# INTERSELF: LEARNING WORLD MODELS UNDER HIDDEN ACTION EXECUTION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

World models typically assume that the action selected by a policy is the action that causes the next transition. This assumption fails when an unobserved execution process repeats, drops, or remaps commands, causing the model to conflate execution dynamics with environmental dynamics. We formalize this mismatch as an action-fidelity problem and quantify the predictive information lost by conditioning only on intended actions. We introduce InterSelf, which augments DreamerV3 with a recurrent latent execution model combining categorical action remapping and temporal persistence. Its effective-action representation conditions both world-model learning and latent imagination, without executed-action labels or corruption identities. We show that this channel covers the evaluated execution failures, while its latent mixtures are identifiable only when different actions leave sufficiently separable effects. Across a 22-game Atari study with three training seeds, InterSelf achieves higher corrupted returns than DreamerV3 in 14 games; in seven action-sensitive games, it outperforms both DreamerV3 and the capacity-matched InterSelf-Identity control. Variation across the remaining games is consistent with the predicted boundary: explicit execution modeling need not help when action effects are weak or redundant. These results establish hidden action execution as a distinct source of world-model misspecification and provide a principled way to model it.

## 1 INTRODUCTION

Learned dynamics support planning, policy optimization, and latent imagination (Sutton, 1991; Ha & Schmidhuber, 2018; Hafner et al., 2020; Schrittwieser et al., 2020; Hafner et al., 2025). Their action input, however, carries a premise that is rarely isolated: the command chosen by the policy is the intervention that generated the observed transition. A sticky controller violates this premise by repeating an earlier action; a faulty actuator may drop or remap the command. When execution is hidden, the learner observes an intention and its consequence but not the intervening event. One conditional transition must then explain two processes—execution and environment dynamics—whose uncertainty has different structure.

This ambiguity is neither visual shift nor undifferentiated transition noise: it changes the semantics of the control variable. Conditioning on intention alone discards predictive information whenever alternative executions have different consequences. It also creates a training–imagination mismatch: the policy is optimized in latent rollouts that treat intentions as executed even though experience was generated through a hidden channel.

We study this action-fidelity failure in pixel-based world-model reinforcement learning. Our approach, INTERSELF, factors the controlled transition into a history-dependent execution process and external world dynamics. A recurrent execution state predicts a near-identity categorical remapping and a persistence probability. Their output is an effective-action distribution that enters the recurrent state-space model (RSSM) during both posterior learning and imagination. The environment's executed action and corruption mode are never used as inputs or targets.

The paper follows one claim through theory and experiment: latent execution is useful to the extent that its consequences are recoverable from interaction history. The information-gap result formalizes what is lost by omitting execution; a representability result connects the learned channel to common actuation failures; an identifiability result shows why benefits depend on the separability of action



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

effects; and a rollout-stability result quantifies how recovery error is amplified by transition sensitivity and imagination depth. This view also predicts a meaningful boundary: when several actions are transition-equivalent, explicit mediation cannot be reliably identified and the identity path may be preferable.

Our contributions are:

1. We formulate hidden action execution as misspecification of the action variable in a world model and quantify the resulting Bayes prediction gap.
2. We introduce a recurrent latent execution channel whose effective action is used consistently for representation learning and imagination, without supervision from the environment's private execution log.
3. We establish exact representations of the evaluated execution channels, give population conditions under which their mixtures are identifiable from action effects, and bound how recovery error propagates through finite-horizon latent imagination.
4. We evaluate 22 Atari games with three independent seeds and an exact capacity-matched identity-channel control, separating benefits of the execution pathway from those of additional representation capacity.

## 2 Related Work

**World-model reinforcement learning.** Model-based RL ranges from integrated planning and learning (Sutton, 1991) to action-conditioned video prediction (Oh et al., 2015) and compact latent dynamics (Ha & Schmidhuber, 2018; Hafner et al., 2019). Dreamer learns behavior through latent imagination (Hafner et al., 2020; 2021; 2025); MuZero and EfficientZero combine learned models with search (Schrittwieser et al., 2020; Ye et al., 2021); IRIS and TD-MPC offer transformer and trajectory-optimization alternatives (Micheli et al., 2023; Hansen et al., 2022; 2024). These methods parameterize controlled dynamics directly by the issued command. A sufficiently expressive recurrent latent model may absorb execution uncertainty into its hidden state, but it does not isolate the command–execution channel or expose an effective action to imagination. InterSelf introduces this factorization as an execution-aware inductive bias, designed to model structured execution uncertainty rather than to impose a task-independent decomposition of environment dynamics.

**Action uncertainty, persistence, and delay.** Robust MDPs optimize against transition uncertainty sets (Iyengar, 2005; Nilim & El Ghaoui, 2005); action-robust RL instead perturbs or adversarially mixes the policy action (Pinto et al., 2017; Tessler et al., 2019). These formulations target robust policy optimization under a specified uncertainty model, rather than inference of the realized execution inside a learned dynamics model. ALE protocols use sticky actions to prevent exploitation of deterministic emulation (Bellemare et al., 2013; Machado et al., 2018). Temporal abstraction and action-repetition methods let the *agent* choose how long an action persists (Sutton et al., 1999; Sharma et al., 2017; Lakshminarayanan et al., 2017). Delayed-RL methods address observations or commands shifted in time using augmented histories, correction, or imitation (Bouteiller et al., 2021; Liotet et al., 2022). Our setting is complementary: in the evaluated channels, persistence, drops, and remappings occur inside the action–transition interface, while their realized executions are not labeled. The goal is a history-dependent predictive action representation shared by posterior learning and imagination, rather than only a robust policy.

**Latent actions and identifiability.** Inverse-dynamics objectives learn controllable representations from temporal effects (Pathak et al., 2017), while latent-action methods learn structure within large or otherwise difficult action spaces (Chandak et al., 2019). Recent video models infer action codes when logged actions are absent (Bruce et al., 2024; Ye et al., 2025). InterSelf instead observes the intended-action vocabulary and infers how each command was executed. Because the issued action is observed, the identity channel is a nested special case, enabling a capacity-matched structural control. Classical mixture-identifiability theory studies conditions under which latent component parameters can be recovered from observed mixtures (Teicher, 1963; Allman et al., 2009). Our theoretical analysis focuses on a conditional identifiability setting: given a fixed dictionary of executed-action effects, it characterizes when the corresponding mixture weights are uniquely identifiable and stably recoverable.

# 3 PROBLEM FORMULATION

## 3.1 HIDDEN ACTION EXECUTION

Consider a partially observed controlled process (Kaelbling et al., 1998) with observations $O_t$, rewards $R_t$, and latent states $W_t$. Before the current command, the agent has history $\mathcal{H}_t = (O_{\leq t}, R_{\leq t}, A_{<t})$. The policy selects an *intended* action $A_t \in \{1, \ldots, K\}$, represented by $\mathbf{a}_t \in \Delta^{K-1}$, whereas the environment may apply an unobserved *executed* action

$$E_t \sim \kappa_\star(\cdot \mid A_t, \mathcal{H}_t). \tag{1}$$

Taking $W_t$ to render external dynamics Markov conditional on $E_t$, the observed controlled transition is

$$\begin{aligned} p_\star(W_{t+1} \mid W_t, A_t, \mathcal{H}_t) = \sum_{e=1}^{K} \kappa_\star(e \mid A_t, \mathcal{H}_t) \\ \cdot\, p_\star(W_{t+1} \mid W_t, E_t = e). \end{aligned} \tag{2}$$

Standard action-conditioned world models feed $A_t$ directly into the transition, effectively tying the applied action to the command. A sufficiently expressive recurrent model may learn the marginal law in Eq. (2), but it does not isolate an execution channel that can be carried into imagination. Persistence, dropped or remapped commands, and actuator failures therefore create a structural mismatch between policy intention and observed action effects.

## 3.2 INFORMATION LOST BY THE INTENDED-ACTION MODEL

Let $X_{t+1}$ be any next-step target, $S_t$ any representation available before the transition, and $\mathcal{R}^*(X \mid U)$ the minimum expected log loss over all predictors conditioned on $U$. The following identity (Cover & Thomas, 2006) quantifies the information available to an oracle that additionally observes execution.

**Theorem 1** (Oracle prediction gap under hidden execution)**.** *Under logarithmic loss,*

$$\begin{aligned} &\mathcal{R}^*(X_{t+1} \mid S_t, A_t) - \mathcal{R}^*(X_{t+1} \mid S_t, A_t, E_t) \\ &\quad = I(X_{t+1}; E_t \mid S_t, A_t) \geq 0. \end{aligned} \tag{3}$$

The gap is positive exactly when execution changes the predictive law beyond what $(S_t, A_t)$ explains. This oracle statement identifies available predictive information; it does not make $E_t$ observable or guarantee finite-sample recovery or improved return.

**Corollary 2** (Predictive value of an execution surrogate)**.** *Let $\widehat{E}_t$ be any additional representation constructed before $X_{t+1}$. If $X_{t+1} \perp \widehat{E}_t \mid (S_t, A_t, E_t)$, then*

$$\begin{aligned} &\mathcal{R}^*(X_{t+1} \mid S_t, A_t) - \mathcal{R}^*(X_{t+1} \mid S_t, A_t, \widehat{E}_t) \\ &= I(X_{t+1}; \widehat{E}_t \mid S_t, A_t) \leq I(X_{t+1}; E_t \mid S_t, A_t). \end{aligned} \tag{4}$$

A useful surrogate must therefore preserve execution-relevant history not already contained in $S_t$; recovering a private execution label is unnecessary. Section 5 states when the associated mixture weights are identifiable from action effects.

## 3.3 WORLD–EXECUTION FACTORIZATION

InterSelf separates an execution representation from external dynamics:

$$\underbrace{\widetilde{\mathbf{a}}_t = q_\psi(E_t = \cdot \mid A_t, S_t, M_t)}_{\text{execution representation}}, \qquad \underbrace{p_\phi(W_{t+1} \mid W_t, \widetilde{\mathbf{a}}_t)}_{\text{world dynamics}}, \tag{5}$$

where $M_t$ summarizes execution-relevant history and $\widetilde{\mathbf{a}}_t \in \Delta^{K-1}$ is the action representation supplied to the world model. The model receives neither $E_t$ nor the active execution mode during training.

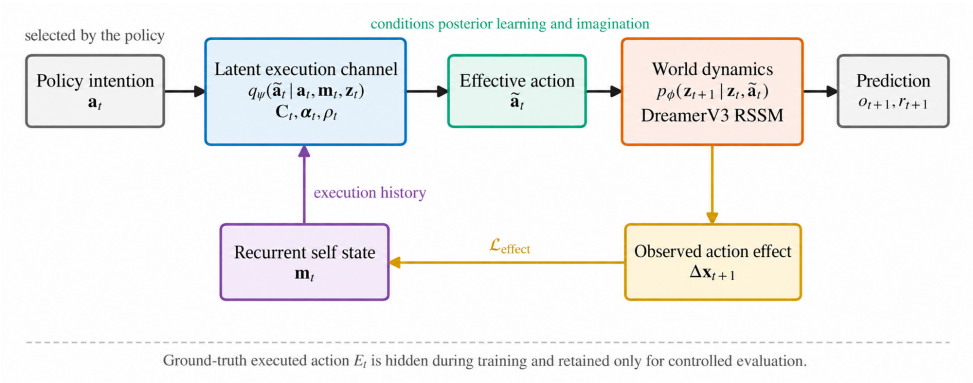



Figure 1: InterSelf separates policy intention from the action representation used by the world model. A recurrent row-stochastic channel produces an effective-action distribution shared by posterior learning and imagination; execution labels and active modes remain hidden.

# 4 INTERSELF

## 4.1 LATENT EXECUTION STATE AND ACTION CHANNEL

DreamerV3 (Hafner et al., 2025) represents history with a deterministic state $\mathbf{h}_t$ and a categorical stochastic state $\mathbf{z}_t$. Its recurrent state-space model (RSSM) applies the intended action directly:

$$\mathbf{h}_{t+1} = f_\phi(\mathbf{h}_t, \mathbf{z}_t, \mathbf{a}_t), \tag{6}$$

$$\widehat{\mathbf{z}}_{t+1} \sim p_\phi(\mathbf{z}_{t+1} \mid \mathbf{h}_{t+1}), \tag{7}$$

$$\mathbf{z}_{t+1} \sim q_\phi(\mathbf{z}_{t+1} \mid \mathbf{h}_{t+1}, \mathbf{x}_{t+1}), \tag{8}$$

where $\mathbf{x}_t = \text{enc}_\phi(O_t)$. InterSelf preserves the Dreamer world model and behavior learning objectives, but replaces the transition action with a learned effective-action representation.

