REPRESENTATION SUFFICIENCY FOR CLOSED-LOOP CONTROL: DECISION RISK AND THE OCCUPANCY BRIDGE

Anonymous authors
Paper under double-blind review

ABSTRACT

Learned representations for robot control are often evaluated through action prediction on demonstrated states, yet accurate prediction need not imply effective closed-loop control. We separate return loss into the information deficit imposed by a representation and the excess loss of a learned policy within the resulting policy class. On a fixed state distribution, action-value gaps bound the minimum decision risk in terms of Bayes optimal-action error. Relating this risk to attainable return requires accounting for the different state occupancies induced by the risk-minimizing decision rule and the best policy available under the representation. A localized correction further tightens the bounds. Under a fixed macro-action interface, removing direction information reduces the exact optimal discounted return from 0.990 to 0.291 ($\gamma = 0.99$ per macro decision). The resulting information deficit is 0.699, whereas decision risk under expert occupancy, scaled to the same return units, is only 0.0246. Since both optimal returns are computed under the same interface, the deficit reflects the representation restriction rather than learned-policy excess. We further show that identical recorded demonstrations can be consistent with learned-policy excess ranging from zero to linear in the horizon, even when the representation permits optimal control. Finally, fixed-interface data aggregation improves state-based, frozen-latent, and visual policies without adding information at execution time, providing evidence that part of the learned-policy excess is recoverable within the same policy class.

1 INTRODUCTION

Learned representations play a central role in robot control by determining what information is available to a policy during execution. Their quality therefore affects not only prediction from data but also the best closed-loop performance a controller can achieve. Yet representations are often evaluated through action prediction on demonstrated states, whereas control depends on the states visited by the policy during execution. This distinction applies to state encoders, world models, and vision-language-action systems, all of which are ultimately evaluated through closed-loop behavior. We ask how much attainable return is lost when a representation makes control-relevant states indistinguishable. The control consequence of a prediction error depends on the associated action-value gap: a nearly optimal action and a missed irreversible opportunity may count equally toward the Bayes error. Moreover, a rule that performs well on demonstrations may fail on states it visits during execution. Poor learned-policy performance may therefore reflect either missing information or failure to exploit available information. We relate Bayes optimal-action error, fixed-distribution decision risk, and the information deficit. Theorem 1 bounds fixed-distribution decision risk using action-value gaps and Bayes optimal-action error. Proposition 1 then connects this risk to attainable return, accounting for the fact that the risk-minimizing decision rule and the best policy available under the representation may visit different states. A localized refinement restricts the correction to control-relevant losses. Theorem 2 shows that recorded demonstrations need not identify a learned-policy excess. Figure 1 summarizes the framework.

Exact finite-process optimization isolates information restriction from learned-policy excess. In the robot macro-action interface, removing direction information reduces optimal return by 0.699, while the scaled decision risk under the expert occupancy is 0.0246, or 3.52% of the deficit. Aligned

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.

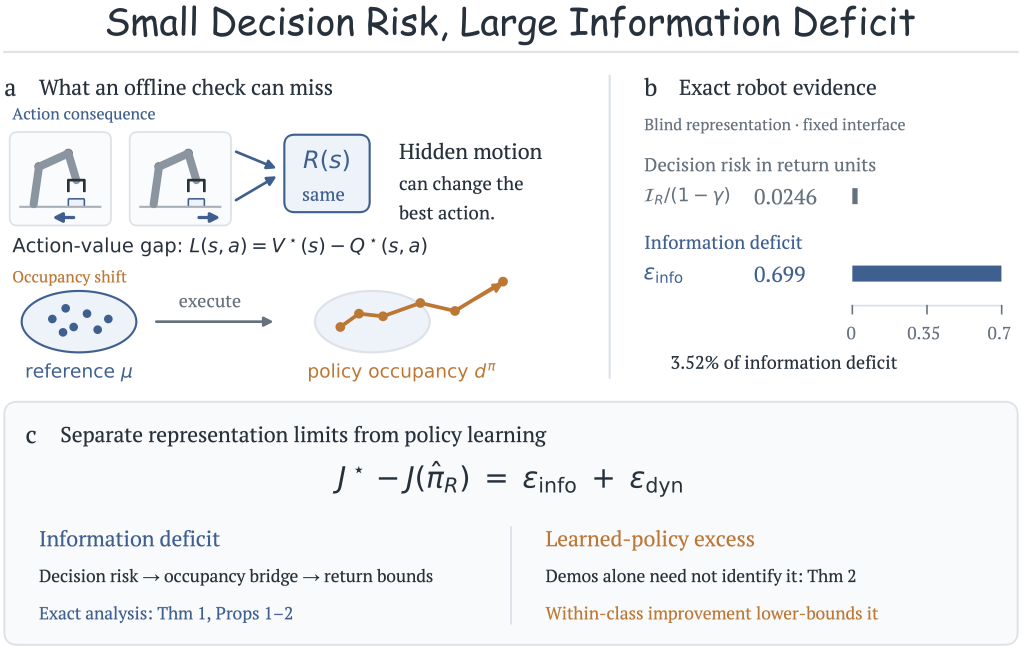

Figure 1: **Framework for representation sufficiency in closed-loop control.** Fixed-distribution decision risk need not reflect the return supported by a representation once policy occupancy changes. The framework connects decision risk to the information deficit and separates this representation-level limit from learned-policy excess.

prediction and execution probes, together with policy-recovery experiments, provide complementary evidence about learned controllers. The research contributions are summarized as follows.

- **A value-weighted bridge from decision ambiguity to return.** Action-value gaps turn Bayes optimal-action error into bounds on fixed-distribution decision risk. The occupancy bridge then relates this risk to the best return achievable under R, with a localized correction for distribution shift.
- **Exact evidence of risk–return separation.** Finite-process examples attain the risk bounds and zero-shift bridge exactly. In robot graphs, a small decision risk under expert occupancy coexists with a large information deficit. The localized bound is substantially tighter than the global-TV bound, and the exact decomposition separates decision-objective mismatch from the signed occupancy-shift term.
- **Identification limits and recovery evidence.** Identical demonstrations can be consistent with learned-policy excess ranging from zero to linear in the horizon, even when the representation permits optimal control. Conversely, same-objective improvement within the fixed policy class lower-bounds the original excess; data aggregation provides such recovery evidence.

Related Work. A central question in representation-based control is which distinctions between states must be preserved for good decisions. State-abstraction theory characterizes information that can be discarded while preserving optimal behavior (Li et al., 2006), and approximate abstraction bounds the resulting performance loss (Abel et al., 2016). Bisimulation metrics relate behavioral similarity to differences in optimal value (Ferns et al., 2004). Value equivalence characterizes models through their Bellman updates over specified policy and value-function classes (Grimm et al., 2020), while sensor-based control limits bound achievable performance given the available observations (Majumdar et al., 2023). For hierarchical goal-conditioned control, Hyeon et al. (2026) show that value-sufficient goal representations need not retain the information required for optimal

action selection. Within this line of work, our focus is the relation between fixed-distribution decision quality and attainable closed-loop return.

Modern robot representation and world-model studies provide empirical motivation for this distinction. Emergent segmentation predicts robustness to visual shifts (Burns et al., 2025), and manipulation centricity correlates with downstream success (Jiang et al., 2025). DINO-WM and JEPA-WM studies evaluate learned latent dynamics through planning performance (Zhou et al., 2025; Tervier et al., 2026). WorldGym uses action-conditioned video rollouts for policy evaluation (Quevedo et al., 2026), while GigaWorld-1 emphasizes long-horizon action fidelity for reliable evaluation (GigaWorld Team et al., 2026). These findings motivate evaluating representations through execution, while leaving the representation-level limit distinct from the performance achieved by a particular learned controller.

Cost-sensitive reductions (Langford & Zadrozny, 2005) and performance-difference identities (Kakade & Langford, 2002) provide tools for connecting local decision loss to return. Loss-class discrepancies in domain adaptation (Mansour et al., 2009) offer a related perspective on measuring distribution shift through control-relevant losses. Offline-RL theory establishes limitations involving coverage and function approximation (Chen & Jiang, 2019; Foster et al., 2022), while DAgger addresses learner-induced shift through interactive data aggregation (Ross et al., 2011). These tools inform our analysis of representation sufficiency and the distinction between information deficit and learned-policy excess.

2 PRELIMINARIES

Control objective. Consider an MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma, \rho_0)$ with bounded rewards, $0 < \gamma < 1$, and return $J(\pi) = \mathbb{E}_{\rho_0,\pi} \sum_{t\geq 0} \gamma^t r(S_t, A_t)$. Let $V^\star, Q^\star$ be the optimal values and $J^\star = \mathbb{E}_{\rho_0} V^\star(S_0)$. Finite-horizon processes are represented by time-augmented states and zero-reward absorbing terminals. Theorem 2 separately uses an undiscounted horizon H.

Representation and policy class. A representation $R : \mathcal{S} \to \mathcal{Z}$ specifies the information available for choosing an action. Let Π_R be the class of all R-measurable policies, and let $\pi_R^\star \in \arg\max_{\pi\in\Pi_R} J(\pi)$ be the best policy available under R. Accordingly, $J(\pi_R^\star)$ is the best return achievable with the information provided by R. A learned policy $\hat{\pi}_R$ is a member of Π_R.

The state includes any history needed to describe the process, and R contains precisely the time, history, and observations exposed to the controller. A macro action is a temporally extended control option selected by the high-level policy and executed by a fixed internal controller over multiple primitive steps. These conventions specify the policy class being compared. The finite-process analysis assumes attained extrema; Appendix A.1.1 gives the approximate-minimizer argument.

