mathbb{E}[o\tilde{u}^T] = 0$. In this section, we characterize the global optimum of $\mathcal{L}_{\mathrm{X-exo}}$ without either assumption. Let $v = [\tilde{o}, \tilde{u}]^T, G = A - I$, and $K = B[C + D, D]$. Then, exogenous-swap reconstruction objective reduces to

$$\mathcal{L}_{\mathrm{X-exo}} = \mathbb{E}[\|u - Go - Kv\|^2]. \tag{29}$$

For fixed K, the stationary condition in G is $\mathbb{E}[(u - Go - Kv)o^T] = 0$, giving

$$G^\star = (\Sigma_{uo} - K\Sigma_{vo})\Sigma^{-1}_{oo}. \tag{30}$$

Writing $u_{\perp o} = u - \Sigma_{uo}\Sigma^{-1}_{oo}o$ and $v_{\perp o}$ similarly, substituting back gives $u - G^\star o - Kv = u_{\perp o} - Kv_{\perp o}$, so the problem is exactly

$$\mathcal{L}_{\mathrm{X-exo}} = \mathbb{E}[\|u_{\perp o} - Kv_{\perp o}\|^2]. \tag{31}$$

The Ao pathway thus absorbs everything linearly predictable from o, and the bottleneck performs reduced-rank regression (RRR) on the residuals. As in Section A.2,

$$\mathbb{E}[\|u_{\perp o} - Kv_{\perp o}\|^2] = \|K\Sigma^{1/2}_{v_{\perp o}v_{\perp o}} - M\|^2_F + \mathrm{const.} \tag{32}$$

where $M = \Sigma_{u_{\perp o}v_{\perp o}}\Sigma^{-1/2}_{v_{\perp o}v_{\perp o}}$. Since multiplying by the invertible $\Sigma^{1/2}_{v_{\perp o}v_{\perp o}}$ leaves the rank unchanged, by Eckart-Young, the optimum is the best rank-d_z truncation $[\cdot]_{d_z}$ of M,

$$K^\star = [M]_{d_z}\Sigma^{-1/2}_{v_{\perp o}v_{\perp o}}, \quad A^\star = I + (\Sigma_{uo} - K^\star\Sigma_{vo})\Sigma^{-1}_{oo}. \tag{33}$$

If the orthogonal assumption is introduced, then $\Sigma_{u_{\perp o}v_{\perp o}} = [0,\ \Sigma_{u\tilde{u}}]$ and therefore $M = [0, \Sigma_{u\tilde{u}}\Sigma^{-1/2}_{\tilde{u}\tilde{u}}]$. The optimum is $A^\star = I$, $C^\star = -D^\star$ and

$$B^\star D^\star = [\Sigma_{u\tilde{u}}\Sigma^{-1/2}_{\tilde{u}\tilde{u}}]_{d_z}\Sigma^{-1/2}_{\tilde{u}\tilde{u}}. \tag{34}$$

C Experimental Details

We use a single RTX 4090 GPU for both linear and practical LAM experiments. Tables 1 and 2 show the hyperparameters of linear and practical LAM experiments, respectively.

D Additional Experiments

D.1 Distracting Control Suite

The grid-world dataset in Section 5 does not fully reflect the complexity of real-world video data. We adopt it primarily to validate our theoretical findings in a controlled nonlinear setting, which is difficult to isolate in real-world datasets. To partially close this gap, we additionally conduct practical LAM experiments on the Distracting Control Suite [29].

Table 1: **Hyperparameters for the linear LAM experiments.** We use the same training setup across all experiments.

Hyperparameters of linear LAM experiments	
d_s	8
d_a	8
d_o	128
num. traj. of train	8k
Batch size	128
Steps/Probe Steps	20k
Learning rate	1e-3

Table 2: **Hyperparameters for the practical LAM** Architecture (top) and training settings (bottom) for the practical LAM.

Hyperparameters of Practical LAM	
ψ_{IDM} conv. channels	[128, 128, 128]
ψ_{IDM} MLP hidden dim.	64
ψ_{FDM} conv. channels	[32, 32, 32, 32]
Latent action dim. (d_z)	32
Codebook size (K)	5
VQ commitment weight (β)	0.25
Optimizer	Adam
Learning rate	3e-4
Batch size	128
Total updates	16k
Gradient clip norm	5.0
$\lambda_{\text{X-exo}}$	1.0
λ_{act}	1.0
Action label fraction	1%

Dataset. We collect offline datasets from the DeepMind Control Suite [30], focusing on the `cheetah-run` environment. We use the pretrained PPO agent to collect 500 expert demonstrations. Note that the behavior of the PPO expert induces a strong correlation between o and a. To mimic in-the-wild videos, we further inject exogenous noise by perturbing the color and viewpoint of each observation. Specifically, we use $(\Delta_{\mathrm{color}}, \Delta_{\mathrm{cam}}) = (0.5, 0.05)$, where Δ_{color} controls the maximum per-channel color perturbation and Δ_{cam} scales the magnitude of camera pose and zoom randomization. Both color and camera parameters drift continuously across timesteps.