InterSelf augments the RSSM carry with a recurrent execution state $\mathbf{m}_t$ and the previous effective-action distribution $\widetilde{\mathbf{a}}_{t-1}$:

$$\boldsymbol{\xi}_t = (\mathbf{h}_t, \mathbf{z}_t, \mathbf{m}_t, \widetilde{\mathbf{a}}_{t-1}). \tag{9}$$

The self state preserves execution-relevant history and influences downstream prediction and control only through the action supplied to the transition. From the pre-transition states, the execution model computes

$$\mathbf{u}_t = g_\psi([\mathbf{m}_t, \text{vec}(\mathbf{z}_t)]) \tag{10}$$

and predicts action-specific remapping gates and row-wise destinations:

$$\boldsymbol{\alpha}_t = \sigma(W_\alpha \mathbf{u}_t + b_\alpha + \alpha_0 \mathbf{1}), \tag{11}$$

$$\mathbf{D}_t[i,:] = \text{softmax}(\text{reshape}(W_D \mathbf{u}_t + b_D)[i,:]) \,. \tag{12}$$

These define a row-stochastic channel centered on identity execution:

$$\mathbf{C}_t = \text{Diag}(\mathbf{1} - \boldsymbol{\alpha}_t) + \text{Diag}(\boldsymbol{\alpha}_t)\mathbf{D}_t, \tag{13}$$

$$\mathbf{C}_t[i,j] = q_\psi(\widetilde{E}_t = j \mid A_t = i, \mathbf{m}_t, \mathbf{z}_t), \tag{14}$$

$$\overline{\mathbf{a}}_t = \mathbf{C}_t^\top \mathbf{a}_t. \tag{15}$$

Here $\widetilde{E}_t$ is the newly mapped action before persistence. The gate $\alpha_{t,i}$ interpolates between the identity row and a learned destination row; the actual off-diagonal mass is $\alpha_{t,i}(1 - \mathbf{D}_t[i,i])$.

To capture history-dependent persistence, InterSelf predicts

$$\rho_t = \sigma(w_\rho^\top \mathbf{u}_t + b_\rho + \rho_0), \qquad \widetilde{\mathbf{a}}_t = (1 - \rho_t)\overline{\mathbf{a}}_t + \rho_t \widetilde{\mathbf{a}}_{t-1}. \tag{16}$$

This is the marginal recursion of a categorical process that either applies a newly mapped action or carries forward the previous execution. Its unrolling forms a normalized temporal kernel over earlier mapped actions (Appendix A).

### 4.2 COUPLED REPRESENTATION LEARNING

The world transition consumes the effective-action distribution:

$$\mathbf{h}_{t+1} = f_\phi(\mathbf{h}_t, \mathbf{z}_t, \widetilde{\mathbf{a}}_t), \tag{17}$$

$$\widehat{\mathbf{z}}_{t+1} \sim p_\phi(\mathbf{z}_{t+1} \mid \mathbf{h}_{t+1}), \tag{18}$$

$$\mathbf{z}_{t+1} \sim q_\phi(\mathbf{z}_{t+1} \mid \mathbf{h}_{t+1}, \mathbf{x}_{t+1}). \tag{19}$$

Thus $\widetilde{\mathbf{a}}_t$ is a differentiable soft action interface, rather than an exact mixture evaluated through the nonlinear RSSM. Proposition 14 shows that this interface is exact under affine action dependence and bounds its nonlinear Jensen gap by transition curvature and mixture uncertainty. After posterior inference, the execution state is updated by

$$\mathbf{m}_{t+1} = G_\psi^{\text{self}}(\mathbf{m}_t, \mathbf{z}_{t+1}, \mathbf{a}_t, \widetilde{\mathbf{a}}_{t-1}, \widetilde{\mathbf{a}}_t). \tag{20}$$

The current channel depends only on pre-transition states; the posterior consequence updates $\mathbf{m}_{t+1}$ for the next step. The exact gated map is given in Appendix B.

Executed actions are never training targets. Instead, an auxiliary predictor ties the latent execution representation to the observed encoder displacement $\Delta\mathbf{x}_{t+1} = \mathbf{x}_{t+1} - \mathbf{x}_t$:

$$\widehat{\Delta\mathbf{x}}_{t+1} = g_\eta^{\text{eff}}([\mathbf{h}_{t+1}, \text{vec}(\mathbf{z}_{t+1}), \mathbf{m}_{t+1}, \widetilde{\mathbf{a}}_t]), \tag{21}$$

$$\mathcal{L}_{\text{eff},t+1} = \frac{1 - d_{t+1}}{d_x} \left\| \widehat{\Delta\mathbf{x}}_{t+1} - \text{sg}(\Delta\mathbf{x}_{t+1}) \right\|_2^2. \tag{22}$$

This is an effect-consistency signal, not supervised recovery of $E_t$.

The identity-centered regularizer is

$$\mathcal{L}_{\text{reg},t} = \frac{\|\mathbf{C}_t - I\|_F^2}{K^2} + \lambda_\alpha \frac{\|\boldsymbol{\alpha}_t\|_2^2}{K} + \lambda_\rho \rho_t^2. \tag{23}$$

With the standard DreamerV3 prediction, dynamics, and representation losses, the complete world-model objective is

$$\begin{aligned} \mathcal{L}_{\text{IS}} = \mathbb{E} \sum_t \big[ & \mathcal{L}_{\text{pred},t} + \mathcal{L}_{\text{dyn},t} + \lambda_{\text{rep}} \mathcal{L}_{\text{rep},t} \\ & + \lambda_{\text{eff}} \mathcal{L}_{\text{eff},t} + \lambda_{\text{reg}} \mathcal{L}_{\text{reg},t} \big]. \end{aligned} \tag{24}$$

As $\boldsymbol{\alpha}_t \to \mathbf{0}$ and $\rho_t \to 0$, the standard Dreamer action path is recovered.

### 4.3 CONSISTENT IMAGINATION AND CAPACITY CONTROL

The actor produces intended actions from the standard Dreamer feature, $A_t \sim \pi_\theta(A_t \mid \mathbf{h}_t, \mathbf{z}_t)$. Each imagined transition then applies the same execution channel used during posterior learning:

$$\mathbf{a}_t \xrightarrow{\mathcal{T}_\psi} \widetilde{\mathbf{a}}_t \xrightarrow{f_\phi} (\mathbf{h}_{t+1}, \widehat{\mathbf{z}}_{t+1}, \mathbf{m}_{t+1}). \tag{25}$$

The prior sample $\widehat{\mathbf{z}}_{t+1}$ replaces the posterior sample in Eq. (20); actor and critic optimization otherwise follows DreamerV3. Learning and imagination therefore use the same action semantics.

**Capacity-matched identity control.** INTERSELF-IDENTITY retains the execution state, auxiliary objectives, regularization, architecture, and training protocol, but supplies the intended action to the RSSM:

$$\mathbf{h}_{t+1}^{\text{ID}} = f_\phi(\mathbf{h}_t, \mathbf{z}_t, \mathbf{a}_t). \tag{26}$$

It is exactly parameter matched to InterSelf, so their comparison isolates whether the learned execution representation improves the world transition.

## 5 WHEN IS EXECUTION MODELING INFORMATIVE?

The oracle gap in Theorem 1 does not imply that a hidden channel is recoverable. We now connect the InterSelf parameterization to the evaluated interventions, characterize when their mixtures are recoverable from effects, and trace recovery error into latent imagination. Proofs are given in Appendix A.

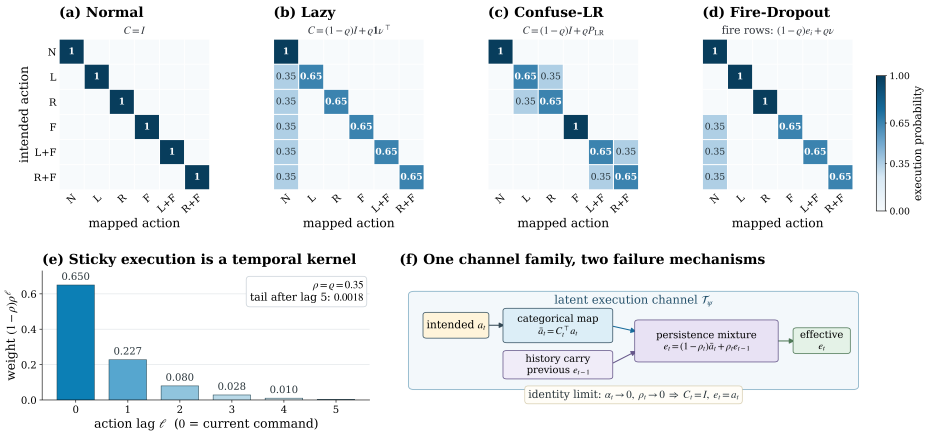



Figure 2: **Execution-channel coverage.** The categorical map represents substitution and dropout (top), while persistence produces a geometric temporal kernel (bottom). Both recover the identity action path.

### 5.1 Representing common execution failures

Let $\boldsymbol{\nu}$ be the one-hot no-op action, $P_{\mathrm{LR}}$ the permutation that exchanges left and right semantics, and $\boldsymbol{f} \in \{0,1\}^K$ the indicator of fire-containing actions.

**Theorem 3** (Coverage of the evaluated execution channels)**.** *Allowing boundary values of $\boldsymbol{\alpha}_t$ and $\rho_t$, the closure of the InterSelf family contains, for every $\varrho \in [0,1]$,*

$$C_{\mathrm{normal}} = I, \tag{27}$$

$$C_{\mathrm{lazy}} = (1-\varrho)I + \varrho \mathbf{1}\boldsymbol{\nu}^\top, \tag{28}$$

$$C_{\mathrm{LR}} = (1-\varrho)I + \varrho P_{\mathrm{LR}}, \tag{29}$$

$$C_{\mathrm{fire}} = I + \varrho \mathrm{Diag}(\boldsymbol{f})(\mathbf{1}\boldsymbol{\nu}^\top - I). \tag{30}$$

*Moreover, $C_t = I$ and $\rho_t = \varrho$ reproduce the sticky-action marginal*

$$q_t = (1-\varrho)\mathbf{a}_t + \varrho q_{t-1}. \tag{31}$$

Thus $C_t$ represents within-step failures and $\rho_t$ temporal persistence. This is a capacity result, not a recovery claim; Figure 2 visualizes the family at the evaluation value $\varrho = 0.35$.

### 5.2 Identifiability from action effects

Let $Y_{t+1} = \Phi(X_t, X_{t+1}) \in \mathbb{R}^d$ be an observable or latent action-effect statistic; the encoder displacement in Eq. (22) is one instance. At predictive state $s$, define

$$\boldsymbol{\mu}_e(s) = \mathbb{E}[Y_{t+1} \mid S_t = s, E_t = e], \qquad e \in \{1, \dots, K\}. \tag{32}$$

We assume *effect sufficiency*: for every supported $(a, h)$,

$$\mathbb{E}[Y_{t+1} \mid S_t = s, E_t = e, A_t = a, \mathcal{H}_t = h] = \boldsymbol{\mu}_e(s). \tag{33}$$

Writing $\boldsymbol{u} = (\kappa_1, \dots, \kappa_{K-1})^\top$ and $\kappa_K = 1 - \mathbf{1}^\top \boldsymbol{u}$, the conditional mean effect is

$$\boldsymbol{y}(s,a,h) = \boldsymbol{\mu}_K(s) + B_s \boldsymbol{u}(a,h), \quad B_s = \big[\boldsymbol{\mu}_1 - \boldsymbol{\mu}_K, \dots, \boldsymbol{\mu}_{K-1} - \boldsymbol{\mu}_K\big]. \tag{34}$$

**Theorem 4** (Identifiability and stability from action effects)**.** *Under Eq. (33) with fixed effect signatures, if* $\mathrm{rank}(B_s) = K-1$*, then $\kappa(\cdot \mid a, h)$ is unique and the least-squares estimate from any $\widehat{\boldsymbol{y}}$ satisfies*

$$\|\widehat{\boldsymbol{u}} - \boldsymbol{u}\|_2 \le \frac{\|\widehat{\boldsymbol{y}} - \boldsymbol{y}\|_2}{\sigma_{\min}(B_s)}. \tag{35}$$

*If* $\mathrm{rank}(B_s) < K-1$*, local uniqueness fails for interior $\boldsymbol{u}$.*

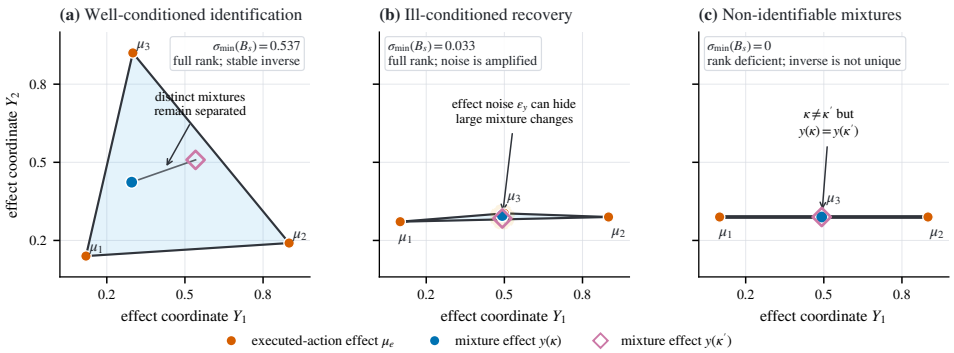



Figure 3: **Geometry of Theorem 4.** The execution simplex maps to the convex hull of action-effect signatures. Affine independence gives a stable inverse (left), poor conditioning amplifies error (center), and rank deficiency makes mixtures observationally equivalent (right).

The quantity $\sigma_{\min}(B_s)$ measures the geometric separation among execution effects: larger values yield more stable inversion, whereas small values increase sensitivity to estimation error and rank deficiency eliminates identifiability. This analysis characterizes when execution-aware modeling can provide identifiable gains over directly modeling intended actions. Appendix A further extends the analysis to simplex-constrained mixtures and dictionary mismatch.