Action-value gap. Define

$$\mathcal{A}^\star(s) = \arg\max_a Q^\star(s,a), \qquad L(s,a) = V^\star(s) - Q^\star(s,a) \geq 0. \tag{1}$$

The gap is zero for every optimal action and measures the return loss from taking a once and then continuing optimally. Thus L already reflects future recovery opportunities. We use $\Delta_{\min}$ and $\Delta_{\max}$ for lower and upper bounds on positive action-value gaps over the support specified in each result.

Evaluation distributions. A distribution μ specifies where decisions are assessed; d^π denotes normalized discounted policy occupancy. The demonstration-side distribution is denoted d_{exp}, with its empirical sampling convention stated in each experiment. The complete law of recorded demonstration trajectories is $\mathcal{L}_D$. This distinction separates the distribution used to assess decision risk from the recorded information available for offline identification.

3 REPRESENTATION SUFFICIENCY FOR CONTROL

A representation affects both the actions available to a controller and the states it visits in closed loop. We first separate representation-induced loss from the suboptimality of a learned policy, then relate fixed-distribution decision risk to attainable return through policy occupancy. Finally, we show that the remaining learned-policy excess cannot be identified from recorded demonstrations alone.

Lemma 1 (Information deficit and learned-policy excess)**.** *For every learned policy $\hat{\pi}_R \in \Pi_R$,*

$$J^\star - J(\hat{\pi}_R) = \underbrace{J^\star - J(\pi_R^\star)}_{\varepsilon_{\text{info}}(R)} + \underbrace{J(\pi_R^\star) - J(\hat{\pi}_R)}_{\varepsilon_{\text{dyn}}(R,\hat{\pi}_R)}. \tag{2}$$

Both terms are nonnegative.

The proof is given in Appendix A.1. The first term is the *information deficit*: the loss that remains even for the best policy compatible with the representation. The second is the *learned-policy excess*: the performance still recoverable within the same policy class.

3.1 Value-Weighted Decision Risk

We first evaluate the representation on a fixed state distribution μ. Define the optimal representation-constrained decision risk

$$\mathcal{I}_R(\mu) = \inf_{f:\mathcal{Z}\to\mathcal{A}} \mathbb{E}_{S\sim\mu} L(S, f(R(S))). \tag{3}$$

For finite actions, define the corresponding Bayes optimal-action error

$$P_e^\star(R,\mu) = \mathbb{E}_{Z\sim R_{\#}\mu}\left[1 - \max_a \Pr\{a \in \mathcal{A}^\star(S) \mid Z\}\right]. \tag{4}$$

All optimal actions are counted as correct, so this quantity measures ambiguity in optimal action choice rather than disagreement with a particular demonstrator.

Theorem 1 (Action-Value-Gap Bounds for Decision Risk)**.** *Suppose the action space is finite and every strictly suboptimal action on the support of μ satisfies*

$$0 < \Delta_{\min} \le L(s,a) \le \Delta_{\max} < \infty.$$

Then

$$\Delta_{\min} P_e^\star(R,\mu) \le \mathcal{I}_R(\mu) \le \Delta_{\max} P_e^\star(R,\mu). \tag{5}$$

Consequently, the two risks vanish together. If all actions are optimal on the support of μ, both are zero without a positive-gap assumption.

The proof is given in Appendix A.1.1. The result shows that action ambiguity alone does not determine decision risk: the same Bayes error can have very different consequences depending on the associated action-value gaps.

If $R_c = g \circ R_f$, then

$$\mathcal{I}_{R_c}(\mu) \ge \mathcal{I}_{R_f}(\mu), \qquad P_e^\star(R_c,\mu) \ge P_e^\star(R_f,\mu).$$

For continuous actions, the same argument applies whenever a nonnegative action loss ℓ satisfies $c_1\ell(s,a) \le L(s,a) \le c_2\ell(s,a)$ on the relevant support; see Appendix A.1.1.

3.2 The Occupancy Bridge

Fixed-distribution decision risk does not by itself determine closed-loop return, because the policy may visit a different state distribution. Define the policy occupancy and expected action-value gap by

$$d^\pi(s) = (1-\gamma)\sum_{t\ge 0}\gamma^t \Pr_{\rho_0}^{\pi}(S_t = s), \qquad \ell_\pi(s) = \mathbb{E}_{a\sim\pi(\cdot|R(s))} L(s,a). \tag{6}$$

The performance-difference identity (Kakade & Langford, 2002) gives

$$J^\star - J(\pi) = \frac{\mathbb{E}_{d^\pi}\ell_\pi}{1-\gamma}. \tag{7}$$

Thus, $\varepsilon_{\text{info}}$ depends on the occupancy of $\pi_R^\star$, whereas $\mathcal{I}_R(\mu)$ is optimized under the fixed distribution μ.

Proposition 1 (Occupancy bridge)**.** *Let $f^\star_\mu$ minimize $\mathcal{I}_R(\mu)$, with its actions specified also outside the support of μ. Define*

$$\eta^- = \mathrm{TV}(d^{\pi^\star_R}, \mu), \qquad \eta^+ = \mathrm{TV}(d^{f^\star_\mu}, \mu). \tag{8}$$

If $0 \le L(s,a) \le \Delta_{\max}$ for all actions on the union of the three distribution supports, then

$$\frac{\mathcal{I}_R(\mu) - \Delta_{\max}\eta^-}{1-\gamma} \le \varepsilon_{\mathrm{info}}(R) \le \frac{\mathcal{I}_R(\mu) + \Delta_{\max}\eta^+}{1-\gamma}. \tag{9}$$

If $\eta^- = \eta^+ = 0$, both bounds equal $\mathcal{I}_R(\mu)/(1-\gamma)$.

The proof is given in Appendix A.2. The lower and upper correction terms measure the mismatch between the reference state distribution μ and the occupancies induced by $\pi^\star_R$ and $f^\star_\mu$, respectively.

Corollary 1 (Bayes Error and Information Deficit)**.** *Under the assumptions of Theorem 1 and Proposition 1,*

$$\max\left\{0, \frac{\Delta_{\min}P^\star_e - \Delta_{\max}\eta^-}{1-\gamma}\right\} \le \varepsilon_{\mathrm{info}} \le \frac{\Delta_{\max}(P^\star_e + \eta^+)}{1-\gamma}, \tag{10}$$

where $P^\star_e = P^\star_e(R, \mu)$.

Choosing $\mu = d_{\exp}$ gives the demonstration-side version of the bound. Hence low demonstration risk implies low information deficit only when the relevant occupancy shift is also controlled.

3.3 Localized Bounds and Risk–Return Decomposition

The global TV correction treats all shifted probability mass as if it incurred the largest action loss. A tighter correction accounts for where the shift occurs. For finite states, define

$$W^-(\nu,\mu) = \sum_s [\mu(s) - \nu(s)]_+ \Delta(s), \qquad W^+(\nu,\mu) = \sum_s [\nu(s) - \mu(s)]_+ \Delta(s), \tag{11}$$

where $\Delta(s) = \max_a L(s,a)$, and

$$D^-_R(\nu,\mu) = \left[\sup_{\pi\in\Pi_R} (\mathbb{E}_\mu - \mathbb{E}_\nu)\ell_\pi\right]_+, \qquad D^+_R(\nu,\mu) = \left[\sup_{\pi\in\Pi_R} (\mathbb{E}_\nu - \mathbb{E}_\mu)\ell_\pi\right]_+. \tag{12}$$

Proposition 2 (Localized correction)**.** *Under the assumptions of Proposition 1,*

$$\frac{\mathcal{I}_R(\mu) - D^-_R(d^{\pi^\star_R}, \mu)}{1-\gamma} \le \varepsilon_{\mathrm{info}} \le \frac{\mathcal{I}_R(\mu) + D^+_R(d^{f^\star_\mu}, \mu)}{1-\gamma}. \tag{13}$$

Moreover,

$$0 \le D^\pm_R \le W^\pm \le \Delta_{\max}\,\mathrm{TV}(\nu,\mu). \tag{14}$$

The proof is given in Appendix A.3. The refinement removes occupancy differences that do not affect any action-value-gap function realizable by Π_R.

This can yield a strict improvement over the global TV correction. In the two-state example of Appendix A.3, all shifted mass lies on zero-gap terminal states, so $D^\pm_R = W^\pm = 0$ even though the total variation is positive. The localized bounds therefore recover the exact information deficit, while the global lower bound can be vacuous.

The same distinction can be expressed exactly. Define

$$B_R(\pi;\mu) = \mathbb{E}_\mu \ell_\pi - \mathcal{I}_R(\mu) \ge 0, \tag{15}$$

and

$$C_R(\pi;\mu) = (\mathbb{E}_{d^\pi} - \mathbb{E}_\mu)\ell_\pi. \tag{16}$$

Then

$$(1-\gamma)\varepsilon_{\mathrm{info}} = \mathcal{I}_R(\mu) + B_R(\pi^\star_R;\mu) + C_R(\pi^\star_R;\mu). \tag{17}$$

Here B_R measures the decision-objective mismatch under the reference distribution, while C_R captures the signed change in expected action-value gap induced by the occupancy shift. Together, they separate two mechanisms by which a small decision risk can coexist with a large closed-loop information deficit. The individual terms are defined for a complete policy; Section 4.3 examines their sensitivity to off-path policy completion on the exact robot graphs.

3.4 Offline Non-Identifiability of Learned-Policy Excess

The previous results concern $\varepsilon_{\mathrm{info}}$. The remaining term $\varepsilon_{\mathrm{dyn}}$ is fundamentally different: even complete recorded demonstrations need not determine the performance of a deployed policy.