Model. We follow the architecture of LAPO [28] to implement the practical latent action model. Both ψ_{IDM} and ψ_{FDM} are implemented with nonlinear layers consisting of several convolution layers with GELU and Tanh activations. We remove the FSQ bottleneck used in the original implementation since it hindered overall performance in our experiments, and instead employ a strong information bottleneck ($d_z = 8$). To investigate the quality of latent actions beyond the linear probing, we additionally train a latent action policy $\pi(z|o)$ and an action decoder $\pi_{\mathrm{Dec}}(a|z)$ with a simple MSE objective. This protocol mirrors common scenarios in which latent actions are used to supervise VLA training. We implement $\pi(z|o)$ with convolution layers and a linear head with ReLU6 activation, while $\pi_{\mathrm{Dec}}(a|z)$ is a 3-layer MLP. We evaluate the learned policy and report the normalized return. As an upper bound, we also train a behavior-cloning policy $\pi_{\mathrm{BC}}(a|o)$ directly on the labeled dataset.

Training objectives. We compare the vanilla LAM trained with $\mathcal{L}_{\mathrm{LAM}}$ against the auxiliary objectives studied in Section 4: exogenous-swap reconstruction $\mathcal{L}_{\text{X-exo}}$, and robust-target prediction objective $\mathcal{L}_{\xi\text{-robust}}$ instantiated with action labels and optical flow as η-robust targets. In addition, motivated by Cui et al. [31], we consider a variant where a vision encoder $f(\cdot)$ is jointly trained with the LAM. For each image I, $f(\cdot)$ extracts a representation $o = f(I)$ that the LAM uses as its observation, and for the next observation $o' = f(I')$, $f(\cdot)$ is treated as stop-gradient and updated via

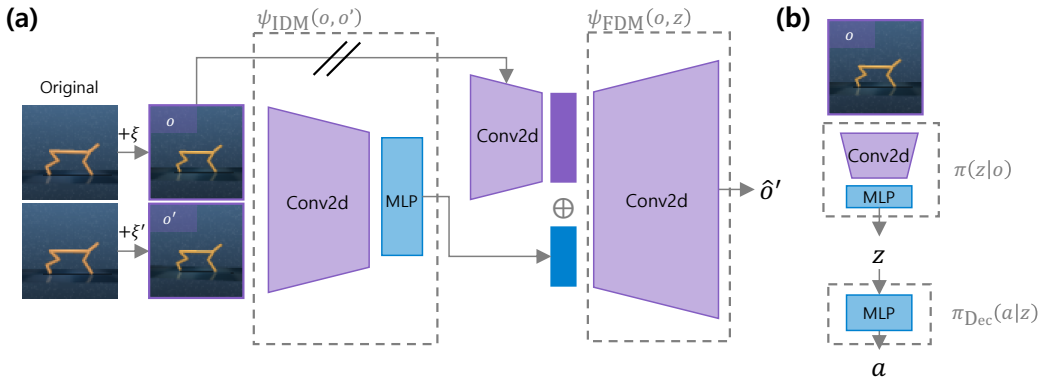


Figure 7: **Architecture of practical LAM on DCS.** (a) We first collect clean demonstrations from a PPO agent and inject exogenous noise into each observation. Both ψ_{IDM} and ψ_{FDM} are implemented with several convolution layers. (b) The latent action policy $\pi(z|o)$ predicts z from o using convolution layers, and $\pi_{\text{Dec}}(a|z)$ decodes the prediction of the latent action policy into a ground-truth action.

an exponential moving average (EMA) of the online encoder. Cui et al. [31] report that this representation improves downstream manipulation performance, suggesting it captures more endogenous information. Within our framework, this corresponds to $f(\cdot)$ providing a representation with low ξ-sensitivity δ_h, which is predicted to reduce exogenous noise energy and improve LAM training (Equation (14)).

Results. Results are reported in Table 3. The EMA-encoder variant ($+f(\cdot)$) substantially outperforms the vanilla LAM. In our framework, this can be interpreted as $f(\cdot)$ learning a representation with low ξ-sensitivity δ_h, which is predicted to improve LAM training (Equation (14)). The auxiliary training objectives ($\mathcal{L}_{\text{X-exo}}$, action prediction, and flow prediction) also yield consistent improvements over the vanilla LAM, in agreement with our theoretical predictions.