### 5.3 From effect recovery to imagination stability

Identifiability matters only if recovery error remains controlled inside latent rollouts. Let $\boldsymbol{r}_t$ collect the variables carried by a coupled rollout and let $\boldsymbol{u}_t$ denote the $K-1$ free coordinates of its effective-action mixture, with feasible set $\mathcal{U} = \{\boldsymbol{u} \succeq 0 : \mathbf{1}^\top \boldsymbol{u} \leq 1\}$. Conditional on common latent innovations, consider a reference transition $F_t$ and a learned approximation $\widehat{F}_t$ satisfying

$$\|\widehat{F}_t(\boldsymbol{r}, \boldsymbol{u}) - F_t(\boldsymbol{r}, \boldsymbol{u})\|_2 \leq \eta_t, \tag{36}$$

$$\|\widehat{F}_t(\boldsymbol{r}, \boldsymbol{u}) - \widehat{F}_t(\boldsymbol{r}', \boldsymbol{u}')\|_2 \leq L_s \|\boldsymbol{r} - \boldsymbol{r}'\|_2 + L_u \|\boldsymbol{u} - \boldsymbol{u}'\|_2. \tag{37}$$

**Theorem 5** (Recovery-to-imagination stability)**.** *Let $\boldsymbol{r}_{t+1} = F_t(\boldsymbol{r}_t, \boldsymbol{u}_t)$ and $\widehat{\boldsymbol{r}}_{t+1} = \widehat{F}_t(\widehat{\boldsymbol{r}}_t, \widehat{\boldsymbol{u}}_t)$ start from the same state. If Theorem 4 holds along the reference rollout and $\|\widehat{\boldsymbol{y}}_t - \boldsymbol{y}_t\|_2 \leq \epsilon_t$, set $\widehat{\boldsymbol{u}}_t = \Pi_{\mathcal{U}}[B_{s_t}^\dagger(\widehat{\boldsymbol{y}}_t - \boldsymbol{\mu}_K(s_t))]$. Then*

$$\|\widehat{\boldsymbol{r}}_H - \boldsymbol{r}_H\|_2 \leq \sum_{t=0}^{H-1} L_s^{H-1-t} \left( \eta_t + \frac{L_u \epsilon_t}{\sigma_{\min}(B_{s_t})} \right). \tag{38}$$

*If $L_s < 1$, $\eta_t \leq \eta$, $\epsilon_t \leq \epsilon$, and $\sigma_{\min}(B_{s_t}) \geq \sigma_\star > 0$, the right-hand side is at most $(1 - L_s^H)(\eta + L_u \epsilon / \sigma_\star)/(1 - L_s)$.*

The bound isolates three governing factors of recovery stability: world-model approximation error $\eta_t$, execution-effect conditioning through $\sigma_{\min}(B_{s_t})$, and dynamical amplification characterized by $(L_s, L_u)$. It therefore predicts selective benefits of execution modeling: informative execution effects improve recovery only when they are sufficiently separated and dynamically stable. Appendix A provides the full proof, a matrix-form error propagation analysis, and a return-stability corollary.

## 6 Experiments

### 6.1 Protocol and statistical unit

**Environment.** We evaluate pixel-based agents in the Arcade Learning Environment (Bellemare et al., 2013; Machado et al., 2018) across 22 games. Training lasts one million environment steps,

except Breakout and Space Invaders at two million; all methods share preprocessing, action repeat, action space, and episode initialization. Each method–game–seed run was trained on a single NVIDIA GeForce RTX 4090 GPU.

**Hidden execution protocol.** Training switches among NORMAL, STICKY, LAZY, CONFUSE-LR, and FIRE-DROPOUT. A mode is sampled uniformly at reset and resampled with probability $5\times10^{-4}$ per step; corruptions occur with probability $\varrho = 0.35$. Agents observe intended actions but never executed actions or mode identities. Formal channel definitions are in Appendix C.

**Fixed-mode evaluation.** Policies are evaluated under clean execution and fixed corruptions, with native ALE stickiness disabled. All four corruptions are used except for Pong and Breakout (LAZY, CONFUSE-LR) and Space Invaders (LAZY, FIRE-DROPOUT). For game $g$,

$$J_g^{\mathrm{corr}} = \frac{1}{|\mathcal{M}_g|} \sum_{m\in\mathcal{M}_g} J_{g,m}. \tag{39}$$

**Methods and aggregation.** We compare DREAMERV3, the capacity-matched INTERSELF-IDENTITY, and InterSelf. Each method–game condition uses three independent training seeds and 50 completed episodes per fixed mode. Tables average episodes within each seed and then the three seed-level values; training seeds are the statistical unit, and win counts are descriptive. This avoids treating evaluation episodes as independent training runs (Henderson et al., 2018; Agarwal et al., 2021).

### 6.2 HIDDEN EXECUTION BREAKS ACTION FIDELITY

DreamerV3's corrupted mean is below its clean score in 18 of 22 games. Among the 21 positive-clean games, the median and mean relative degradations are $9.6\%$ and $13.4\%$; drops reach $57.9\%$ in Kung Fu Master and $40.3\%$ in River Raid. Because only the hidden execution channel changes, this pattern isolates a failure of action fidelity rather than visual domain shift.

### 6.3 CONTROL WITH LATENT EXECUTION

Across the full suite, InterSelf exceeds DreamerV3 in 12 clean and 14 corrupted conditions and obtains the best corrupted return in 9 games. The corresponding counts are 14, 12, and 7 for InterSelf-Identity, and 6 best-corrupted games for DreamerV3 (Table 1).

Table 1: Descriptive comparison over all 22 games. Counts use three-seed means; *best corrupted* compares all three methods.

| Method | Higher clean than DreamerV3 | Higher corrupted than DreamerV3 | Best corrupted among three |
|---|---|---|---|
| DreamerV3 | – | – | 6/22 |
| InterSelf-Identity | 14/22 | 12/22 | 7/22 |
| InterSelf | 12/22 | **14/22** | **9/22** |

Table 2 reports every game where InterSelf's corrupted mean is at least $5\%$ above both controls. Seven games satisfy this transparent criterion, with gains over InterSelf-Identity ranging from $9.4\%$ to $85.1\%$; InterSelf also exceeds DreamerV3 in all seven and improves clean return in five.

Table 2: Games where InterSelf's corrupted mean is at least $5\%$ above both controls. Scores average fixed modes and then three training seeds (50 episodes per seed–mode pair); Appendix D reports the full suite.

| | | Clean | | | Corrupted mean | | |
|---|---|---|---|---|---|---|---|
| Game | Steps | DreamerV3 | IS-Identity | InterSelf | DreamerV3 | IS-Identity | InterSelf |
| Space Invaders | 2M | 538.50 | 455.20 | **542.70** | 434.00 | 458.35 | **520.45** |
| Galaxian | 1M | 1339.80 | 1640.00 | **1676.80** | 1238.80 | 1454.40 | **1590.50** |
| Phoenix | 1M | 2237.60 | 2299.00 | **2907.40** | 1633.95 | 1561.25 | **2204.25** |
| Demon Attack | 1M | 167.50 | 280.10 | **612.50** | 157.65 | 260.85 | **480.88** |
| Chopper Command | 1M | 1100.00 | 1504.00 | **1648.00** | 1410.50 | 1226.00 | **1576.00** |
| River Raid | 1M | **4319.20** | 2822.20 | 4171.00 | 2577.50 | 2429.80 | **3089.10** |
| Seaquest | 1M | 741.20 | 432.00 | **1141.20** | 770.70 | 468.00 | **866.10** |

### 6.4 Mechanism and scope

The experimental comparisons form a layered test of action mediation. The clean-to-corrupted contrast first establishes that hidden execution makes the issued command an unreliable control variable: DreamerV3 degrades in 18 of 22 games. InterSelf-Identity then controls for the alternative explanation that improvements arise merely from the recurrent execution branch, auxiliary objectives, or additional representation capacity. Its gains over DreamerV3 in 12 corrupted conditions show that this is a substantive control rather than an artificially weak baseline.

The decisive comparison is therefore between InterSelf and InterSelf-Identity. The two models are exactly parameter- and topology-matched and share the same objectives and training protocol; they differ only in whether the inferred effective action mediates the RSSM transition. Across the full suite, InterSelf obtains the highest corrupted return in nine games, compared with seven for InterSelf-Identity and six for DreamerV3. The transparently selected cases in Table 2 further expose the strongest mediation-specific gains, while Appendix D reports all remaining outcomes.

The resulting heterogeneity is consistent with the theoretical conditions. Theorem 4 predicts stable recovery only when execution effects are sufficiently separated, while Theorem 5 additionally requires recovery error not to be amplified excessively by model error or transition sensitivity. Returns are therefore not treated as direct estimates of $\sigma_{\min}(B_s)$; rather, the theory explains why explicit mediation can produce substantial gains in some environments while the nested identity path remains competitive in others. Together, suite breadth tests prevalence, the matched control tests action mediation, and the theory characterizes when that mediation should be informative.

## 7 Conclusion

Hidden execution breaks the usual equivalence between the command selected by a policy and the action that generates the next observation. InterSelf addresses this mismatch by separating a recurrent execution channel from world dynamics and using the resulting effective-action representation consistently in posterior learning and imagination. Our analysis quantifies the predictive information lost when execution is omitted, establishes coverage of categorical and temporally persistent execution failures, and characterizes how effect separation and transition sensitivity govern recovery and finite-horizon imagination error. Across 22 Atari games, hidden execution degrades DreamerV3 in most environments, while InterSelf exceeds it in 14 corrupted conditions and achieves the best corrupted return in nine. Together with the capacity-matched identity control, these results support attributing the strongest improvements to action mediation rather than additional representation capacity alone. The gains are necessarily environment-dependent: when execution effects are weak or observationally redundant, the nested identity path can remain preferable. InterSelf is therefore best understood as an inductive bias for recoverable hidden execution, rather than a universal decomposition of environment dynamics. Extending this framework to continuous control and developing online measures of effect conditioning are promising directions.

## AI USE STATEMENT

In this work, we used generative AI tools for proofreading and correcting typographical errors. Additionally, we used generative AI tools for minor Python code generation. All LLM-generated code was reviewed and tested for correctness by the authors. Generative AI tools were not used for research idea generation or theoretical formulations. The authors take responsibility for the final content of this work, including any text, claims, or artifacts produced with the aid of generative AI.

## ETHICS STATEMENT

This work studies reinforcement-learning methodology in simulated Atari environments. It does not involve human participants, personal data, or newly released datasets. We are not aware of material ethical risks specific to this study beyond the general considerations associated with reinforcement-learning research. The authors have read and adhere to the ICLR Code of Ethics.

## REPRODUCIBILITY STATEMENT

The code for reproducing our results is provided anonymously at `https://anonymous.4open.science/r/InterSelf-B1CE`.

## REFERENCES

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc G. Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In *Advances in Neural Information Processing Systems*, volume 34, pp. 29304–29320, 2021.

Elizabeth S. Allman, Catherine Matias, and John A. Rhodes. Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, 37(6A):3099–3132, 2009. doi: 10.1214/09-AOS689.

Marc G. Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47:253–279, 2013. doi: 10.1613/jair.3912.

Yann Bouteiller, Simon Ramstedt, Giovanni Beltrame, Christopher J. Pal, and Jonathan Binas. Reinforcement learning with random delays. In *International Conference on Learning Representations*, 2021.

Jake Bruce, Michael D. Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Maria Elisabeth Bechtle, Feryal Behbahani, Stephanie C. Y. 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. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 4603–4623, 2024.

Yash Chandak, Georgios Theocharous, James Kostas, Scott Jordan, and Philip Thomas. Learning action representations for reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 941–950, 2019.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2 edition, 2006. doi: 10.1002/047174882X.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 2555–2565, 2019.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020.

Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. In *International Conference on Learning Representations*, 2021.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640:647–653, 2025. doi: 10.1038/s41586-025-08744-2.

Nicklas Hansen, Hao Su, and Xiaolong Wang. Temporal difference learning for model predictive control. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 8387–8406, 2022.

Nicklas Hansen, Hao Su, and Xiaolong Wang. TD-MPC2: Scalable, robust world models for continuous control. In *International Conference on Learning Representations*, 2024.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pp. 3207–3214, 2018. doi: 10.1609/aaai.v32i1.11694.

Garud N. Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2): 257–280, 2005. doi: 10.1287/moor.1040.0129.

Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1–2):99–134, 1998. doi: 10.1016/S0004-3702(98)00023-X.

Aravind S. Lakshminarayanan, Sahil Sharma, and Balaraman Ravindran. Dynamic action repetition for deep reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, pp. 2133–2139, 2017. doi: 10.1609/aaai.v31i1.10918.

Pierre Liotet, Davide Maran, Lorenzo Bisi, and Marcello Restelli. Delayed reinforcement learning by imitation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 13528–13556, 2022.

Marlos C. Machado, Marc G. Bellemare, Erik Talvitie, Joel Veness, Matthew Hausknecht, and Michael Bowling. Revisiting the arcade learning environment: Evaluation protocols and open problems for general agents. *Journal of Artificial Intelligence Research*, 61:523–562, 2018. doi: 10.1613/jair.5699.

Vincent Micheli, Eloi Alonso, and François Fleuret. Transformers are sample-efficient world models. In *International Conference on Learning Representations*, 2023.

Arnab Nilim and Laurent El Ghaoui. Robust control of markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005. doi: 10.1287/opre.1050.0216.