The following result uses an undiscounted finite horizon H and the complete recorded-demonstration law $\mathcal{L}_D$.

Theorem 2 (Offline non-identifiability)**.** *For every finite horizon $H \geq 4$, there exist two MDPs with rewards in $[0, 1]$, the same representation R, the same recorded-demonstration law $\mathcal{L}_D$, and the same deployed policy $\hat{\pi}_R$, such that $\varepsilon_{\mathrm{info}} = 0$ in both, while*

$$\varepsilon_{\mathrm{dyn}}(M_1, \hat{\pi}_R) = 0, \qquad \varepsilon_{\mathrm{dyn}}(M_2, \hat{\pi}_R) = H - 3. \tag{18}$$

Thus $\varepsilon_{\mathrm{dyn}}$ is not identifiable from $(R, \mathcal{L}_D, \hat{\pi}_R)$, and no upper bound based only on this information can be uniformly $o(H)$.

The proof is given in Appendix A.4. Both environments admit an optimal R-policy, yet an unobserved transition changes the return of the same deployed policy without changing the recorded demonstrations. Additional coverage or knowledge of the dynamics is therefore required to resolve the learned-policy excess.

Section 4 therefore uses two evaluation regimes: exact finite-process optimization for representation-level quantities, and execution-based probes for learned-policy behavior.

4 Experiments

The experiments follow the distinction between decision risk, attainable return, and learned-policy excess. Exact finite models examine the action-gap bounds and the occupancy bridge. Learned-policy experiments then assess prediction–execution associations and performance recovered within a fixed information and action interface.

4.1 Evaluation Questions and Protocols

We ask how Bayes error relates to decision risk, when decision risk differs from the information deficit, and what execution reveals about a learned policy beyond offline prediction. The first two questions use exact values and policy-class optimization; the third uses fitted predictors, reference controllers, and data-aggregation interventions.

The robot finite processes are induced from robomimic Lift snapshots (Mandlekar et al., 2022). Each of four development base blocks supplies 16 object-motion directions with the robot initialization held fixed. Five nested representations, $R_{16}, R_8, R_4, R_2, R_{\mathrm{blind}}$, progressively merge direction cells. The high-level controller selects from a fixed dictionary of 20 macro actions and receives base identity, the representation cell, macro time, and the public action prefix. Low-level feedback is fixed across representations (Appendix B.2).

We use two complementary protocols. *H2* has two macro decisions and measures risk on the uniform distribution over the 64 initial mode states. *H3* has three macro decisions and measures restricted optimal returns and discounted occupancies. They differ in dynamics and reward timing as well as depth. H2 uses terminal-success and primitive-duration-discounted objectives; H3 uses $\gamma = 0.99$ per macro decision. All reported optima refer to the declared finite control interface. Appendix B records graph construction, solver checks, and protocol selection.

Learned-policy diagnostics use robomimic Lift, and recovery experiments use LIBERO-Object (Liu et al., 2023). Their measurements are summarized over the stated policy seeds or task–seed settings, with shared evaluation initial conditions within each protocol. Appendices C–E give the training and evaluation details.

4.2 Fixed-Distribution decision risk

H2 evaluates $P_e^\star$ and $\mathcal{I}_R$ on the same root-state distribution, treating every optimal action as correct. For the terminal-success objective, all positive root gaps equal one, so Theorem 1 gives equality.

Table 1: H2 fixed-distribution decision risks on four development blocks. Terminal-success risks coincide. The primitive-duration-discounted objective has its own optimal-action sets and gap bounds.

Representation	Terminal $P_e^\star = \mathcal{I}_R$	Discounted $\mathcal{I}_R$	$\Delta_{\min} P_e^\star$	$\Delta_{\max} P_e^\star$
R_{16}	0.000	0.000	0.000	0.000
R_8	0.406	0.293	0.275	0.319
R_4	0.656	0.472	0.443	0.516
R_2	0.781	0.560	0.528	0.614
R_{blind}	0.844	0.608	0.570	0.663

Both risks increase from zero to 0.844 as direction information is removed (Table 1). The six modes with zero full-state success value remain in the evaluation support; the full-state success coverage is $58/64$. Primitive-duration discounting changes the positive gaps to approximately $[0.676, 0.786]$. The separately minimized decision risks remain within the corresponding bounds, with zero recorded violation under either objective. These are applications of the pointwise action-loss comparison; the discounted occupancy bridge is evaluated separately below. Complementary 12×12 gridworld calculations have positive gaps of $[0.0107, 0.857]$ in an irreversible regime and $[0.017, 0.0975]$ in a recoverable regime. Both satisfy the bounds across the recorded representation sweep (Appendix B.1), illustrating the role of action-value gap beyond Bayes optimal-action error.

4.3 FROM DECISION RISK TO ATTAINABLE RETURN

An exact four-state MDP first isolates the zero-shift case of Proposition 1. Actions affect rewards but not transitions, so all policies share the reference occupancy. At $\gamma = 0.9$, the coarse representations yield $\mathcal{I}_R = 0.448$ and $\varepsilon_{\text{info}} = 4.48$; both sides of the bridge attain the same nonzero deficit. Appendix B.1 reports the enumeration and numerical checks.

H3 then examines what changes when state visitation depends on the policy. Exhaustive optimization gives a full-state macro-discounted return of 0.990 and a best blind return of 0.291, yielding an information deficit of

$$\varepsilon_{\text{info}} = 0.699.$$

Throughout the H3 analysis, μ denotes the normalized expert occupancy, and we write $\mathcal{I}_R$ for $\mathcal{I}_R(\mu)$. Expressed in return units, the corresponding decision risk is $\mathcal{I}_R/(1-\gamma) = 0.0246$, only 3.52% of the information deficit. Thus, small decision risk under the expert occupancy can coexist with a large loss in attainable closed-loop return. Because both returns are exact optima within the same fixed macro-action interface, this gap is the information deficit rather than learned-policy excess.

Across the four development base blocks, blind information deficits range from 0.684 to 0.744. All 64 graphs pass complete-action and Bellman audits, every block has full-state success coverage, and no block admits a common optimal sequence across all 16 modes.

Table 2: Exact H3 information deficits and lower-bound endpoints at $\gamma = 0.99$. $\mathcal{I}_R/(1-\gamma)$ is evaluated under the normalized expert occupancy. L_G, L_W, and L_D are the global-TV, statewise W^-, and policy-class D_R^- lower-bound endpoints, respectively, evaluated at the same closest-TV return-optimal policy. Negative values are not clipped.

Representation	$\varepsilon_{\text{info}}$	$\mathcal{I}_R/(1-\gamma)$	L_G	L_W	L_D
R_{16}	0.0000	0.0000	0.0000	0.0000	0.0000
R_8	0.0042	0.0022	-0.8381	-0.4245	-0.4239
R_4	0.0404	0.0080	-1.8423	-0.8774	-0.8619
R_2	0.4539	0.0145	-1.7765	-0.8116	-0.5409
R_{blind}	0.6989	0.0246	-1.8515	-0.8866	-0.5075

Localized lower bounds. Table 2 compares the global (L_G), statewise (L_W), and policy-class (L_D) lower endpoints at the same closest-TV return optimum. For R_{blind}, localized correction raises the endpoint from -1.8515 to -0.8866 and then to -0.5075, reducing the global correction penalty by

71.6%. The additional $W^- \to D_R^-$ improvement is 0.3791 in return units and persists when success and failure terminals are encoded separately.

The localized correction therefore provides a materially tighter lower bound on the robot graphs, although all four coarsened representations retain negative lower endpoints. Directly minimizing D_R^- over return-optimal policies further raises the blind endpoint to -0.4503, but does not make it positive. Appendix B.5 reports the policy-selection and sensitivity analyses.

Decision-objective mismatch and occupancy shift. The exact decomposition in Equation (17) explains why $\mathcal{I}_R/(1-\gamma)$ is much smaller than the information deficit. For the closest-TV blind policy with risk-minimizing off-path completion, the scaled terms are

$$\frac{\mathcal{I}_R}{1-\gamma} = 0.0246, \qquad \frac{B_R}{1-\gamma} = 0.3438, \qquad \frac{C_R}{1-\gamma} = 0.3305.$$

Together they recover the exact information deficit (Figure 2c). Here B_R measures the decision-objective mismatch under the reference distribution, while C_R is the signed occupancy-shift term. Both effects are substantial for the blind representation. Neither term is learned-policy excess: the policy already maximizes return within Π_R.

Dependence on policy completion. The split between B_R and C_R is defined for a complete policy and can depend on its off-path actions. Replacing only the unreachable-history actions by uniform randomization leaves executed occupancy, return, and $B_R + C_R$ unchanged, while changing the blind scaled terms to $B_R/(1-\gamma) = 0.8931$ and $C_R/(1-\gamma) = -0.2188$. Figure 2 uses the risk-minimizing completion; Appendix B.5 reports the completion and terminal-encoding sensitivity analyses.