Table 3: **Normalized return across auxiliary objectives on `cheetah-run`.** Results are reported as $\text{mean}_{\pm\text{std}}$ over 100 episodes.

Metric	$\pi_{\text{BC}}(a\|o)$	$\mathcal{L}_{\text{LAM}}$	$+f(\cdot)$	$+\mathcal{L}_{\text{X-exo}}$	+Action-pred	+Flow-pred
Normalized Return	$0.18_{\pm 0.12}$	$0.05_{\pm 0.04}$	$0.15_{\pm 0.06}$	$0.16_{\pm 0.07}$	$0.15_{\pm 0.11}$	$0.18_{\pm 0.10}$
One-sided p	–	–	<0.001	<0.001	<0.001	<0.001

D.2 Validation in Real-world dataset

Table 4: Result of latent action learning across the methods in Bridge V2 dataset. Results are shown in $\text{mean}_{\pm\text{std}}$ across 5 training seeds.

Method	Trans. NMSE	PSNR$(\hat{o}', o')$ ↑	PSNR$(\hat{o}', \tilde{o}')$ ↓	ΔPSNR	95% CI
$\mathcal{L}_{\text{LAM}}$	$0.419_{\pm 0.007}$	$20.46_{\pm 0.22}$	$21.15_{\pm 0.23}$	$+0.69_{\pm 0.44}$	$[+0.150, +1.230]$
$+\lambda\mathcal{L}_{\xi\text{-robust}}$	$0.120_{\pm 0.004}$	$22.76_{\pm 0.11}$	$18.81_{\pm 0.05}$	$-3.95_{\pm 0.12}$	$[-4.094, -3.796]$
$+\mathcal{L}_{\text{X-exo}}$	$0.359_{\pm 0.019}$	$22.27_{\pm 0.08}$	$19.09_{\pm 0.10}$	$-3.18_{\pm 0.13}$	$[-3.343, -3.008]$
$+f_{\text{dino}}(\cdot)$	$0.315_{\pm 0.014}$	$22.56_{\pm 0.09}$	$19.20_{\pm 0.08}$	$-3.36_{\pm 0.08}$	$[-3.459, -3.262]$

Nonlinear LAM To validate our theoretical results on real-world data, we train LAPA's LAM from scratch on the Bridge V2 dataset. We train LAMs with $\mathcal{L}_{\text{LAM}}$, $\mathcal{L}_{\text{LAM}} + \lambda\mathcal{L}_{\xi\text{-robust}}$(action for η-robust target), and $\mathcal{L}_{\text{LAM}} + \mathcal{L}_{\text{X-exo}}$, and evaluate two quantities.

1. Translation NMSE: how well the latent action recovers the ground-truth end-effector translation with a single linear layer. We exclude the rotation and gripper axes because no model

predicts them better than the mean predictor, whose NMSE is approximately 1, at this scale, where every LAM is trained from scratch on a small dataset.

2. $\Delta\mathrm{PSNR}$: the future leakage diagnostic of Figure 3, evaluated on the ξ'-`swap`. We compute $\mathrm{PSNR}(\hat{o}', \tilde{o}') - \mathrm{PSNR}(\hat{o}', o')$, where $\hat{o}'$ is the output of ξ'-`swap`. A positive value means the latent action reproduces the swapped future, i.e., it leaks the future exogenous state.

We additionally include a frozen DINOv2 encoder $f_{\mathrm{dino}}(\cdot)$ as a low-sensitivity backbone, which tests the effect of encoder-sensitivity in reliable latent action learning. The training procedure is identical across all objectives, with the same architecture and hyperparameters, so that the objective (and the encoder initialization) is the only variable. Each model is trained for 10k steps and we report the early-stopped checkpoint selected by validation reconstruction loss.

Table 4 shows that the vanilla LAM leaks the future exogenous state, while the LAMs trained with the three remedies do not. This is significant under a one-sided t-test with $n = 5$, giving $p = 0.012$ for $\mathcal{L}_{\mathrm{LAM}}$ and $p < 0.001$ for remaining methods. Also, none of the four 95% confidence intervals contains zero.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Please refer to Section 1 and Abstract.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See Section 6.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: See Section 4 and Appendix A

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: See Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: See supplementary materials for linear LAM and pratical LAM experiments.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: See Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: See Figure 4, 5, and Table 3.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our work doesn't deviate from NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: Our work does not have any potential to negative social impacts.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: We don't release the data or models. Also, we used only open-source datasets and models.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification:

    (a) LAPO: github
    (b) LAOM: github

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: We don't release any asset in this work.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: We include neither crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: We include neither crowdsourcing nor research.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are not involved in this research.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.