Junhyuk Oh, Xiaoxiao Guo, Honglak Lee, Richard L. Lewis, and Satinder Singh. Action-conditional video prediction using deep networks in atari games. In *Advances in Neural Information Processing Systems*, volume 28, pp. 2845–2853, 2015.

Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pp. 2778–2787, 2017.

Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. Robust adversarial reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pp. 2817–2826, 2017.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy Lillicrap, and David Silver. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588: 604–609, 2020. doi: 10.1038/s41586-020-03051-4.

Sahil Sharma, Aravind S. Lakshminarayanan, and Balaraman Ravindran. Learning to repeat: Fine grained action repetition for deep reinforcement learning. In *International Conference on Learning Representations*, 2017.

Richard S. Sutton. Dyna, an integrated architecture for learning, planning, and reacting. *ACM SIGART Bulletin*, 2(4):160–163, 1991. doi: 10.1145/122344.122377.

Richard S. Sutton, Doina Precup, and Satinder Singh. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112(1–2):181–211, 1999. doi: 10.1016/S0004-3702(99)00052-1.

Henry Teicher. Identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 34(4): 1265–1269, 1963. doi: 10.1214/aoms/1177703862.

Chen Tessler, Yonathan Efroni, and Shie Mannor. Action robust reinforcement learning and applications in continuous control. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 6215–6224, 2019.

Seonghyeon Ye, Joel Jang, Byeongguk Jeon, Sejune Joo, Jianwei Yang, Baolin Peng, Ajay Mandlekar, Reuben Tan, Yu-Wei Chao, Bill Yuchen Lin, Lars Liden, Kimin Lee, Jianfeng Gao, Luke Zettlemoyer, Dieter Fox, and Minjoon Seo. Latent action pretraining from videos. In *International Conference on Learning Representations*, 2025.

Weirui Ye, Shaohuai Liu, Thanard Kurutach, Pieter Abbeel, and Yang Gao. Mastering atari games with limited data. In *Advances in Neural Information Processing Systems*, volume 34, pp. 25476–25488, 2021.

# A PROOFS

We state the population objects used by the results before giving the proofs. The analysis separates five questions that are easily conflated: whether execution carries predictive information, whether the architecture can represent the relevant channel, whether its mixture is identifiable from a chosen effect statistic, whether the soft-action interface introduces approximation bias, and how recovery error propagates through imagination. These are population and conditional-stability statements.

## A.1 ASSUMPTIONS AND POPULATION QUANTITIES

**Regular conditional laws.** All random variables are defined on a common probability space and take values in standard Borel spaces, so the conditional laws below exist. For the log-loss results, these laws are dominated by a common measure and the stated conditional entropies are finite. These conditions justify exchanging conditional expectations and the usual entropy–KL decomposition.

**Norms and almost-sure scope.** Vector norms are Euclidean unless a subscript states otherwise, matrix norms are the corresponding induced norms, and all conditional identities and bounds hold on the support of the conditioning variables, up to null sets. Whenever a conditional expectation is used as an $L^2$ projection, the corresponding random vector is assumed square-integrable. Empty products equal one and empty sums equal zero.

**Predictive state and effect sufficiency.** The state $S_t$ need only be sufficient for the prediction target used in a given statement; it need not equal the learned RSSM state. For Theorem 4, we additionally assume Eq. (33). Equivalently at the level of conditional means, once $(S_t, E_t)$ is fixed, $(A_t, \mathcal{H}_t)$ does not further change the mean of $Y_{t+1}$. A stronger sufficient condition is

$$Y_{t+1} \perp (A_t, \mathcal{H}_t) \mid (S_t, E_t). \tag{40}$$

**Conditional dictionary.** The vectors $\{\boldsymbol{\mu}_e(s)\}_{e=1}^{K}$ are fixed population effect signatures indexed by the known action vocabulary. Theorem 4 identifies mixture weights *conditional on this dictionary*; it does not claim that an unrestricted latent model jointly recovers both the dictionary and semantically named execution labels. Proposition 13 quantifies what happens when the dictionary is estimated with error.

**Coupled rollout convention.** For Theorem 5, $\boldsymbol{r}_t$ may contain the full augmented latent carry. A stochastic rollout is analyzed conditional on a coupling of its latent innovations, so $F_t$ and $\widehat{F}_t$ denote the resulting one-step maps. The bound is pathwise under Eqs. (36)–(37); taking expectations preserves it whenever the displayed errors are integrable. The free coordinate is embedded as $\boldsymbol{\kappa}(\boldsymbol{u}) = (\boldsymbol{u}^\top, 1 - \mathbf{1}^\top \boldsymbol{u})^\top$. Hence an $L_e$-Lipschitz transition in the full simplex vector induces $L_u \le \sqrt{K} L_e$ in the coordinates used by the theorem, since

$$\|\boldsymbol{\kappa}(\boldsymbol{u}) - \boldsymbol{\kappa}(\boldsymbol{u}')\|_2^2 = \|\boldsymbol{u} - \boldsymbol{u}'\|_2^2 + \left(\mathbf{1}^\top(\boldsymbol{u} - \boldsymbol{u}')\right)^2 \le K\|\boldsymbol{u} - \boldsymbol{u}'\|_2^2. \tag{41}$$

In fact, the linear part of this embedding is $T = [I_{K-1}; -\mathbf{1}^\top]$, for which $T^\top T = I_{K-1} + \mathbf{1}\mathbf{1}^\top$ and $\|T\|_2 = \sqrt{\lambda_{\max}(T^\top T)} = \sqrt{K}$; the constant is therefore sharp in these coordinates.

### A.2 Information gap and recoverable predictive information

*Proof of Theorem 1.* Let $U$ be any conditioning variable and let $p$ denote the true conditional law. Under logarithmic loss,

$$\begin{aligned}
\mathcal{R}^*(X \mid U) &= \inf_q \mathbb{E}[-\log q(X \mid U)] && (42)\\
&= H(X \mid U) + \inf_q \mathbb{E}_U[\mathrm{KL}(p(\cdot \mid U) \| q(\cdot \mid U))] && (43)\\
&= H(X \mid U). && (44)
\end{aligned}$$

With $U = (S_t, A_t)$, Eq. (44) gives

$$\begin{aligned}
&\mathcal{R}^*(X_{t+1} \mid S_t, A_t) - \mathcal{R}^*(X_{t+1} \mid S_t, A_t, E_t) && (45)\\
&= H(X_{t+1} \mid S_t, A_t) - H(X_{t+1} \mid S_t, A_t, E_t) && (46)\\
&= I(X_{t+1}; E_t \mid S_t, A_t) \ge 0. && (47)
\end{aligned}$$

□

*Proof of Corollary 2.* Set $U = (S_t, A_t)$, $X = X_{t+1}$, $E = E_t$, and $\widehat{E} = \widehat{E}_t$. Applying Eq. (44) with conditioning variables $U$ and $(U, \widehat{E})$ gives

$$\begin{aligned}
\mathcal{R}^*(X \mid U) - \mathcal{R}^*(X \mid U, \widehat{E}) &= H(X \mid U) - H(X \mid U, \widehat{E}) && (48)\\
&= I(X; \widehat{E} \mid U). && (49)
\end{aligned}$$

The assumption $X \perp \widehat{E} \mid (U, E)$ is equivalent to $I(X; \widehat{E} \mid U, E) = 0$. The conditional mutual-information chain rule therefore yields

$$\begin{aligned}
I(X; E, \widehat{E} \mid U) &= I(X; E \mid U) + I(X; \widehat{E} \mid U, E) && (50)\\
&= I(X; E \mid U), && (51)\\
I(X; E, \widehat{E} \mid U) &= I(X; \widehat{E} \mid U) + I(X; E \mid U, \widehat{E}) && (52)\\
&\ge I(X; \widehat{E} \mid U). && (53)
\end{aligned}$$

Combining Eqs. (51)–(53) proves Eq. (4). □

The conditional mutual information also has the pointwise predictive form

$$I(X; E \mid U) = \mathbb{E}_{U,E}[\mathrm{KL}(p(X \mid U, E) \,\|\, p(X \mid U))]\,. \tag{54}$$

Hence the gap vanishes if and only if $p(X \mid U, E) = p(X \mid U)$ for $p(U, E)$-almost every $(U, E)$. This gives the boundary case used in the paper: if alternative executions are predictively equivalent at the chosen state, no execution representation can lower Bayes log loss.

Under the conditional Markov condition in Corollary 2, the unrecovered part of the oracle gap is exact rather than merely bounded:

$$I(X; E \mid U) - I(X; \widehat{E} \mid U) = I(X; E \mid U, \widehat{E}). \tag{55}$$

Indeed, Eqs. (51)–(53) and the chain rule give the equality directly. Thus a surrogate need not reproduce a private label; it is sufficient that conditioning on it leaves little execution-dependent predictive information.

**Proposition 6** (Squared-loss prediction gap)**.** *Let $X \in L^2(\mathbb{R}^p)$ and define $\mathcal{R}_2^*(X \mid U) = \inf_g \mathbb{E}\|X - g(U)\|_2^2$. Then*

$$\mathcal{R}_2^*(X \mid U) - \mathcal{R}_2^*(X \mid U, E) = \mathbb{E}\,\|\mathbb{E}[X \mid U, E] - \mathbb{E}[X \mid U]\|_2^2\,. \tag{56}$$

*If $X \perp \widehat{E} \mid (U, E)$, then the corresponding gain from $\widehat{E}$ is no larger than the right-hand side of Eq. (56).*

*Proof.* Write $m_U = \mathbb{E}[X \mid U]$ and $m_{U,E} = \mathbb{E}[X \mid U, E]$. Conditional expectation is the orthogonal projection of $X$ onto the closed subspace of measurable square-integrable random variables. Thus, for every square-integrable $U$-measurable $g$,

$$\mathbb{E}\|X - g(U)\|_2^2 = \mathbb{E}\|X - m_U\|_2^2 + \mathbb{E}\|m_U - g(U)\|_2^2, \tag{57}$$

because the cross term vanishes:

$$\begin{aligned}&\mathbb{E}\langle X - m_U, m_U - g(U)\rangle \\ &\quad = \mathbb{E}[\langle \mathbb{E}[X - m_U \mid U], m_U - g(U)\rangle] = 0.\end{aligned} \tag{58}$$

Next decompose $X - m_U = (X - m_{U,E}) + (m_{U,E} - m_U)$. The corresponding cross term is zero since $\mathbb{E}[X - m_{U,E} \mid U, E] = 0$, and hence

$$\mathbb{E}\|X - m_U\|_2^2 = \mathbb{E}\|X - m_{U,E}\|_2^2 + \mathbb{E}\|m_{U,E} - m_U\|_2^2. \tag{59}$$

Subtracting the two Bayes risks proves Eq. (56).

For the surrogate claim, let $m_{U,\widehat{E}} = \mathbb{E}[X \mid U, \widehat{E}]$. Conditional independence and the tower property give

$$m_{U,\widehat{E}} = \mathbb{E}\Big[m_{U,E} \mid U, \widehat{E}\Big]\,. \tag{60}$$

Because $m_U$ is $U$-measurable, conditional Jensen yields

$$\mathbb{E}\|m_{U,\widehat{E}} - m_U\|_2^2 = \mathbb{E}\,\Big\|\mathbb{E}[m_{U,E} - m_U \mid U, \widehat{E}]\Big\|_2^2 \tag{61}$$

$$\leq \mathbb{E}\|m_{U,E} - m_U\|_2^2. \tag{62}$$

This proves the surrogate inequality. It is the squared-loss analogue of Corollary 2 and directly covers mean effect prediction. □

### A.3 Representation of the evaluated channels

*Proof of Theorem 3.* The InterSelf channel has the row-wise form

$$C = \mathrm{Diag}(\mathbf{1} - \boldsymbol{\alpha}) + \mathrm{Diag}(\boldsymbol{\alpha})D. \tag{63}$$

For normal execution, set $\boldsymbol{\alpha} = \mathbf{0}$. For lazy execution, set $\boldsymbol{\alpha} = \varrho\mathbf{1}$ and $D = \mathbf{1}\boldsymbol{\nu}^\top$; then

$$C = \mathrm{Diag}((1 - \varrho)\mathbf{1}) + \mathrm{Diag}(\varrho\mathbf{1})\mathbf{1}\boldsymbol{\nu}^\top \tag{64}$$

$$= (1 - \varrho)I + \varrho\mathbf{1}\boldsymbol{\nu}^\top. \tag{65}$$

For left–right confusion, the same substitution with $D = P_{\mathrm{LR}}$ gives

$$C = (1 - \varrho)I + \varrho P_{\mathrm{LR}}. \tag{66}$$

For fire dropout, set $\boldsymbol{\alpha} = \varrho\boldsymbol{f}$ and choose $D[i, :] = \boldsymbol{\nu}^\top$ whenever $f_i = 1$. Row $i$ is therefore

$$C[i, :] = e_i^\top + \varrho f_i(\boldsymbol{\nu}^\top - e_i^\top), \tag{67}$$

which stacks into Eq. (30).