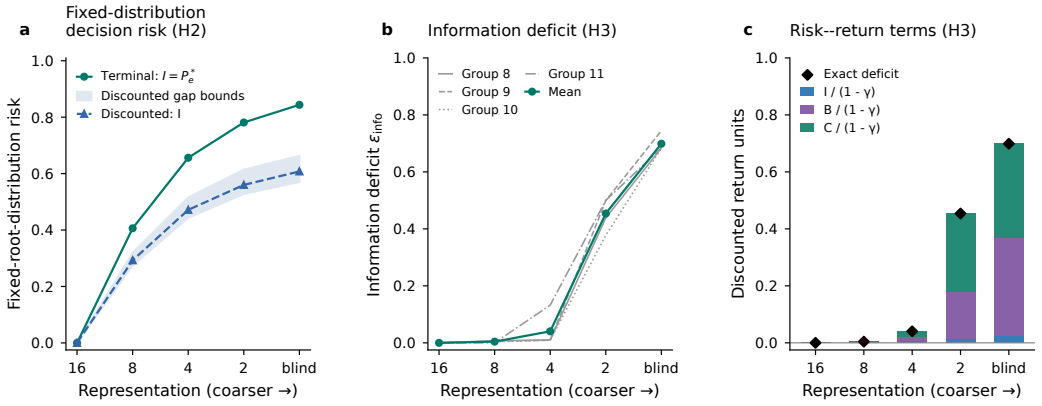

Figure 2: Exact decision risk and information deficit. (a) H2 root-state risks and primitive-duration-discounted action-gap bounds. (b) H3 information deficits for four development blocks and their equal-weight mean. (c) Exact H3 decomposition for closest-TV return optima with risk-minimizing off-path completion: bars show the scaled $\mathcal{I}_R$, B_R, and signed C_R terms, and diamonds mark their sum, the exact information deficit. Because C_R is signed, the R_8 occupancy-shift term is slightly negative. H2 and H3 use different dynamics, reward timing, and decision depth.

4.4 Learned-Policy Diagnostics and Within-Class Improvement

Aligned prediction and execution. On Lift, action prediction and teacher-continuation loss are evaluated from the same 30 held-out states, stratified across four trajectory phases. The representations are $R_\delta = (R_{\text{prop}}, Q_\delta(p_{\text{obj}}))$, with $\delta \in \{0, 0.5, 2, 5, \infty\}$ cm and three policy seeds. A random forest predicts demonstration actions, yielding normalized RMSE E_{act}. For $K \in \{1, 8, 32, 64, 128, 200\}$, the student acts for up to K primitive steps before a full-state BC teacher completes the remaining budget of $H = 200$. The teacher has a richer object input than the student, including the $\delta = 0$ setting. The loss L_K is the teacher-only discounted first-success value minus the handoff value, with $\gamma = 0.99$ per primitive step (Appendix C).

Mean action errors vary only from 0.544 to 0.563, whereas the mean full-horizon loss is 0.0583 with full-precision object position and 0.259 when object position is removed; the intermediate quantization levels are not monotone. Across the five representations, the mean of the three within-seed Spearman correlations between E_{act} and L_K changes from -0.133 at $K = 1$ to 0.767 at $K = 200$. The association therefore strengthens overall with longer autonomous execution, although not monotonically across all horizons. These quantities are empirical diagnostics rather than estimates of $P_e^\star$, $\mathcal{I}_R$, or ϵ_{info}.

Policy improvement within a fixed interface. DAgger-style aggregation (Ross et al., 2011) labels learner-visited states with a full-state BC teacher while keeping the student's representation, policy class, action interface, and evaluation objective fixed. LIBERO-Object experiments use object-blind robot state, a frozen 16-dimensional learned state representation, and frozen ResNet-18 visual features (Appendix D). Table 3 reports baseline and final-iteration success under the evaluation protocol associated with each setting.

Table 3: Fixed-interface recovery on LIBERO-Object. Entries are empirical mean success rates. Replication units and evaluation budgets differ across rows; each gain compares baseline with the final iteration within its own protocol.

Student input	Runs	Episodes/run	Horizon	BC	Final
Object-blind state	5 seeds	15	500	0.000	0.493
Frozen state latent	9 task–seed runs	12	300	0.111	0.278
Frozen visual features	3 seeds	12	300	0.083	0.333

For the success-probability objective, the corresponding mean observed gains are 0.493, 0.167, and 0.250. Equation (20) implies that a positive expected gain under the same objective lower-bounds the original policy excess; the values in Table 3 are finite-sample estimates of such gains. These interventions therefore provide evidence of performance recoverable within the fixed information and action interface. The frozen-latent best-iteration result is reported separately in Appendix D.

Cross-distribution transfer. Appendix E evaluates a common action predictor trained on demonstrations at policy-induced states. Object-blind transfer excess exceeds the full-state value in 18 of 20 task–seed pairs, which provides a separate test of sensitivity to distribution shift. This analysis does not estimate the oracle occupancy corrections. Chunked-execution comparisons, which change both the action interface and evaluation budget, are reported separately in Appendix D.

5 Discussion and Conclusion

Representation sufficiency for control depends on more than action prediction on demonstrated states. The consequence of an ambiguity depends on its action-value gap and whether the policy visits states where it matters. Fixed-distribution decision risk and the occupancy bridge make both effects explicit. In the exact robot graphs, removing information that contributes little risk under expert occupancy can still substantially reduce the best attainable return.

For learned controllers, a separate question is how much performance remains recoverable within the same policy class. Exact optimization isolates the loss imposed by the representation, but offline demonstrations alone need not determine a learned policy's gap to the best policy available under that representation. Aligned execution probes and fixed-interface policy-improvement experiments provide complementary evidence about this gap.

Our exact robot analysis covers four development base blocks and a finite macro-action interface, while the learned-policy experiments use simulated manipulation tasks. Localized occupancy corrections substantially tighten the global bounds on these graphs, though the lower bounds remain loose. Further work should extend the analysis to broader state coverage and richer control interfaces. Evaluating representations for control requires distinguishing the return that available information can support from the return that learning actually achieves.

AI USE STATEMENT

Large language models were used to assist with language editing, manuscript organization, literature search and citation checking, mathematical discussion, and code and result inspection. The empirical results reported here come from the recorded experiments. Responsibility for the mathematical statements, citations, experimental interpretation, and final manuscript rests with the authors.

REPRODUCIBILITY STATEMENT

Proofs are provided in Appendix A, and experimental protocols in Appendices B–E. The accompanying package provides a tested CPU workflow for the key theoretical checks, exact H2/H3 analyses, and recorded-result summaries. Its README documents dependencies, usage, provenance, and scope; simulator data generation and policy retraining are not included.

REFERENCES

David Abel, David Hershkowitz, and Michael Littman. Near optimal behavior via approximate state abstraction. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48, pp. 2915–2923. PMLR, 2016.

Kaylee Burns, Zach Witzel, Jubayer Ibn Hamid, Tianhe Yu, Chelsea Finn, and Karol Hausman. What makes pre-trained visual representations successful for robust manipulation? In *Proceedings of The 8th Conference on Robot Learning*, volume 270, pp. 4525–4545. PMLR, 2025.

Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 1042–1051. PMLR, 2019.

Norm Ferns, Prakash Panangaden, and Doina Precup. Metrics for finite Markov decision processes. In *Proceedings of the Twentieth Conference on Uncertainty in Artificial Intelligence*, pp. 162–169. AUAI Press, 2004. URL https://arxiv.org/abs/1207.4114.

Dylan J. Foster, Akshay Krishnamurthy, David Simchi-Levi, and Yunzong Xu. Offline reinforcement learning: Fundamental barriers for value function approximation. In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178, pp. 3489–3489. PMLR, 2022.

GigaWorld Team, Angyuan Ma, Boyuan Wang, Bohan Li, Chaojun Ni, Guo Li, Guan Huang, Guosheng Zhao, Hao Li, Hengtao Li, Jingyu Liu, Jiwen Lu, Qiuping Deng, Tingdong Yu, Xuancheng Xu, Xinyu Zhou, Xiuwei Xu, Xinze Chen, Xiaofeng Wang, Xiaoyu Tian, Yang Wang, Yifan Chang, Yukun Zhou, Yun Ye, Zhenyu Wu, Zhanqian Wu, and Zheng Zhu. GigaWorld-1: A roadmap to build world models for robot policy evaluation. *arXiv preprint arXiv:2607.02642*, 2026.

Christopher Grimm, André Barreto, Satinder Singh, and David Silver. The value equivalence principle for model-based reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5541–5552, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Jinu Hyeon, Woobin Park, Hongjoon Ahn, and Taesup Moon. Action-sufficient goal representations. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Guangqi Jiang, Yifei Sun, Tao Huang, Huanyu Li, Yongyuan Liang, and Huazhe Xu. Robots pre-train robots: Manipulation-centric robotic representation from large-scale robot datasets. In *International Conference on Learning Representations*, pp. 81885–81905, 2025.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*, pp. 267–274, 2002.

John Langford and Bianca Zadrozny. Relating reinforcement learning performance to classification performance. In *Proceedings of the 22nd International Conference on Machine Learning*, pp. 473–480, 2005.

Lihong Li, Thomas J. Walsh, and Michael L. Littman. Towards a unified theory of state abstraction for MDPs. In *Proceedings of the Ninth International Symposium on Artificial Intelligence and Mathematics*, 2006. URL https://thomasjwalsh.net/pub/aima06Towards.pdf.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. LIBERO: Benchmarking knowledge transfer for lifelong robot learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 44776–44791, 2023.

Anirudha Majumdar, Zhiting Mei, and Vincent Pacelli. Fundamental limits for sensor-based robot control. *The International Journal of Robotics Research*, 42(12):1051–1069, 2023.

Ajay Mandlekar, Danfei Xu, Josiah Wong, Soroush Nasiriany, Chen Wang, Rohun Kulkarni, Li Fei-Fei, Silvio Savarese, Yuke Zhu, and Roberto Martín-Martín. What matters in learning from offline human demonstrations for robot manipulation. In *Proceedings of the 5th Conference on Robot Learning*, volume 164, pp. 1678–1690. PMLR, 2022.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009.

Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. CALVIN: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *IEEE Robotics and Automation Letters*, 7(3):7327–7334, 2022.

Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3M: A universal visual representation for robot manipulation. In *Proceedings of The 6th Conference on Robot Learning*, volume 205, pp. 892–909. PMLR, 2023.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Julian Quevedo, Ansh Kumar Sharma, Yixiang Sun, Varad Suryavanshi, Percy Liang, and Sherry Yang. WorldGym: World model as an environment for policy evaluation. In *International Conference on Learning Representations*, pp. 78932–78957, 2026.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15, pp. 627–635. PMLR, 2011.

Basile Tervier, Tsung-Yen Yang, Jean Ponce, Adrien Bardes, and Yann LeCun. What drives success in physical planning with joint-embedding predictive world models? *Transactions on Machine Learning Research*, 2026.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. DINO-WM: World models on pre-trained visual features enable zero-shot planning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 79115–79135. PMLR, 2025.

APPENDIX

Appendix Contents

A. Proofs and Supporting Results 12
A.1 Return-Loss Decomposition and Decision Risk 12
A.2 Proof of Proposition 1 13
A.3 Localized Bounds and Risk–Return Decomposition 13
A.4 Proof of Theorem 2 14
A.5 Approximate Values, Execution Probes, and Policy Improvement 14
B. Exact Evaluation Protocols 14
B.1 Finite-MDP Evaluation 14
B.2 Robot Counterfactuals and Action Interface 15
B.3 H2/H3 Objectives and Exact Optimization 12
B.4 Occupancy Computation and Tie-Breaking 12
B.5 Localized Bounds, Decomposition, and Sensitivity 16
C. Aligned Prediction–Execution Protocol 19
D. Policy Improvement and Execution Protocols 19
E. Additional Representation Diagnostics 20
E.1 Static Probes and Cross-Distribution Transfer 20

A PROOFS AND SUPPORTING RESULTS

A.1 RETURN-LOSS DECOMPOSITION AND DECISION RISK

Equation (2) decomposes the return loss of a learned policy into the information deficit and the learned-policy excess, using the best policy available under R as the intermediate reference.

Proof. Add and subtract $J(\pi_R^\star)$, and use optimality in the full-state and restricted policy classes, respectively. □

A.1.1 PROOF OF THEOREM 1

For a decision rule f, define $e_\mu(f) = \Pr_\mu\{f(R(S)) \notin \mathcal{A}^\star(S)\}$ and $r_\mu(f) = \mathbb{E}_\mu L(S, f(R(S)))$. Conditioning on $Z = R(S)$ shows that minimizing e_μ selects an action maximizing $\Pr\{a \in \mathcal{A}^\star(S) \mid Z\}$, which gives Equation (4). The events associated with different optimal actions can overlap; their conditional probabilities need not sum to one.

Proof. For every rule f, the pointwise loss is bounded above and below by the corresponding gap constant times $\mathbf{1}\{f(R(s)) \notin \mathcal{A}^\star(s)\}$. Taking expectations and then the infimum proves the lower bound. Evaluating the decision risk at a Bayes-error minimizer proves the upper bound. The Bayes-error and decision-risk minimizers need not coincide. □

The upper and lower bounds apply to different optimizations. Writing f_e for a Bayes-error minimizer and f_L for a decision-risk minimizer makes the upper argument explicit: $r_\mu(f_L) \le r_\mu(f_e) \le \Delta_{\max} e_\mu(f_e)$. For the lower argument, every f has $r_\mu(f) \ge \Delta_{\min} P_e^\star$. If an infimum is not attained, choose an ϵ-minimizer and let $\epsilon \downarrow 0$. For randomized rules the conditional risk is linear in the action probabilities, so randomization cannot improve the minimum one-decision risk in a finite action space. The zero-risk equivalence follows when the lower gap is positive.

Representation refinement and continuous actions. When $R_c = g \circ R_f$, every R_c-measurable decision rule is also R_f-measurable, giving both risk monotonicity inequalities. This condition concerns partitions, not the numerical ordering of quantization widths. For continuous actions, the same proof applies when the comparison loss ℓ satisfies $c_1\ell(s,a) \le L(s,a) \le c_2\ell(s,a)$.

A.2 Proof of Proposition 1

The Bellman equation gives

$$L(s,a) = V^\star(s) - r(s,a) - \gamma \mathbb{E}[V^\star(S') \mid s,a].$$

Summing along π through time $T-1$ telescopes:

$$\mathbb{E}_\pi \sum_{t=0}^{T-1} \gamma^t L(S_t, A_t) = \mathbb{E}_{\rho_0} V^\star(S_0) - \mathbb{E}_\pi \sum_{t=0}^{T-1} \gamma^t r_t - \gamma^T \mathbb{E}_\pi V^\star(S_T).$$

Bounded rewards make the last term vanish as $T \to \infty$. The left-hand side equals $\mathbb{E}_{d^\pi}\ell_\pi/(1-\gamma)$, proving Equation (7).

For probability measures μ, ν and $0 \le h \le b$, $|(\mathbb{E}_\nu - \mathbb{E}_\mu)h| \le b\ \mathrm{TV}(\nu,\mu)$. The constant is b, rather than $2b$, because h is nonnegative and its range has width b. The union-support hypothesis ensures that it applies at states visited by either comparison policy.

Proof. For a function bounded in $[0, \Delta_{\max}]$, changing its distribution from μ to ν changes its expectation by at most $\Delta_{\max}\,\mathrm{TV}(\nu,\mu)$. Apply this to $\ell_{\pi_R^\star}$ in (7) and use $\mathbb{E}_\mu \ell_{\pi_R^\star} \ge \mathcal{I}_R(\mu)$ for the lower bound. For the upper bound, first use $J(\pi_R^\star) \ge J(f_\mu^\star)$, then apply the same argument to $f_\mu^\star$, for which $\mathbb{E}_\mu \ell_{f_\mu^\star} = \mathcal{I}_R(\mu)$. □

Corollary 1 follows by inserting the two sides of Theorem 1 into Proposition 1 and using $\varepsilon_{\mathrm{info}} \ge 0$. The rule $f_\mu^\star$ must be completed on unobserved cells before its occupancy is defined. When minimizers are tied, this completion and the tie rule are part of the bridge calculation.

A.3 Localized Bounds and Risk–Return Decomposition

Proof. Repeat the proof of Proposition 1, bounding each signed expectation difference by its defining supremum. Since $0 \le \ell_\pi(s) \le \Delta(s)$, discarding the negative contributions gives $D_R^\pm \le W^\pm$. Bounding $\Delta(s)$ by $\Delta_{\max}$ and using equality of the total probability masses gives the TV bound. □

For a finite observation partition, maximizing each observation-cell contribution independently yields

$$\sup_{\pi \in \Pi_R} (\mathbb{E}_\mu - \mathbb{E}_\nu)\ell_\pi = \sum_z \max_a \sum_{s: R(s)=z} (\mu(s) - \nu(s)) L(s,a).$$

Taking the positive part gives D_R^-; reversing μ and ν gives D_R^+. The formula uses the full decision-rule class on the specified partition. For a parametrically restricted class, the supremum is taken over that class instead.

In the terminal-shift example, initial states each have occupancy $(1-\gamma)/2$ under all policies. The full-state policy puts terminal mass γ at the good terminal. Any best restricted rule puts $\gamma/2$ at each terminal, so TV is $\gamma/2$. Both terminal action gaps are zero. It follows that $W^\pm = D_R^\pm = 0$, even though TV is positive. The decision risk is $(1-\gamma)/2$, so the localized interval collapses to $1/2$ and the global lower endpoint equals $(1-2\gamma)/(2(1-\gamma))$.

For completeness, adding and subtracting $\mathbb{E}_\mu \ell_\pi$ in the performance identity yields

$$(1-\gamma)(J^\star - J(\pi)) = \mathcal{I}_R(\mu) + \big(\mathbb{E}_\mu \ell_\pi - \mathcal{I}_R(\mu)\big) + (\mathbb{E}_{d^\pi} - \mathbb{E}_\mu)\ell_\pi.$$

At $\pi = \pi_R^\star$ this is Equation (17). Its second term is a difference between decision objectives, and its third is a signed distribution difference. Both terms are evaluated for the selected optimal policy, including its off-support actions.

A.4 Proof of Theorem 2

Proof. Use states $\{s_0, s_A, s_B, g\}$, initial state s_0, and reward $r(s,a) = \mathbf{1}\{s = g\}$. Let $R(s_A) = R(s_B) = z$, with $R(s_0)$ and $R(g)$ distinct from z and each other. In both MDPs, a_0 takes s_0 to s_A, a takes s_A to g, and a' takes s_B to g. Unspecified actions at these three states self-loop. At g, $c^\star$ and all actions other than c self-loop in both environments; c self-loops in $\mathcal{M}_1$ but goes to s_B in $\mathcal{M}_2$. The distinct actions satisfy $a' \neq a$ and $c^\star \neq c$.

Record the expert path $s_0 \xrightarrow{a_0} s_A \xrightarrow{a} g$ and stop recording at the first visit to g, while the environment continues. The complete recorded laws, including rewards and censoring, are identical. Choose the deployed policy with actions a_0 at $R(s_0)$, a at z, and c at $R(g)$. Its return is $H-2$ in $\mathcal{M}_1$ and one in $\mathcal{M}_2$, where it becomes trapped in s_B. An R-policy that instead chooses $c^\star$ at g obtains $H-2$ in both, which is also the full-state optimum. Hence both information deficits are zero, while the learned-policy excesses differ by $H-3$. Any functional of the common offline triple must have the same output in the two environments, proving the assertion. □

The construction holds the law of the complete recorded trajectory fixed, including its stopping rule. Recording ends at the first visit to g; task evaluation continues for H action steps. This gives $H-2$ rewarding steps for a policy staying at g and one for the failing deployed policy. The unobserved transition under c is therefore sufficient to change the learned-policy excess without changing either recorded data or representation sufficiency. The result is about identification from this offline information; additional coverage or known dynamics can distinguish the environments.

A.5 Approximate Values, Execution Probes, and Policy Improvement

The exact finite-process experiments evaluate $Q^\star$, optimal-action sets, and restricted optima within the declared action and information interfaces. Learned continuous-control probes instead use a reference. For $\widehat{V} = \max_a \widehat{Q}$, if $\|\widehat{Q} - Q^\star\|_\infty \leq \alpha$ on the relevant support, then $|\max_a \widehat{Q}(s,a) - V^\star(s)| \leq \alpha$. The triangle inequality gives $|\widehat{L}(s,a) - L(s,a)| \leq 2\alpha$. Taking expectations and optimizing over the same decision-rule class yields $|\widehat{\mathcal{I}}_R - \mathcal{I}_R| \leq 2\alpha$. For a fixed policy followed by an optimal continuation after K decisions, Bellman telescoping gives

$$\mathbb{E}\left[V^\star(S_0) - \sum_{t=0}^{K-1} \gamma^t r_t - \gamma^K V^\star(S_K)\right] = \sum_{t=0}^{K-1} \gamma^t \mathbb{E} L(S_t, A_t). \tag{19}$$

Retaining the terminal $\gamma^K V^\star(S_K)$ gives the discounted cumulative action-value gap. Finally, for $\hat{\pi}_R, \tilde{\pi}_R \in \Pi_R$, subtraction gives

$$\Delta_{\mathrm{rec}} = J(\tilde{\pi}_R) - J(\hat{\pi}_R) = \varepsilon_{\mathrm{dyn}}(R, \hat{\pi}_R) - \varepsilon_{\mathrm{dyn}}(R, \tilde{\pi}_R). \tag{20}$$

Thus a positive gain lower-bounds the original excess for the same return objective.

B Exact Evaluation Protocols

B.1 Finite-MDP Evaluation

The gridworld experiment uses a 12×12 state grid at $\gamma = 0.95$ and computes action-value gaps, Bayes errors, and decision risks for each recorded quantization. The irreversible regime has positive gaps in $[0.0107, 0.857]$, and the recoverable regime has gaps in $[0.017, 0.0975]$. Both satisfy the bounds across the recorded representation sweep. The smaller maximum gap in the recoverable regime illustrates why Bayes error alone does not determine decision risk. The separate bridge experiment has four states, $\gamma = 0.9$, a uniform initial distribution over nonterminal states, and representation widths $1, 2, 4$. All deterministic stationary representation-measurable policies are enumerated. In the action-independent-transition regime, every policy has the same normalized occupancy, giving the equality case in Section 4.3. Finite-prefix identities are also checked at $K = 1, 2, 4, 8, 16, 32$. These calculations use the exact finite models and their declared policy classes. Across the recorded bridge regimes, all original two-sided bounds hold, with maximum performance-identity residual below 1.2×10^{-16}.

Table 4: H3 information deficits by development base block. These are four base blocks, not four independently trained policy seeds.

Base block	R_8	R_4	R_2	R_{blind}
8	0.007	0.010	0.439	0.684
9	0.006	0.009	0.499	0.744
10	0.004	0.010	0.377	0.684
11	0.000	0.132	0.500	0.684

B.2 Robot Counterfactuals and Action Interface

The source consists of 200 robomimic Lift proficient-human trajectories. A trajectory split with seed 9212027 assigns 160 to training, 15 to development, 15 to validation, and 10 to audit. A base selector chooses a non-grasp prefix state in the first half of a trajectory with end-effector height 0.07–0.14 m above the cube and sufficient table margin. Each base snapshot generates 16 modes at angles $2\pi m/16$, $m = 0, \ldots, 15$, while retaining the same robot initialization. Nested angular cells merge adjacent direction modes. The reported data use development base groups 8–11.

The action dictionary contains 16 directional interception options and four fixed controls: down-open, close, lift-close, and no-op-open. A directional option uses its selected direction, elapsed option time, robot end-effector position, and frozen task constants. Its low-level controller does not read live cube pose, cube velocity, contact, grasp, reward, or hidden mode. It closes after reaching the configured proximity conditions and terminates after four close-dwell steps, immediately before lifting, or on an absorbing environment terminal. Fixed controls last eight primitive steps; directional options have a maximum of 60. The frozen parameters are a 0.1 s lead, 0.012 m horizontal and 0.022 m vertical closing thresholds, proportional gain 20, task height 0.82 m, and option-motion time increment 0.00816 s. The simulator control frequency is 20 Hz; the option-motion increment is a separate controller parameter.

The high-level information interface comprises base identity, representation cell, macro time, and public action prefix, with terminal flags ending execution. Object position, velocity, and mode are hidden at the blind level. The option's proprioceptive feedback is fixed by its implementation. Exact high-level optimization therefore concerns this interface, rather than every possible continuous feedback controller built from the same sensors.

B.3 H2/H3 Objectives and Exact Optimization

H2 has two macro decisions. The root-risk analysis uses a uniform distribution over the 64 initial mode states and evaluates both terminal success and primitive-duration-discounted success. All root actions are retained, including the six modes for which every available continuation has zero success value. H3 has three decisions and uses the selected recoverable regime: initial radius 0.04 m, speed 0.05 m/s, and escape boundary 0.12 m. Its macro-discounted objective uses $\gamma = 0.99$ per macro decision. H2 and H3 differ in dynamics and reward timing as well as depth. The canonical correct-action and symmetric codebook-coverage thresholds are both 0.90. The H3 regime was selected on training group 0, evaluated on development group 8, and then repeated on groups 9–11; validation and audit splits remained unused.

Shared-prefix action trees are expanded exhaustively, and full-state values follow by backward dynamic programming. Within each H3 observation cell, the 20^3 macro sequences are evaluated under the conditional mode distribution. Sequence mixtures implement randomized choices within the same public-history interface. The 16 modes and 20 root actions give 320 computational shards per base block, 1,280 in total. Full action coverage, terminal handling, and Bellman residuals are checked on the resulting graphs.

B.4 Occupancy Computation and Tie-Breaking

Normalized occupancy includes the absorbing tail, and terminal states are merged for the primary H3 comparison. The reference is the frozen full-state expert occupancy. To compute the most

favorable global-TV lower endpoint among tied return optima, the analysis mixes return-optimal sequences within each observation cell and minimizes TV by a linear program. Each cell has its own normalization constraint; the joint objective evaluates the aggregate occupancy rather than averaging separately optimized group TV values. The resulting η^- values for $R_8, R_4, R_2, R_{\text{blind}}$ are 0.0084, 0.0185, 0.0179, and 0.0188. The support-wide upper action gap is one. Combining these corrections with the risk values gives Table 2. The performance-identity residual is below 1.8×10^{-15} and the LP objective reconstruction residual below 7×10^{-16}. The ratio of $\mathcal{I}_R/(1-\gamma)$ to the blind information deficit is

$$\frac{0.000246/(1-0.99)}{0.699} = 0.0352\ldots.$$

B.5 Localized Bounds, Decomposition, and Sensitivity

Scope and reference policy. This supplementary analysis reuses all 64 completed development graphs under the frozen K20/H3 protocol. It is an analysis of previously observed development instances. The original requirement of global-TV lower bounds above 0.01 at two or more coarsened levels remains unmet; validation and audit splits remain unused. Each initial mode has probability $1/64$. The expert randomizes uniformly over full-state optimal actions, using the original 10^{-8} optimal-action tolerance. Its normalized occupancy μ includes the absorbing tail. Throughout this subsection, $m = 1-\gamma = 0.01$ and $\Delta_{\max} = 1$ on the full graph.

Same-policy comparison and upper bounds. The main comparison fixes the closest-TV return optimum and evaluates $G_- = \Delta_{\max}\,\mathrm{TV}(d^{\pi^\star_R}, \mu)$, W^-, and D_R^- at that same occupancy. For the upper side, each observation cell selects an action minimizing $A_z(a) = \sum_{s:R(s)=z} \mu(s)L(s,a)$; ties use the smallest action ID. This rule also completes cells with zero reference mass. Forward propagation gives $d^{f^\star_\mu}$, at which G_+, W^+, D_R^+ are evaluated. The lower and upper occupancies therefore generally differ. Endpoints are $(\mathcal{I}_R - G_-)/m$, $(\mathcal{I}_R - W^-)/m$, $(\mathcal{I}_R - D_R^-)/m$, and their upper-side counterparts with plus signs. Table 5 reports the upper endpoints, along with a second risk minimizer that randomizes uniformly over minimizing actions. For R_{16}, the latter agrees with the expert on its visited states and gives a zero upper endpoint; the deterministic rule can have nonzero occupancy discrepancy despite the same optimal return. Neither rule is claimed to maximize return over all risk minimizers.

Table 5: H3 upper endpoints for the deterministic risk minimizer (smallest action ID at ties). U_D^{unif} instead uses uniform randomization over risk-minimizing actions, including zero-reference-mass cells. These choices need not maximize return among risk minimizers.