Finally, let $q_{t-1}$ be the conditional marginal law of the preceding execution. Sticky execution gives

$$q_t(e) = \mathbb{P}(E_t = e \mid A_{0:t}, \mathcal{H}_t) \tag{68}$$

$$= (1 - \varrho)\mathbf{1}\{e = A_t\} + \varrho q_{t-1}(e). \tag{69}$$

With $C_t = I$ and $\rho_t = \varrho$, Eq. (16) is the same vector recursion. To justify the closure statement explicitly, for a target gate $g \in (0, 1)$ choose logit $\ell(g) = \log(g/(1 - g))$; for $g = 0$ or $g = 1$, use respectively the sequences $-n$ or $n$. Their sigmoid images converge to the desired boundary value. The softmax rows are handled analogously by sending the selected coordinate logit to $n$ while keeping the others fixed. Continuity of Eqs. (63) and (16) then yields each displayed boundary channel as a limit of implemented parameters. □

**Theorem 7** (Finite-horizon temporal-kernel equivalence)**.** *Fix $t \geq 0$. Unrolling Eq. (16) gives*

$$\widetilde{\mathbf{a}}_t = w_{-1,t}\widetilde{\mathbf{a}}_{-1} + \sum_{k=0}^{t} w_{k,t}\overline{\mathbf{a}}_k, \quad \begin{cases} w_{-1,t} = \prod_{j=0}^{t} \rho_j, \\ w_{k,t} = (1-\rho_k)\prod_{j=k+1}^{t} \rho_j. \end{cases} \tag{70}$$

*The weights belong to $\Delta^{t+1}$. Conversely, every $\boldsymbol{\lambda} = (\lambda_{-1}, \lambda_0, \ldots, \lambda_t) \in \Delta^{t+1}$ is represented by some gate sequence $(\rho_0, \ldots, \rho_t) \in [0,1]^{t+1}$. Whenever $\Lambda_k = \sum_{j=-1}^{k} \lambda_j > 0$, one such sequence is*

$$\rho_k = \frac{\Lambda_{k-1}}{\Lambda_k}, \qquad k = 0, \ldots, t, \tag{71}$$

*with gates preceding the first positive cumulative mass arbitrary.*

*Proof.* We first prove Eq. (70) by induction. At $t = 0$, Eq. (16) reads

$$\widetilde{\mathbf{a}}_0 = \rho_0\widetilde{\mathbf{a}}_{-1} + (1-\rho_0)\overline{\mathbf{a}}_0, \tag{72}$$

which has the stated coefficients. Suppose the formula holds at time $t-1$. Substitution into the recursion gives

$$\widetilde{\mathbf{a}}_t = (1-\rho_t)\overline{\mathbf{a}}_t + \rho_t \left( w_{-1,t-1}\widetilde{\mathbf{a}}_{-1} + \sum_{k=0}^{t-1} w_{k,t-1}\overline{\mathbf{a}}_k \right) \tag{73}$$

$$= (1-\rho_t)\overline{\mathbf{a}}_t + (\rho_t w_{-1,t-1})\widetilde{\mathbf{a}}_{-1} + \sum_{k=0}^{t-1} (\rho_t w_{k,t-1})\overline{\mathbf{a}}_k. \tag{74}$$

The identities $\rho_t w_{-1,t-1} = \prod_{j=0}^{t} \rho_j$ and $\rho_t w_{k,t-1} = (1-\rho_k)\prod_{j=k+1}^{t} \rho_j$ establish the induction step. All coefficients are nonnegative. Their sum $S_t$ satisfies

$$S_t = (1-\rho_t) + \rho_t S_{t-1}, \qquad S_{-1} = 1, \tag{75}$$

and hence $S_t = 1$ by induction. For the converse, note that $\Lambda_t = 1$. If $\lambda_{-1} > 0$, define every gate by Eq. (71). Then

$$\prod_{j=k+1}^{t} \rho_j = \prod_{j=k+1}^{t} \frac{\Lambda_{j-1}}{\Lambda_j} = \frac{\Lambda_k}{\Lambda_t} = \Lambda_k. \tag{76}$$

Therefore

$$(1-\rho_k)\prod_{j=k+1}^{t} \rho_j = \frac{\lambda_k}{\Lambda_k}\Lambda_k = \lambda_k, \qquad \prod_{j=0}^{t} \rho_j = \lambda_{-1}. \tag{77}$$

It remains to handle a zero-mass prefix. Let $m = \min\{k \in \{0, \ldots, t\} : \lambda_k > 0\}$, which exists when $\lambda_{-1} = 0$. Set

$$\rho_k = \begin{cases} 1, & k < m, \\ 0, & k = m, \\ \Lambda_{k-1}/\Lambda_k, & k > m. \end{cases} \tag{78}$$

The factor $\rho_m = 0$ makes $w_{-1,t}$ and every $w_{k,t}$ with $k < m$ equal to zero. For $k \geq m$, the same telescoping calculation applies because $\Lambda_k > 0$, and gives $w_{k,t} = \lambda_k$. Thus every boundary case is represented as well. □

The theorem is a fixed-horizon expressivity statement: kernels at different horizons must share a time-consistent gate sequence and therefore cannot be specified independently. It nevertheless shows that the scalar persistence recursion is not restricted to a stationary geometric kernel.

**Lemma 8** (Temporal-tail control)**.** *Fix $0 \leq L \leq t$ and suppose $0 \leq \rho_j \leq \bar{\rho} < 1$. In the unrolled kernel of Eq. (70), the total weight at lags strictly larger than $L$ is at most $\bar{\rho}^{L+1}$. If $\rho_j = \rho$ for $j = t-L, \ldots, t$, this mass is exactly $\rho^{L+1}$.*

*Proof.* Using Eq. (70), the total mass carried from before time $t - L$ (including the initial state) is

$$M_{>L} = w_{-1,t} + \sum_{k=0}^{t-L-1} w_{k,t}$$
$$= \left( \prod_{j=t-L}^{t} \rho_j \right) \left[ \prod_{j=0}^{t-L-1} \rho_j + \sum_{k=0}^{t-L-1} (1 - \rho_k) \prod_{j=k+1}^{t-L-1} \rho_j \right]. \tag{79}$$

The bracket is the sum of the normalized kernel at time $t - L - 1$, and therefore equals one. Consequently,

$$M_{>L} = \prod_{j=t-L}^{t} \rho_j \leq \bar{\rho}^{L+1}. \tag{80}$$

For constant $\rho$, equality gives $M_{>L} = \rho^{L+1}$ directly at every finite horizon. At the evaluation value $\rho = 0.35$, mass beyond lag five is $0.35^6 \approx 1.84 \times 10^{-3}$, as visualized in Figure 2. □

### A.4 Identifiability and stability

*Proof of Theorem 4.* The feasible coordinates are

$$\mathcal{U} = \left\{ \boldsymbol{u} \in \mathbb{R}^{K-1} : \boldsymbol{u} \succeq 0, \mathbf{1}^\top \boldsymbol{u} \leq 1 \right\}. \tag{81}$$

Fix a supported condition $(s, a, h)$, suppress the state argument on the effect signatures, and write $\kappa_e = \mathbb{P}(E_t = e \mid S_t = s, A_t = a, \mathcal{H}_t = h)$. The law of iterated expectation and effect sufficiency give Eq. (34) directly:

$$\boldsymbol{y} = \sum_{e=1}^{K} \kappa_e \mathbb{E}[Y_{t+1} \mid s, e, a, h] \tag{82}$$
$$= \sum_{e=1}^{K} \kappa_e \boldsymbol{\mu}_e(s) \qquad \text{(by Eq. (33))} \tag{83}$$
$$= \boldsymbol{\mu}_K + \sum_{e=1}^{K-1} \kappa_e (\boldsymbol{\mu}_e - \boldsymbol{\mu}_K) \tag{84}$$
$$= \boldsymbol{\mu}_K + B_s \boldsymbol{u}. \tag{85}$$

If $\operatorname{rank}(B_s) = K - 1$, then $B_s^\dagger B_s = I$ and hence

$$\boldsymbol{u} = B_s^\dagger (\boldsymbol{y} - \boldsymbol{\mu}_K), \tag{86}$$

which proves uniqueness. Applying the same inverse to an estimated effect gives

$$\|\widehat{\boldsymbol{u}} - \boldsymbol{u}\|_2 = \|B_s^\dagger (\widehat{\boldsymbol{y}} - \boldsymbol{y})\|_2 \tag{87}$$
$$\leq \|B_s^\dagger\|_2 \|\widehat{\boldsymbol{y}} - \boldsymbol{y}\|_2 \tag{88}$$
$$= \frac{\|\widehat{\boldsymbol{y}} - \boldsymbol{y}\|_2}{\sigma_{\min}(B_s)}. \tag{89}$$

If $\operatorname{rank}(B_s) < K - 1$, choose $\boldsymbol{v} \neq \mathbf{0}$ with $B_s \boldsymbol{v} = \mathbf{0}$. For an interior $\boldsymbol{u}$, every coordinate is positive and the simplex slack $r = 1 - \mathbf{1}^\top \boldsymbol{u}$ is positive. Define

$$\epsilon_0 = min \left\{ \min_{i : v_i \neq 0} \frac{u_i}{|v_i|}, \begin{cases} r/|\mathbf{1}^\top \boldsymbol{v}|, & \mathbf{1}^\top \boldsymbol{v} \neq 0, \\ +\infty, & \mathbf{1}^\top \boldsymbol{v} = 0 \end{cases} \right\} > 0. \tag{90}$$

For any $0 < \epsilon < \epsilon_0$, both $\boldsymbol{u}_\pm = \boldsymbol{u} \pm \epsilon \boldsymbol{v}$ satisfy $\boldsymbol{u}_\pm \succeq 0$ and $\mathbf{1}^\top \boldsymbol{u}_\pm \leq 1$, so they lie in $\mathcal{U}$. They are distinct, but

$$\boldsymbol{\mu}_K + B_s \boldsymbol{u}_+ = \boldsymbol{\mu}_K + B_s \boldsymbol{u}_- = \boldsymbol{y}. \tag{91}$$

Thus distinct feasible mixtures have the same conditional mean effect. □

**Corollary 9** (Necessary effect dimension)**.** *Full-rank identification from a $d$-dimensional mean effect requires $d \geq K-1$.*

*Proof.* Because $B_s \in \mathbb{R}^{d\times(K-1)}$, $\mathrm{rank}(B_s) \leq d$. The rank condition in Theorem 4 is therefore impossible when $d < K-1$. □

**Corollary 10** (Directional amplification)**.** *Under the full-rank condition, let $B_s = P\Sigma Q^\top$ be a thin singular-value decomposition, with $\Sigma = \mathrm{Diag}(\sigma_1, \ldots, \sigma_{K-1})$ and left singular vectors $\boldsymbol{p}_j$. For $\boldsymbol{\delta}_y = \widehat{\boldsymbol{y}} - \boldsymbol{y}$, the unprojected least-squares error obeys*

$$\widetilde{\boldsymbol{u}} - \boldsymbol{u} = Q\Sigma^{-1}P^\top \boldsymbol{\delta}_y, \tag{92}$$

$$\|\widetilde{\boldsymbol{u}} - \boldsymbol{u}\|_2^2 = \sum_{j=1}^{K-1} \frac{(\boldsymbol{p}_j^\top \boldsymbol{\delta}_y)^2}{\sigma_j^2}. \tag{93}$$

*Proof.* Full column rank implies $B_s^\dagger = Q\Sigma^{-1}P^\top$. Substituting this identity into $\widetilde{\boldsymbol{u}} - \boldsymbol{u} = B_s^\dagger \boldsymbol{\delta}_y$ proves Eq. (92). Orthogonality of $Q$ then gives

$$\|Q\Sigma^{-1}P^\top \boldsymbol{\delta}_y\|_2^2 = \|\Sigma^{-1}P^\top \boldsymbol{\delta}_y\|_2^2 = \sum_{j=1}^{K-1} \frac{(\boldsymbol{p}_j^\top \boldsymbol{\delta}_y)^2}{\sigma_j^2}, \tag{94}$$

which proves Eq. (93). Thus the $\sigma_{\min}^{-1}$ bound is attained only when the effect error is aligned with a least-separated singular direction; components orthogonal to $\mathrm{col}(B_s)$ do not affect the recovered coordinates. □

**Proposition 11** (Simplex-constrained stability)**.** *Under the full-rank condition, let*

$$\widetilde{\boldsymbol{u}} = B_s^\dagger(\widehat{\boldsymbol{y}} - \boldsymbol{\mu}_K), \qquad \widehat{\boldsymbol{u}} = \Pi_{\mathcal{U}}(\widetilde{\boldsymbol{u}}), \tag{95}$$

*where $\Pi_{\mathcal{U}}$ is Euclidean projection. Then*

$$\|\widehat{\boldsymbol{u}} - \boldsymbol{u}\|_2 \leq \frac{\|\widehat{\boldsymbol{y}} - \boldsymbol{y}\|_2}{\sigma_{\min}(B_s)}. \tag{96}$$

*Proof.* The feasible set $\mathcal{U}$ is closed and convex and contains $\boldsymbol{u}$. Let $\widehat{\boldsymbol{u}} = \Pi_{\mathcal{U}}(\widetilde{\boldsymbol{u}})$. The variational characterization of Euclidean projection states that

$$\langle \widetilde{\boldsymbol{u}} - \widehat{\boldsymbol{u}}, \boldsymbol{v} - \widehat{\boldsymbol{u}} \rangle \leq 0 \quad \text{for every } \boldsymbol{v} \in \mathcal{U}. \tag{97}$$