Representation	$J(f^\star_\mu)$	U_G	U_W	U_D	U_D^{unif}
R_{16}	0.9898	0.2430	0.2428	0.2428	0.0000
R_8	0.7886	0.6610	0.4598	0.4598	0.4464
R_4	0.4331	1.2826	0.7257	0.7243	0.6444
R_2	0.2475	1.5863	0.8438	0.8377	0.7836
R_{blind}	0.0153	1.9924	1.0943	1.0794	1.0521

Localized discrepancy and policy selection. We compute $D_R^\pm$ using the finite-partition formula in Appendix A.3, maximizing over all 20 actions in each public-history cell. The positive part is applied only after summing all cell maxima. The maximizing rule is a discrepancy witness; it need not generate the input occupancy or maximize return. To distinguish localized correction from policy selection, we separately minimize TV, W^-, and D_R^- over the same return-optimal sequence mixtures (Table 6). If their occupancy is $\nu(x) = Mx$, minimizing D_R^- is a linear program with free variables t_z and $d \geq 0$:

$$t_z \geq \sum_{s:R(s)=z} [\mu(s) - (Mx)(s)]L(s,a) \quad \text{for every } a, \qquad d \geq \sum_z t_z.$$

The objective is $\min d$, subject to the original per-cell mixture simplexes. Constraining each t_z to be nonnegative would compute a different discrepancy. Every retained sequence attained its cell's

maximum return in floating-point evaluation; no positive candidate shortfall was recorded. Even the best D_R^- lower endpoints remain negative for all four coarsened representations.

Table 6: Policy selection within the same return-optimal set. The first three columns give L_D after separately minimizing TV, W^-, and D_R^-. The last columns give the decomposition for the closest-TV policy with risk-minimizing off-path completion; $m = 1 - \gamma$.

Representation	$L_D^{\min \mathrm{TV}}$	$L_D^{\min W}$	$L_D^{\min D}$	B_R/m	C_R/m
R_{16}	0.000000	0.000000	0.000000	0.000000	0.000000
R_8	-0.423928	-0.421741	-0.375594	0.002082	-0.000071
R_4	-0.861865	-0.869379	-0.857927	0.011756	0.020615
R_2	-0.540918	-0.540895	-0.522957	0.164929	0.274431
R_{blind}	-0.507466	-0.499797	-0.450272	0.343777	0.330466

Policy completion and decomposition. Sequence weights must be converted into behavior rules before evaluating loss on the expert distribution. For a public prefix h, the probability of action a is the total mixture weight of sequences extending ha divided by that of sequences extending h. When the denominator is zero, the main completion uses an action minimizing A_z, with the same smallest-ID tie rule. This is an observation-cell rule, not a hidden-mode choice. We independently compute $\mathbb{E}_\mu \ell_\pi$ and $\mathbb{E}_{d^\pi} \ell_\pi$ and then obtain $B_R = \mathbb{E}_\mu \ell_\pi - \mathcal{I}_R$ and $C_R = \mathbb{E}_{d^\pi} \ell_\pi - \mathbb{E}_\mu \ell_\pi$. Table 7 changes only off-path completion to uniform actions or to an action maximizing A_z. This leaves executed occupancy, return, and $B_R + C_R$ unchanged, but can reverse the sign of C_R. The risk-minimum and risk-maximum completions bound the allocation induced by those unreachable cells for the fixed on-path behavior; they are not global extrema over all return-optimal policies.

Table 7: Off-path completion of the same closest-TV policy. All columns use return units ($B_R/m, C_R/m$). Changing actions only at unreachable public histories leaves executed occupancy, return, and $B_R + C_R$ unchanged.

Representation	$B_{\min}/m$	$C_{\min}/m$	B_{unif}/m	C_{unif}/m	$B_{\max}/m$	$C_{\max}/m$
R_8	0.0021	-0.0001	0.2670	-0.2650	0.4204	-0.4184
R_4	0.0118	0.0206	0.5715	-0.5392	0.8869	-0.8546
R_2	0.1649	0.2744	0.6625	-0.2232	0.9424	-0.5030
R_{blind}	0.3438	0.3305	0.8931	-0.2188	1.2063	-0.5321

Temporal and terminal contributions. We also sum $\mathcal{I}_R, B_R, C_R$ separately at each macro time. The initial distributions agree, so $C_{R,0} = 0$; Figure 3 shows the later contributions for the main completion. For the blind policy, the global-to-W^- improvement is 0.9649 return units, of which 99.99% comes from zero-gap terminal mass. The further $W^- \to D_R^-$ improvement is 0.3791 and involves nonterminal observation cells. Separating success and failure terminals changes blind TV from 0.01876 to 0.70100, but leaves $W^\pm, D_R^\pm, \mathcal{I}_R, B_R, C_R$, and return unchanged because terminal action losses are zero. Thus terminal encoding strongly affects the global envelope without accounting for the additional policy-class improvement. Table 8 gives independent per-block calculations; the joint analysis optimizes the aggregate occupancy, not an average of blockwise TV or clipped discrepancies.

Numerical validation and remaining slack. All 75 selection LPs passed, with maximum raw primal–dual gap 1.56×10^{-16}. The analysis checks probability conservation, sequence-to-behavior equivalence, the performance identity, $0 \le D_R^\pm \le W^\pm \le G_\pm$, bound validity, and completion and terminal invariances; its maximum recorded residual is 5.23×10^{-13}. An independent implementation reconstructs values and replays policies from the source graphs, with maximum residual 2.34×10^{-15}. It checks all joint policy/completion combinations, all per-block closest-TV completions, and both risk minimizers in every scope. These are finite-instance calculations without Monte Carlo confidence intervals. The manifest, complete behavior rules, occupancies, and per-state contributions are retained in the accompanying analysis artifacts. For a selected return optimum, the remaining lower slack

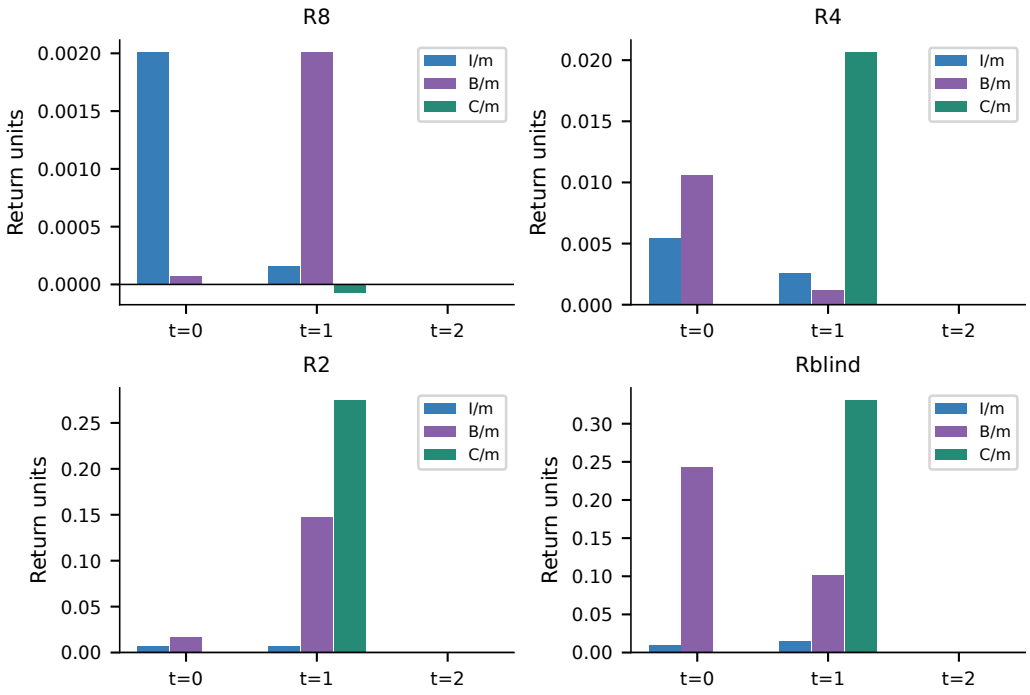

Figure 3: Time-resolved H3 decomposition for the closest-TV return optimum with risk-minimizing off-path completion. All contributions are divided by $m = 1 - \gamma$. The common initialization implies $C_{R,0} = 0$. These contributions condition on the specified complete policy.

Table 8: Each base block is independently normalized and optimized. These are finite-instance values, not independent training-seed estimates; $m = 1 - \gamma$.

Block / representation	$\varepsilon_{\text{info}}$	L_D	B_R/m	C_R/m
8 / R_8	0.0074	-0.4294	0.0037	0.0000
8 / R_4	0.0099	-0.9832	0.0031	0.0000
8 / R_2	0.4387	-0.5519	0.1244	0.3063
8 / R_{blind}	0.6837	-0.4894	0.2450	0.4288
9 / R_8	0.0056	-0.4765	0.0028	-0.0003
9 / R_4	0.0093	-0.9895	0.0040	-0.0100
9 / R_2	0.4993	-0.4773	0.2647	0.1989
9 / R_{blind}	0.7444	-0.6241	0.6510	0.0246
10 / R_8	0.0037	-0.4210	0.0019	0.0000
10 / R_4	0.0099	-0.9538	0.0046	0.0003
10 / R_2	0.3774	-0.6751	0.2069	0.1638
10 / R_{blind}	0.6837	-0.5501	0.3263	0.3475
11 / R_8	0.0000	-0.3713	0.0000	0.0000
11 / R_4	0.1324	-0.5210	0.0353	0.0922
11 / R_2	0.5000	-0.4594	0.0637	0.4288
11 / R_{blind}	0.6837	-0.3663	0.1528	0.5210

equals $[B_R + C_R + D_R^-]/m$. For a selected risk minimizer f, the upper slack is

$$U_D - \varepsilon_{\text{info}} = \frac{D_R^+(d^f, \mu) - C_R(f; \mu)}{m} + J(\pi_R^\star) - J(f).$$

Thus upper looseness can reflect both the discrepancy envelope and the return of the chosen risk minimizer. The $\mathcal{I}_R/(1-\gamma)$ is only 3.52% of the blind deficit, so even a zero lower-side correction could not make this lower bound tight on the current reference state distribution.