Setting $\boldsymbol{v} = \boldsymbol{u}$ and expanding gives

$$\|\widetilde{\boldsymbol{u}} - \boldsymbol{u}\|_2^2 = \|\widetilde{\boldsymbol{u}} - \widehat{\boldsymbol{u}}\|_2^2 + \|\widehat{\boldsymbol{u}} - \boldsymbol{u}\|_2^2 + 2\,\langle \widetilde{\boldsymbol{u}} - \widehat{\boldsymbol{u}}, \widehat{\boldsymbol{u}} - \boldsymbol{u} \rangle \tag{98}$$

$$\geq \|\widehat{\boldsymbol{u}} - \boldsymbol{u}\|_2^2. \tag{99}$$

The final term is bounded exactly as in Eq. (35). □

**Proposition 12** (Finite-sample effect recovery)**.** *At a fixed supported condition, suppose*

$$Y_i = \boldsymbol{\mu}_K + B_s \boldsymbol{u} + \boldsymbol{\xi}_i, \qquad i = 1, \ldots, n, \tag{100}$$

*where the $\boldsymbol{\xi}_i$ are conditionally independent with zero mean and covariance $\Sigma$. Let $\overline{Y}_n = n^{-1}\sum_i Y_i$, $\widetilde{\boldsymbol{u}} = B_s^\dagger(\overline{Y}_n - \boldsymbol{\mu}_K)$, and $\widehat{\boldsymbol{u}} = \Pi_{\mathcal{U}}(\widetilde{\boldsymbol{u}})$. Under the full-rank condition,*

$$\mathbb{E}\|\widetilde{\boldsymbol{u}} - \boldsymbol{u}\|_2^2 = \frac{1}{n}\,\mathrm{tr}\big(B_s^\dagger \Sigma (B_s^\dagger)^\top\big) \tag{101}$$

$$\leq \frac{\mathrm{tr}(\Sigma)}{n\,\sigma_{\min}^2(B_s)}. \tag{102}$$

*The projected estimate obeys the same upper bound. Moreover, for every $r > 0$,*

$$\mathbb{P}(\|\widehat{\boldsymbol{u}} - \boldsymbol{u}\|_2 \geq r) \leq \frac{\mathrm{tr}(\Sigma)}{n\,\sigma_{\min}^2(B_s) r^2}. \tag{103}$$

*Proof.* Since $B_s^\dagger B_s = I$,

$$\widetilde{\boldsymbol{u}} - \boldsymbol{u} = B_s^\dagger \overline{\boldsymbol{\xi}}_n, \qquad \mathrm{Cov}(\overline{\boldsymbol{\xi}}_n) = \frac{\Sigma}{n}. \tag{104}$$

For a zero-mean vector $Z$ and matrix $A$, $\mathbb{E}\|AZ\|_2^2 = \mathrm{tr}(A\,\mathrm{Cov}(Z)A^\top)$, which gives Eq. (101). Because $(B_s^\dagger)^\top B_s^\dagger \preceq \|B_s^\dagger\|_2^2 I$ and $\Sigma \succeq 0$,

$$\mathrm{tr}\big(B_s^\dagger \Sigma (B_s^\dagger)^\top\big) = \mathrm{tr}\Big(\Sigma^{1/2}(B_s^\dagger)^\top B_s^\dagger \Sigma^{1/2}\Big) \tag{105}$$

$$\leq \|B_s^\dagger\|_2^2\,\mathrm{tr}(\Sigma) = \frac{\mathrm{tr}(\Sigma)}{\sigma_{\min}^2(B_s)}. \tag{106}$$

Projection cannot increase the distance to the feasible $\boldsymbol{u}$ by Eq. (99). Applying Markov's inequality to the projected squared error proves Eq. (103). Thus poor effect conditioning incurs a quadratic sample penalty, even before neural approximation or optimization error is considered. □

**Proposition 13** (Estimated-dictionary perturbation)**.** *Let $\widehat{B}_s = B_s + \Delta B$, $\widehat{\boldsymbol{\mu}}_K = \boldsymbol{\mu}_K + \boldsymbol{\delta}_K$, and $\widehat{\boldsymbol{y}} = \boldsymbol{y} + \boldsymbol{\delta}_y$. If $\|\Delta B\|_2 < \sigma_{\min}(B_s)$, then $\widehat{B}_s$ has full column rank, and the unprojected estimator $\widehat{\boldsymbol{u}} = \widehat{B}_s^\dagger(\widehat{\boldsymbol{y}} - \widehat{\boldsymbol{\mu}}_K)$ satisfies*

$$\|\widehat{\boldsymbol{u}} - \boldsymbol{u}\|_2 \leq \frac{\|\boldsymbol{\delta}_y\|_2 + \|\boldsymbol{\delta}_K\|_2 + \|\Delta B\|_2\|\boldsymbol{u}\|_2}{\sigma_{\min}(B_s) - \|\Delta B\|_2}. \tag{107}$$

*Projection onto $\mathcal{U}$ cannot increase this bound.*

*Proof.* Weyl's singular-value inequality gives

$$\sigma_{\min}(\widehat{B}_s) \geq \sigma_{\min}(B_s) - \|\Delta B\|_2 > 0. \tag{108}$$

Thus $\widehat{B}_s$ has full column rank and $\widehat{B}_s^\dagger \widehat{B}_s = I$. Using $\boldsymbol{y} - \boldsymbol{\mu}_K = B_s\boldsymbol{u}$,

$$\widehat{\boldsymbol{u}} - \boldsymbol{u} = \widehat{B}_s^\dagger\big[(\widehat{\boldsymbol{y}} - \widehat{\boldsymbol{\mu}}_K) - \widehat{B}_s\boldsymbol{u}\big] \tag{109}$$

$$= \widehat{B}_s^\dagger(\boldsymbol{\delta}_y - \boldsymbol{\delta}_K - \Delta B\boldsymbol{u}). \tag{110}$$

Submultiplicativity, the triangle inequality, and $\|\widehat{B}_s^\dagger\|_2 = 1/\sigma_{\min}(\widehat{B}_s)$ yield Eq. (107). If $\widehat{\boldsymbol{u}}_{\mathrm{proj}} = \Pi_{\mathcal{U}}(\widehat{\boldsymbol{u}})$, then Eq. (99), applied with $\widetilde{\boldsymbol{u}} = \widehat{\boldsymbol{u}}$, gives

$$\|\widehat{\boldsymbol{u}}_{\mathrm{proj}} - \boldsymbol{u}\|_2 \leq \|\widehat{\boldsymbol{u}} - \boldsymbol{u}\|_2, \tag{111}$$

which proves the projection claim without requiring the estimated dictionary to equal the population dictionary. □

**Identifiability scope.** The recovered mixture is defined at the level of predictive execution effects rather than semantic labels. When effect signatures are jointly learned with unconstrained latent assignments, the model is identifiable only up to a permutation of latent components, since permuting labels and mixture coordinates leaves the observable law unchanged. Similarly, rank conditions cannot separate executed actions that induce identical effect distributions. Thus, the theorem characterizes recovery of a predictive execution mixture under identifiable effect structures, while semantic interpretation of latent states requires additional supervision or external anchors.

### A.5 Approximation induced by the soft-action interface

The external transition under a categorical execution mixture is formally a mixture of action-conditioned transitions. InterSelf instead supplies the mixture vector to a differentiable transition map. The following proposition isolates the resulting approximation term.

**Proposition 14** (Soft-action exactness and curvature gap)**.** *Fix a state $s$ and let $\mathcal{F}_s : \Delta^{K-1} \to \mathbb{R}^p$ extend to a twice continuously differentiable map on a neighborhood of the simplex. Suppose that for every $\boldsymbol{a} \in \Delta^{K-1}$ and $\boldsymbol{v} \in \mathbb{R}^K$ tangent to the simplex,*

$$\|D^2\mathcal{F}_s(\boldsymbol{a})[\boldsymbol{v},\boldsymbol{v}]\|_2 \leq M_s\|\boldsymbol{v}\|_2^2. \tag{112}$$

*Let $E \sim \boldsymbol{\kappa}$ and let $\mathbf{e}_E$ denote its one-hot vector. Then*

$$\|\mathbb{E}[\mathcal{F}_s(\mathbf{e}_E)] - \mathcal{F}_s(\boldsymbol{\kappa})\|_2 \leq \frac{M_s}{2}\left(1 - \|\boldsymbol{\kappa}\|_2^2\right). \tag{113}$$

*In particular, the soft-action interface is exact when $\mathcal{F}_s$ is affine in its action argument and the gap vanishes for deterministic execution.*

*Proof.* Set $\boldsymbol{A} = \mathbf{e}_E$ and $\boldsymbol{H} = \boldsymbol{A} - \boldsymbol{\kappa}$. Since $\mathbf{1}^\top \boldsymbol{A} = \mathbf{1}^\top \boldsymbol{\kappa} = 1$, $\mathbf{1}^\top \boldsymbol{H} = 0$, so $\boldsymbol{H}$ is tangent to the simplex. Moreover, $\boldsymbol{\kappa} + \tau \boldsymbol{H} = (1-\tau)\boldsymbol{\kappa} + \tau \boldsymbol{A} \in \Delta^{K-1}$ for every $\tau \in [0,1]$. Taylor's theorem with integral remainder at $\boldsymbol{\kappa} = \mathbb{E}[\boldsymbol{A}]$ therefore gives

$$\begin{aligned}\mathcal{F}_s(\boldsymbol{A}) = \mathcal{F}_s(\boldsymbol{\kappa}) + D\mathcal{F}_s(\boldsymbol{\kappa})(\boldsymbol{A} - \boldsymbol{\kappa}) \\ + \int_0^1 (1-\tau) D^2 \mathcal{F}_s(\boldsymbol{\kappa} + \tau(\boldsymbol{A} - \boldsymbol{\kappa}))[\boldsymbol{A} - \boldsymbol{\kappa}, \boldsymbol{A} - \boldsymbol{\kappa}], d\tau.\end{aligned} \tag{114}$$

Because $\mathbb{E}[\boldsymbol{H}] = \mathbf{0}$, the expectation of the linear term is zero. Subtracting $\mathcal{F}_s(\boldsymbol{\kappa})$, taking expectations, and using the triangle inequality for the expectation and integral yields

$$\|\mathbb{E}[\mathcal{F}_s(\boldsymbol{A})] - \mathcal{F}_s(\boldsymbol{\kappa})\|_2 \leq \mathbb{E}\left[\int_0^1 (1-\tau)\left\|D^2 \mathcal{F}_s(\boldsymbol{\kappa} + \tau \boldsymbol{H})[\boldsymbol{H}, \boldsymbol{H}]\right\|_2 d\tau\right] \tag{115}$$

$$\leq M_s \int_0^1 (1-\tau) d\tau\, \mathbb{E}\|\boldsymbol{H}\|_2^2 \tag{116}$$

$$= \frac{M_s}{2}\mathbb{E}\|\boldsymbol{A} - \boldsymbol{\kappa}\|_2^2. \tag{117}$$

Because $\boldsymbol{A}$ is one-hot and $\mathbb{E}[\boldsymbol{A}] = \boldsymbol{\kappa}$,

$$\mathbb{E}\|\boldsymbol{A} - \boldsymbol{\kappa}\|_2^2 = \mathbb{E}\|\boldsymbol{A}\|_2^2 - 2\boldsymbol{\kappa}^\top \mathbb{E}[\boldsymbol{A}] + \|\boldsymbol{\kappa}\|_2^2 \tag{118}$$

$$= 1 - \|\boldsymbol{\kappa}\|_2^2 = \mathrm{tr}(\mathrm{Cov}(\boldsymbol{A})). \tag{119}$$

which proves Eq. (113). For an affine map the Hessian is zero and the Taylor remainder vanishes identically, so equality of the two transition constructions is exact. If execution is deterministic, then $\boldsymbol{\kappa}$ is a simplex vertex and $1 - \|\boldsymbol{\kappa}\|_2^2 = 0$. □

The factor $1 - \|\boldsymbol{\kappa}\|_2^2$ is the categorical Gini impurity. Thus the approximation is controlled jointly by how uncertain execution is and how nonlinear the transition is along action directions.

### A.6 Propagation from effect recovery to imagination

*Proof of Theorem 5.* Let $d_t = \|\widehat{\boldsymbol{r}}_t - \boldsymbol{r}_t\|_2$ and $\delta_t = \|\widehat{\boldsymbol{u}}_t - \boldsymbol{u}_t\|_2$. Adding and subtracting $\widehat{F}_t(\boldsymbol{r}_t, \boldsymbol{u}_t)$ and applying Eqs. (36)–(37) gives

$$\begin{aligned}d_{t+1} &\leq \|\widehat{F}_t(\widehat{\boldsymbol{r}}_t, \widehat{\boldsymbol{u}}_t) - \widehat{F}_t(\boldsymbol{r}_t, \boldsymbol{u}_t)\|_2 + \|\widehat{F}_t(\boldsymbol{r}_t, \boldsymbol{u}_t) - F_t(\boldsymbol{r}_t, \boldsymbol{u}_t)\|_2 \\ &\leq L_s d_t + L_u \delta_t + \eta_t.\end{aligned} \tag{120}$$

By Proposition 11 and the assumed effect error,

$$\delta_t \leq \frac{\epsilon_t}{\sigma_{\min}(B_{s_t})}. \tag{121}$$

Define

$$q_t = \eta_t + \frac{L_u \epsilon_t}{\sigma_{\min}(B_{s_t})}. \tag{122}$$

Equations (120)–(121) give $d_{t+1} \leq L_s d_t + q_t$. We now prove by induction that, for every $h \in \{1, \ldots, H\}$,

$$d_h \leq \sum_{t=0}^{h-1} L_s^{h-1-t} q_t. \tag{123}$$

For $h = 1$, $d_0 = 0$ gives $d_1 \le q_0$. If the claim holds at $h$, then

$$d_{h+1} \le L_s d_h + q_h \tag{124}$$

$$\le L_s \sum_{t=0}^{h-1} L_s^{h-1-t} q_t + q_h \tag{125}$$

$$= \sum_{t=0}^{h} L_s^{h-t} q_t, \tag{126}$$

which proves the induction step. Setting $h = H$ and substituting Eq. (122) proves Eq. (38). Under the uniform bounds, $q_t \le q := \eta + L_u \epsilon / \sigma_\star$, so

$$d_H \le q \sum_{j=0}^{H-1} L_s^j = q \frac{1 - L_s^H}{1 - L_s}, \tag{127}$$

as $L_s < 1$. □

The constant $L_s$ is a concise worst-case sensitivity. If instead the $t$th learned transition is $L_{s,t}$-Lipschitz in its state argument, the same induction gives the sharper path-dependent form

$$d_H \le \sum_{t=0}^{H-1} \left( \prod_{j=t+1}^{H-1} L_{s,j} \right) q_t. \tag{128}$$

Thus a local error is amplified only by sensitivities of transitions that follow it.

The accumulation can also be written as a lower-triangular linear operator. Define

$$q_t = \eta_t + \frac{L_u \epsilon_t}{\sigma_{\min}(B_{s_t})}, \quad \boldsymbol{d} = (d_1, \dots, d_H)^\top, \quad \boldsymbol{q} = (q_0, \dots, q_{H-1})^\top. \tag{129}$$

Then, componentwise,

$$\boldsymbol{d} \preceq \mathsf{T}_H(L_s)\boldsymbol{q}, \qquad \mathsf{T}_H(L_s) = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ L_s & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ L_s^{H-1} & \cdots & L_s & 1 \end{bmatrix}. \tag{130}$$

The matrix makes the horizon effect explicit: each local error is propagated by all later state sensitivities. Define the finite geometric factor

$$\mathsf{G}_m(x) = \sum_{\ell=0}^{m-1} x^\ell = \begin{cases} (1 - x^m)/(1 - x), & x \ne 1, \\ m, & x = 1. \end{cases} \tag{131}$$

Since $L_s \ge 0$, the last row has the largest row sum and

$$\|\mathsf{T}_H(L_s)\|_\infty = \sum_{j=0}^{H-1} L_s^j = \mathsf{G}_H(L_s). \tag{132}$$

**Corollary 15** (Finite-horizon return stability)**.** *For a fixed coupled intended-action sequence, let $J_H = \sum_{t=0}^{H-1} \gamma^t g_t(\boldsymbol{r}_t)$ and $\widehat{J}_H = \sum_{t=0}^{H-1} \gamma^t \widehat{g}_t(\widehat{\boldsymbol{r}}_t)$, where $\gamma \in [0, 1]$. If $\widehat{g}_t$ is $L_R$-Lipschitz and $|\widehat{g}_t(\boldsymbol{r}) - g_t(\boldsymbol{r})| \le \zeta_t$, then*

$$|\widehat{J}_H - J_H| \le \sum_{t=0}^{H-1} \gamma^t (\zeta_t + L_R d_t), \tag{133}$$

*where each $d_t$ is bounded by Eq. (123). Consequently,*

$$|\widehat{J}_H - J_H| \le \sum_{t=0}^{H-1} \gamma^t \zeta_t + L_R \sum_{j=0}^{H-2} \gamma^{j+1} \mathsf{G}_{H-j-1}(\gamma L_s) q_j. \tag{134}$$

*Proof.* For each $t$, adding and subtracting $\widehat{g}_t(\boldsymbol{r}_t)$ gives

$$|\widehat{g}_t(\widehat{\boldsymbol{r}}_t) - g_t(\boldsymbol{r}_t)| \le |\widehat{g}_t(\widehat{\boldsymbol{r}}_t) - \widehat{g}_t(\boldsymbol{r}_t)| + |\widehat{g}_t(\boldsymbol{r}_t) - g_t(\boldsymbol{r}_t)| \tag{135}$$

$$\le L_R d_t + \zeta_t. \tag{136}$$

Multiplying by $\gamma^t$, summing, and using the triangle inequality proves Eq. (133). Since $d_0 = 0$, substituting Eq. (123) yields

$$|\widehat{J}_H - J_H| \le \sum_{t=0}^{H-1} \gamma^t \zeta_t + L_R \sum_{t=1}^{H-1} \gamma^t \sum_{j=0}^{t-1} L_s^{t-1-j} q_j \tag{137}$$

$$= \sum_{t=0}^{H-1} \gamma^t \zeta_t + L_R \sum_{j=0}^{H-2} q_j \sum_{t=j+1}^{H-1} \gamma^t L_s^{t-1-j} \tag{138}$$

$$= \sum_{t=0}^{H-1} \gamma^t \zeta_t + L_R \sum_{j=0}^{H-2} \gamma^{j+1} q_j \sum_{\ell=0}^{H-j-2} (\gamma L_s)^\ell, \tag{139}$$

which is Eq. (134) by Eq. (131). □

This corollary controls evaluation of a fixed coupled action sequence. It is not a policy-improvement theorem: if the actor selects different intentions after the rollouts diverge, an additional policy-sensitivity term is required.

### A.7 Validity and nesting of the action path

**Lemma 16** (Simplex preservation). *If $\mathbf{a}_t, \widetilde{\mathbf{a}}_{t-1} \in \Delta^{K-1}$, $\boldsymbol{\alpha}_t \in [0,1]^K$, every row of $\mathbf{D}_t$ lies in $\Delta^{K-1}$, and $\rho_t \in [0,1]$, then $\overline{\mathbf{a}}_t, \widetilde{\mathbf{a}}_t \in \Delta^{K-1}$.*

*Proof.* For every row $i$,

$$C_t[i,j] \ge 0, \tag{140}$$

$$\sum_j C_t[i,j] = (1 - \alpha_{t,i}) + \alpha_{t,i} \sum_j D_t[i,j] = 1. \tag{141}$$

Hence $C_t$ is row-stochastic and $C_t \mathbf{1} = \mathbf{1}$. Since $\mathbf{a}_t \succeq 0$ and $C_t \succeq 0$ entrywise, $\overline{\mathbf{a}}_t = C_t^\top \mathbf{a}_t \succeq 0$, while

$$\mathbf{1}^\top \overline{\mathbf{a}}_t = \mathbf{1}^\top C_t^\top \mathbf{a}_t = (C_t \mathbf{1})^\top \mathbf{a}_t = \mathbf{1}^\top \mathbf{a}_t = 1. \tag{142}$$

Thus $\overline{\mathbf{a}}_t \in \Delta^{K-1}$. Finally,

$$\widetilde{\mathbf{a}}_t = (1 - \rho_t)\overline{\mathbf{a}}_t + \rho_t \widetilde{\mathbf{a}}_{t-1} \in \Delta^{K-1}. \tag{143}$$

The last inclusion follows because the simplex is convex and $\rho_t \in [0,1]$. □

Theorem 7 additionally shows that the persistence recursion remains a normalized simplex mixture at every horizon and gives an explicit inverse from finite-horizon temporal weights to gates.

**Lemma 17** (Deviation from the identity path). *For a one-hot intended action $\mathbf{a}_t = e_i$,*

$$\|\overline{\mathbf{a}}_t - \mathbf{a}_t\|_1 = 2\alpha_{t,i}(1 - D_t[i,i]) \le 2\alpha_{t,i}, \qquad \|\widetilde{\mathbf{a}}_t - \mathbf{a}_t\|_1 \le 2\alpha_{t,i} + 2\rho_t. \tag{144}$$

*Proof.* The selected row of the categorical map gives $\overline{\mathbf{a}}_t = (1 - \alpha_{t,i})e_i + \alpha_{t,i} D_t[i,:]^\top$. Because $D_t[i,:]$ is a probability vector,

$$\|\overline{\mathbf{a}}_t - e_i\|_1 = \alpha_{t,i} \|D_t[i,:]^\top - e_i\|_1 \tag{145}$$

$$= \alpha_{t,i} \left( |D_t[i,i] - 1| + \sum_{j \ne i} D_t[i,j] \right) \tag{146}$$

$$= 2\alpha_{t,i}(1 - D_t[i,i]) \le 2\alpha_{t,i}. \tag{147}$$

Using Eq. (16) and again bounding the distance between simplex elements by two,

$$\|\widetilde{\mathbf{a}}_t - e_i\|_1 \le (1-\rho_t)\|\overline{\mathbf{a}}_t - e_i\|_1 + \rho_t\|\widetilde{\mathbf{a}}_{t-1} - e_i\|_1 \tag{148}$$

$$\le 2(1-\rho_t)\alpha_{t,i} + 2\rho_t \le 2\alpha_{t,i} + 2\rho_t. \tag{149}$$

□

Finally, Lemma 17 gives the first two limits below. If the neural transition $f_\phi$ is continuous in its action argument, as for the standard finite-weight RSSM parameterization, continuity gives the third:

$$\|\boldsymbol{\alpha}_t\|_\infty \to 0 \Longrightarrow C_t \to I \Longrightarrow \overline{\mathbf{a}}_t \to \mathbf{a}_t, \tag{150}$$

$$(\|\boldsymbol{\alpha}_t\|_\infty, \rho_t) \to (0,0) \Longrightarrow \widetilde{\mathbf{a}}_t \to \mathbf{a}_t, \tag{151}$$

$$f_\phi(\mathbf{h}_t, \mathbf{z}_t, \widetilde{\mathbf{a}}_t) \longrightarrow f_\phi(\mathbf{h}_t, \mathbf{z}_t, \mathbf{a}_t). \tag{152}$$

Thus the ordinary DreamerV3 action path is contained as the identity limit.

## B Additional Model Details

The compact update in Eq. (20) uses the post-transition stochastic state as evidence about the just-observed action effect. Define

$$\mathbf{v}_{t+1} = [\mathbf{m}_t, \mathrm{vec}(\mathbf{z}_{t+1}), \mathbf{a}_t, \widetilde{\mathbf{a}}_{t-1}, \widetilde{\mathbf{a}}_t], \qquad \mathbf{w}_{t+1} = g_\psi^{\mathrm{self}}(\mathbf{v}_{t+1}). \tag{153}$$

The implementation then applies

$$\widehat{\mathbf{m}}_{t+1} = \tanh(W_m \mathbf{w}_{t+1} + b_m), \tag{154}$$

$$\boldsymbol{\gamma}_{t+1} = \sigma(W_g \mathbf{w}_{t+1} + b_g), \tag{155}$$

$$\mathbf{m}_{t+1} = \boldsymbol{\gamma}_{t+1} \odot \widehat{\mathbf{m}}_{t+1} + (\mathbf{1} - \boldsymbol{\gamma}_{t+1}) \odot \mathbf{m}_t. \tag{156}$$

During posterior learning, $\mathbf{z}_{t+1}$ is sampled from $q_\phi(\mathbf{z}_{t+1} \mid \mathbf{h}_{t+1}, \mathbf{x}_{t+1})$. During imagination, the same map uses the prior sample from $p_\phi(\mathbf{z}_{t+1} \mid \mathbf{h}_{t+1})$. Consequently the execution state and effective-action recursion evolve in imagined rollouts; they are not frozen or replaced by the identity channel at policy-optimization time.

The effect loss in Eq. (22) stops gradients through the encoder displacement target. It is therefore a predictive consistency signal, not an inverse-dynamics label. The capacity-matched identity control preserves this update, predictor, and all associated parameters; only the edge carrying $\widetilde{\mathbf{a}}_t$ into $f_\phi$ is replaced by $\mathbf{a}_t$, as stated in Eq. (26).

## C Execution-Environment Protocol

Let $B_t \sim \mathrm{Bernoulli}(\varrho)$ with $\varrho = 0.35$, let $a_{\mathrm{noop}}$ be ALE's no-op action, let $s_{\mathrm{LR}}$ exchange LEFT and RIGHT tokens while preserving compound semantics such as LEFTFIRE, and let $F$ be the set of fire-containing actions. For intended action $A_t$ and previous executed action $E_{t-1}$, the fixed evaluation channels are

$$\textsc{Normal}: \quad E_t = A_t, \tag{157}$$

$$\textsc{Sticky}: \quad E_t = \begin{cases} E_{t-1}, & B_t = 1, \\ A_t, & B_t = 0, \end{cases} \tag{158}$$

$$\textsc{Lazy}: \quad E_t = \begin{cases} a_{\mathrm{noop}}, & B_t = 1, \\ A_t, & B_t = 0, \end{cases} \tag{159}$$

$$\textsc{Confuse-LR}: \quad E_t = \begin{cases} s_{\mathrm{LR}}(A_t), & B_t = 1, \\ A_t, & B_t = 0, \end{cases} \tag{160}$$

$$\textsc{Fire-Dropout}: \quad E_t = \begin{cases} a_{\mathrm{noop}}, & B_t = 1 \text{ and } A_t \in F, \\ A_t, & \text{otherwise.} \end{cases} \tag{161}$$

At training reset, the active channel is sampled uniformly from the five modes. Under stochastic switching, an event occurs independently with probability $5\times10^{-4}$ per environment step and triggers

a fresh uniform draw from the same mode set. Evaluation fixes one mode for the full episode and disables native ALE sticky actions, so no additional unknown persistence process is superimposed on the custom channel.