C Aligned Prediction–Execution Protocol

Data and representations. The Lift aligned sweep uses an independent trajectory split: 200 demonstrations are permuted with seed 0 into 160 training and 40 held-out trajectories. Four relative-phase bins contribute shared held-out simulator states, giving 30 in total (8, 8, 8, and 6 by phase). The same saved states and demonstration-action targets are used across representations and training seeds 0, 1, and 2. Robot features are end-effector position, quaternion, and gripper position (nine dimensions); the representation adds three cube-position coordinates quantized at $\delta = 0, 0.5, 2, 5, \infty$ cm. The full-state teacher receives the robot features and the full 10-dimensional object vector. Thus the $\delta = 0$ student and the teacher have different input interfaces.

Training and action prediction. The Gaussian-mixture head has three 1,024-unit hidden layers and five mixture components. Policies use Adam with learning rate 10^{-3}, batch size 256, and 100 epochs. Training-input standard deviations are floored at 10^{-2} and normalized inputs clipped to $[-10, 10]$. A separate random forest with 80 trees and minimum leaf size five predicts demonstration actions. Its metric is the mean over action dimensions of held-out RMSE divided by the target standard deviation (with a 10^{-6} numerical stabilizer).

Teacher-continuation loss. Each saved simulator state is restored for evaluation. For $K = 1, 8, 32, 64, 128$, the student acts for up to K primitive steps, then the full-state BC teacher completes the remaining budget. The full rollout uses $H = 200$. Success at primitive time τ has value γ^τ, with $\gamma = 0.99$; failure has value zero. The loss L_K is teacher-only value minus handoff value, averaged over the shared states. Values can be negative when the student improves on that reference.

Association measure. Table 10 reports Spearman correlations across the five representations, calculated separately for each policy seed. The summary is the arithmetic mean of the three seed-specific correlations.

D Policy Improvement and Execution Protocols

This section records supplementary experiments on LIBERO-Object (Liu et al., 2023). Fixed-interface recovery uses DAgger-style data aggregation (Ross et al., 2011); execution comparisons use frozen ResNet-18 (He et al., 2016), R3M (Nair et al., 2023), and DINOv2 ViT-S/14 (Oquab et al., 2024) features.

State-based DAgger. The object-blind t1 student receives only robot-state inputs. Six aggregation rounds collect 12 learner rollouts per round, label them with the full-state BC teacher, and retrain on the augmented dataset. Each round uses 60 training epochs; the teacher uses 100. Evaluation uses 15 initial conditions per seed, five seeds, and a 500-step horizon. Baseline and final checkpoints are compared under the same evaluation protocol. Table 11 reports every seed.

Frozen learned state representation. An autoencoder compresses robot inputs to 16 dimensions and is frozen during policy learning. Its encoder and decoder each use two 1,024-unit hidden layers; training runs for 500 epochs with position reconstruction weighted by four. The BC policy uses a five-component Gaussian-mixture head. Five aggregation rounds collect ten rollouts per round; evaluation uses 12 initial conditions and 300 steps. Tasks t1, t3, and t4 each have seeds 0, 1, and 2. The final checkpoint mean is 0.278 compared with baseline 0.111; selecting the best iteration within each run instead gives 0.435. Two initially unsuccessful task–seed runs remain unsuccessful at the final iteration.

Table 9: Aligned Lift sweep: means across three policy seeds on the same 30 held-out states. ∞ removes cube position.

Quantization (cm)	0	0.5	2	5	∞
E_{act}	0.546	0.546	0.544	0.563	0.554
L_{200}	0.058	0.053	0.026	0.258	0.259

Table 10: Aligned Lift sweep: Spearman correlation across five representations within each seed. Both measurements use the same 30 phase-stratified held-out states. All three seed values and their arithmetic mean are shown.

Autonomous horizon K	1	8	32	64	128	200
Seed 0	-0.10	0.30	0.90	0.70	0.90	0.90
Seed 1	-0.60	0.70	0.60	0.70	0.70	0.70
Seed 2	0.30	-0.60	0.70	0.60	0.70	0.70
Mean	-0.133	0.133	0.733	0.667	0.767	0.767

Visual DAgger. A frozen ResNet-18 supplies the visual inputs on t1. Four aggregation rounds collect eight rollouts per round; each evaluation uses 12 initial conditions and a 300-step budget. The three seed baseline rates are $0.167, 0, 0.0833$; final rates are $0.333, 0.25, 0.417$; and teacher rates are $0.333, 0.667, 0.583$. The corresponding means are 0.0833, 0.333, and 0.528.

Chunked execution. Frozen ResNet-18, R3M, and DINOv2 ViT-S/14 features are paired with single-step or eight-action heads on tasks t1, t3, and t4. The runs use 30 demonstrations and 15 evaluation episodes per task–seed setting. ResNet-18 and DINOv2 each have five seeds per task, and R3M has three. The head uses three 1,024-unit hidden layers, five mixture components, and 120 training epochs. The single-step budget is 200 primitive steps and the chunked budget is 300; Table 13 reports the resulting execution-protocol comparison. Because both the action interface and budget change, this is not a matched-horizon causal estimate of chunking alone or a direct application of Equation (20). DINOv2 improves in 14 pairs and ties in one. A separate state-based comparison keeps chunk length at eight and uses 50 evaluations, five seeds, and 500 steps: full-state versus blind means are 0.896 versus 0.164 on t1, and 0.904 versus 0.892 on t2.

E Additional Representation Diagnostics

E.1 Static Probes and Cross-Distribution Transfer

For a demonstration-side diagnostic D, define $\Delta_{\text{static}} = D(R_{\text{prop}}) - D(R_{\text{full}})$ and $\Delta_{\text{CL}} = \text{SR}_{\text{full}} - \text{SR}_{\text{prop}}$. The earlier benchmark summaries report static separations 0.29, 0.28, and 0.320 for Lift, Can, and CALVIN (Mees et al., 2022), with success gaps 0.62, 0.72, and 0.44, respectively. LIBERO-Object has a reported BC gap near 0.38 and a ten-task state-probe separation mean of 0.0132. These summaries describe their respective experimental settings.

The visual masking analyses use two scales. One normalized-error analysis has three tasks and eight probe seeds per task, with masked-minus-full mean -0.0467. A separate unnormalized-MSE analysis uses three probe seeds per task; t1 has mean difference -0.00889. Each describes the sensitivity of its fitted diagnostic, rather than an estimate of information deficit.

Let $\hat{f}_\mu$ be an action predictor fitted on μ. Its transfer error is

$$e_{\mu\to\nu} = \mathbb{E}_{s\sim\nu}\ell(\hat{f}_\mu(R(s)), a_{\text{ref}}(s)),$$

where the reference target and normalization are held fixed within a comparison. Cross-distribution excess subtracts $e_{d_{\text{exp}}\to d_{\text{exp}}}$ from $e_{d_{\text{exp}}\to d_{\hat{\pi}_R}}$. Table 14 lists the archived task–seed values. Fitting a separate low-error predictor on each distribution is compatible with poor transfer of one common predictor. In a representative object-blind t1 run, the within-demonstration, within-policy, and demonstration-to-policy errors are 0.396, 0.314, and 1.072. Across the four tasks and five seed pairs, object-blind cross-distribution excess exceeds the full-state value in 18 of 20 pairs.

Table 11: Object-blind state DAgger on LIBERO-Object t1. Success rates are based on 15 evaluations per seed and condition.

Seed	BC	Final DAgger	Teacher
0	0.000	0.333	0.600
1	0.000	0.267	0.133
2	0.000	0.133	0.267
3	0.000	0.800	0.667
4	0.000	0.933	0.467
Mean	0.000	0.493	0.427

Table 12: Frozen-latent DAgger: all task–seed outcomes. The primary result uses the final iteration, not the best evaluation.

Task	Seed	BC	Final	Best iteration
t1	0	0.000	0.583	0.750
t1	1	0.000	0.250	0.500
t1	2	0.333	0.417	0.583
t3	0	0.583	0.583	0.583
t3	1	0.000	0.000	0.000
t3	2	0.000	0.333	0.417
t4	0	0.000	0.000	0.083
t4	1	0.000	0.250	0.750
t4	2	0.083	0.083	0.250
Mean		0.111	0.278	0.435

Table 13: Frozen-encoder execution comparisons on three LIBERO-Object tasks. Each pair is one task–seed setting, with 15 evaluation episodes per setting. Single-step and chunked policies have different action interfaces and evaluation budgets (200 versus 300 steps).

Encoder	Pairs	Single step	Chunk 8	Mean gain	Improved pairs
ResNet-18	15	0.142	0.542	0.400	15/15
R3M	9	0.200	0.652	0.452	9/9
DINOv2 ViT-S/14	15	0.129	0.533	0.404	14/15

Table 14: Cross-distribution action-prediction excess by task and seed. Each entry is object-blind / full-state.

Task	Seed 0	Seed 1	Seed 2	Seed 3	Seed 4
t0	.463/.233	.411/.317	.447/.271	1.009/.390	.596/.572
t1	.676/.253	.688/.389	.650/.411	.839/.344	.692/.361
t2	.515/.623	.678/.517	.637/.433	.659/.406	.622/.472
t3	.901/.441	.585/.419	.786/.334	.613/.647	.601/.422