The agent receives no field derived from the active mode or executed action. All methods therefore observe the same image, reward, continuation, and intended-action streams.

## D Complete Broad-Suite Results

For method $q$, game $g$, seed $s \in \{1, 2, 3\}$, and fixed mode $m$, let $R^{(e)}_{qgsm}$ be the return of completed evaluation episode $e \in \{1, \ldots, 50\}$. We first average completed episodes within each independently trained seed and then average the three resulting seed-level estimates:

$$\widehat{J}_{qgsm} = \frac{1}{50} \sum_{e=1}^{50} R^{(e)}_{qgsm}, \qquad \widehat{J}_{qgm} = \frac{1}{3} \sum_{s=1}^{3} \widehat{J}_{qgsm}. \tag{162}$$

The corrupted score in Table 3 is $|\mathcal{M}_g|^{-1} \sum_{m \in \mathcal{M}_g} \widehat{J}_{qgm}$, where $\mathcal{M}_g$ contains the non-normal modes evaluated for game $g$. Thus, training seeds—rather than individual evaluation episodes—are the unit of replication.

Table 3: Complete mean episodic returns under clean and fixed corrupted execution. Corrupted scores are unweighted means over the non-normal fixed modes evaluated for each game. Each entry first averages 50 completed evaluation episodes within each independently trained seed and then averages the three resulting seed-level estimates. Higher is better; bold denotes the largest mean within each evaluation condition. Scores remain on their game-specific scales and are not averaged across games.

| | | Clean | | | Corrupted mean | | |
|---|---|---|---|---|---|---|---|
| Game | Steps | DreamerV3 | IS-Identity | InterSelf | DreamerV3 | IS-Identity | InterSelf |
| Pong | 1M | -5.92 | **4.20** | 3.14 | -16.16 | -11.37 | **-11.29** |
| Breakout | 2M | 14.82 | **26.34** | 15.62 | 9.99 | **17.86** | 10.41 |
| Space Invaders | 2M | 538.50 | 455.20 | **542.70** | 434.00 | 458.35 | **520.45** |
| Boxing | 1M | 80.50 | **94.10** | 87.36 | 59.51 | **75.80** | 60.02 |
| Assault | 1M | 796.40 | **853.10** | 790.06 | 692.12 | **755.55** | 704.53 |
| Galaxian | 1M | 1339.80 | 1640.00 | **1676.80** | 1238.80 | 1454.40 | **1590.50** |
| Phoenix | 1M | 2237.60 | 2299.00 | **2907.40** | 1633.95 | 1561.25 | **2204.25** |
| Demon Attack | 1M | 167.50 | 280.10 | **612.50** | 157.65 | 260.85 | **480.88** |
| Beam Rider | 1M | **711.20** | 632.88 | 696.96 | **654.58** | 641.64 | 621.56 |
| Chopper Command | 1M | 1100.00 | 1504.00 | **1648.00** | 1410.50 | 1226.00 | **1576.00** |
| River Raid | 1M | **4319.20** | 2822.20 | 4171.00 | 2577.50 | 2429.80 | **3089.10** |
| Seaquest | 1M | 741.20 | 432.00 | **1141.20** | 770.70 | 468.00 | **866.10** |
| Star Gunner | 1M | **1144.00** | 984.00 | 848.00 | **1037.00** | 999.50 | 868.00 |
| Time Pilot | 1M | 3218.00 | **3522.00** | 3412.00 | 3097.00 | 3184.00 | **3267.00** |
| Centipede | 1M | 8520.60 | 6948.66 | **8752.88** | **6404.14** | 6136.12 | 5454.25 |
| Battle Zone | 1M | **4820.00** | 4140.00 | 3040.00 | **4620.00** | 3650.00 | 2620.00 |
| Alien | 1M | 759.80 | **802.60** | 524.40 | 680.60 | **733.25** | 611.70 |
| Kung Fu Master | 1M | 26856.00 | **34192.00** | 22534.00 | 11308.50 | **13441.50** | 11748.00 |
| Carnival | 1M | 2229.40 | **2622.40** | 2005.20 | **2267.95** | 2137.60 | 1950.60 |
| Name This Game | 1M | **6959.20** | 6771.20 | 6075.80 | **5486.65** | 5307.85 | 5049.05 |
| Wizard of Wor | 1M | 430.00 | **978.00** | 528.00 | 458.00 | **1107.50** | 695.00 |
| Atlantis | 1M | 24158.00 | **26428.00** | 23832.00 | 21842.00 | **24331.00** | 20801.00 |

Table 4: Per-mode decomposition of the complete results in Table 3. Each entry first averages 50 completed evaluation episodes within each independently trained seed and then averages the three resulting seed-level estimates. Higher is better; bold denotes the largest mean in each game–mode row.

| Game | Evaluation mode | DreamerV3 | IS-Identity | InterSelf |
|---|---|---|---|---|
| Pong | *Normal* | -5.92 | **4.20** | 3.14 |
| | Lazy | -13.56 | **-6.60** | -7.38 |
| | Confuse-LR | -18.76 | -16.13 | **-15.20** |
| Breakout | *Normal* | 14.82 | **26.34** | 15.62 |
| | Lazy | 10.92 | **25.84** | 12.06 |
| | Confuse-LR | 9.06 | **9.88** | 8.76 |
| Space Invaders | *Normal* | 538.50 | 455.20 | **542.70** |
| | Lazy | 422.80 | 448.95 | **521.60** |
| | Fire-Dropout | 445.20 | 467.75 | **519.30** |
| Boxing | *Normal* | 80.50 | **94.10** | 87.36 |
| | Sticky | 52.62 | **61.72** | 55.38 |
| | Lazy | 55.04 | **74.76** | 58.68 |
| | Confuse-LR | 60.86 | **85.72** | 57.50 |
| | Fire-Dropout | 69.50 | **81.00** | 68.52 |
| Assault | *Normal* | 796.40 | **853.10** | 790.06 |
| | Sticky | 707.06 | **760.54** | 697.94 |
| | Lazy | 705.54 | **783.24** | 706.84 |
| | Confuse-LR | 571.34 | **660.56** | 660.16 |
| | Fire-Dropout | 784.54 | **817.86** | 753.16 |
| Galaxian | *Normal* | 1339.80 | 1640.00 | **1676.80** |
| | Sticky | 1200.00 | 1548.60 | **1646.20** |
| | Lazy | 1422.40 | 1540.60 | **1785.60** |
| | Confuse-LR | 1114.40 | **1358.60** | 1211.60 |
| | Fire-Dropout | 1218.40 | 1369.80 | **1718.60** |
| Phoenix | *Normal* | 2237.60 | 2299.00 | **2907.40** |
| | Sticky | 1516.20 | 1241.60 | **1966.80** |
| | Lazy | 1913.60 | 1920.80 | **2570.00** |
| | Confuse-LR | 1235.40 | 1145.60 | **1366.40** |
| | Fire-Dropout | 1870.60 | 1937.00 | **2913.80** |
| Demon Attack | *Normal* | 167.50 | 280.10 | **612.50** |
| | Sticky | 155.20 | 270.90 | **464.60** |
| | Lazy | 148.90 | 280.90 | **443.40** |
| | Confuse-LR | 159.50 | 255.00 | **460.80** |
| | Fire-Dropout | 167.00 | 236.60 | **554.70** |
| Beam Rider | *Normal* | **711.20** | 632.88 | 696.96 |
| | Sticky | 664.72 | **687.68** | 598.40 |
| | Lazy | **669.76** | 604.80 | 625.20 |
| | Confuse-LR | 617.36 | **673.12** | 670.72 |
| | Fire-Dropout | **666.48** | 600.96 | 591.92 |
| Chopper Command | *Normal* | 1100.00 | 1504.00 | **1648.00** |
| | Sticky | 1516.00 | 1552.00 | **1898.00** |
| | Lazy | 1412.00 | 1024.00 | **1534.00** |
| | Confuse-LR | **1310.00** | 1278.00 | 1292.00 |
| | Fire-Dropout | 1404.00 | 1050.00 | **1580.00** |
| River Raid | *Normal* | **4319.20** | 2822.20 | 4171.00 |
| | Sticky | 2361.40 | 2331.20 | **2714.60** |
| | Lazy | 2893.00 | 2602.20 | **3304.60** |
| | Confuse-LR | 1891.00 | 2087.20 | **2496.60** |
| | Fire-Dropout | 3164.60 | 2698.60 | **3840.60** |
| Seaquest | *Normal* | 741.20 | 432.00 | **1141.20** |
| | Sticky | 668.00 | 420.80 | **762.00** |
| | Lazy | **788.00** | 497.20 | 774.40 |
| | Confuse-LR | 787.60 | 474.40 | **920.40** |
| | Fire-Dropout | 839.20 | 479.60 | **1007.60** |
| Star Gunner | *Normal* | **1144.00** | 984.00 | 848.00 |
| | Sticky | **1000.00** | 992.00 | 934.00 |
| | Lazy | 888.00 | 934.00 | **962.00** |
| | Confuse-LR | **1182.00** | 1066.00 | 886.00 |
| | Fire-Dropout | **1078.00** | 1006.00 | 690.00 |

Continued on next page

Table 4 (continued)

| Game | Evaluation mode | DreamerV3 | IS-Identity | InterSelf |
|---|---|---|---|---|
| Time Pilot | *Normal* | 3218.00 | **3522.00** | 3412.00 |
| | Sticky | 3504.00 | **3752.00** | 2856.00 |
| | Lazy | **3558.00** | 2934.00 | 3334.00 |
| | Confuse-LR | 2920.00 | 3058.00 | **3524.00** |
| | Fire-Dropout | 2406.00 | 2992.00 | **3354.00** |
| Centipede | *Normal* | 8520.60 | 6948.66 | **8752.88** |
| | Sticky | **6946.28** | 6890.32 | 5713.58 |
| | Lazy | 5001.84 | **5186.04** | 4852.72 |
| | Confuse-LR | **9157.34** | 6915.82 | 6773.24 |
| | Fire-Dropout | 4511.10 | **5552.30** | 4477.44 |
| Battle Zone | *Normal* | **4820.00** | 4140.00 | 3040.00 |
| | Sticky | **5700.00** | 4140.00 | 3020.00 |
| | Lazy | **4100.00** | 2900.00 | 2160.00 |
| | Confuse-LR | 3800.00 | **4040.00** | 2600.00 |
| | Fire-Dropout | **4880.00** | 3520.00 | 2700.00 |
| Alien | *Normal* | 759.80 | **802.60** | 524.40 |
| | Sticky | **719.40** | 709.40 | 579.00 |
| | Lazy | **738.60** | 725.00 | 613.00 |
| | Confuse-LR | 480.40 | 657.40 | **666.40** |
| | Fire-Dropout | 784.00 | **841.20** | 588.40 |
| Kung Fu Master | *Normal* | 26856.00 | **34192.00** | 22534.00 |
| | Sticky | 13800.00 | **15936.00** | 14836.00 |
| | Lazy | 5074.00 | 5970.00 | **7470.00** |
| | Confuse-LR | **9348.00** | 9278.00 | 8458.00 |
| | Fire-Dropout | 17012.00 | **22582.00** | 16228.00 |
| Carnival | *Normal* | 2229.40 | **2622.40** | 2005.20 |
| | Sticky | 2166.40 | **2248.40** | 1939.60 |
| | Lazy | **2482.00** | 2413.20 | 1968.40 |
| | Confuse-LR | **1988.40** | 1587.60 | 1707.60 |
| | Fire-Dropout | **2435.00** | 2301.20 | 2186.80 |
| Name This Game | *Normal* | **6959.20** | 6771.20 | 6075.80 |
| | Sticky | 6370.40 | **6494.00** | 5839.60 |
| | Lazy | **5750.00** | 5407.00 | 5115.20 |
| | Confuse-LR | 3868.60 | **3912.60** | 3880.60 |
| | Fire-Dropout | **5957.60** | 5417.80 | 5360.80 |
| Wizard of Wor | *Normal* | 430.00 | **978.00** | 528.00 |
| | Sticky | 470.00 | **1020.00** | 444.00 |
| | Lazy | 448.00 | **1132.00** | 524.00 |
| | Confuse-LR | 488.00 | 1262.00 | **1294.00** |
| | Fire-Dropout | 426.00 | **1016.00** | 518.00 |
| Atlantis | *Normal* | 24158.00 | **26428.00** | 23832.00 |
| | Sticky | 18012.00 | **20312.00** | 19916.00 |
| | Lazy | 23814.00 | **24380.00** | 19266.00 |
| | Confuse-LR | 20530.00 | **24478.00** | 22854.00 |
| | Fire-Dropout | 25012.00 | **28154.00** | 21168.00 |

*Scope and aggregation.* Normal denotes clean fixed-mode evaluation. The corrupted mean reported in Table 3 is the unweighted arithmetic mean of the non-normal rows shown for each game. Pong and Breakout use Lazy and Confuse-LR; Space Invaders uses Lazy and Fire-Dropout; all remaining games use Sticky, Lazy, Confuse-LR, and Fire-Dropout. The same game-specific mode set is used for every method and seed. For every method–game–mode condition, returns are first averaged over 50 completed evaluation episodes within each seed and then across the three independently trained seeds.