WHEN CAN A WORLD MODEL CONTROL?

Anonymous authors
Paper under double-blind review

ABSTRACT

World models enable agents to simulate and plan, but when do decisions evaluated inside a model remain valid during physical interaction? In this work, we present a normative theoretical framework establishing the conditions under which a world model supports downstream physical control. By formalizing the response map from feedback policies to interaction laws, we show that matching optimal outcome values alone leaves command-consequence correspondence undetermined: a model and a physical system can agree on achievable task scores yet respond differently to the same command. We prove that preserving this policy-specific correspondence tightly bounds downstream planning loss by model substitution error. Through an exact projection onto causal behavior, we decompose a target trajectory's requirements into initial preparation, pre-action information, and post-action physical response. For Gaussian systems, we solve physical realization analytically over all randomized controls, exposing fundamental joint limits on precision, correlation, and sensor timing: a sensor reduces the outcome variance limit only if its reading precedes corrective action. Furthermore, our framework explains a counterintuitive practical failure: why expanding skill libraries can require finer state distinctions and amplify model selection errors. We validate our theoretical predictions across exact finite processes, learned visual models, and continuous contact simulations by systematically varying feedback timing, response coverage, and skill availability. These results connect prediction objectives, latent representations, and execution interfaces directly to the downstream control decisions a world model must support.

1 INTRODUCTION

An agent uses a world model to compare candidate actions before acting in a physical environment (Ha & Schmidhuber, 2018; Schrittwieser et al., 2020; Hansen et al., 2024). Modern learned simulators generate long rollouts and provide synthetic experience for policy improvement (Hafner et al., 2025a;b; Bruce et al., 2024). Feature models support planning (Zhou et al., 2025; Assran et al., 2025), while joint world and action architectures predict future states together with behavior (Ye et al., 2026; Li et al., 2026a). Yet a high predicted task score or high visual fidelity does not by itself show that a selected policy will produce the predicted consequences during execution (Zhang et al., 2026; Yang et al., 2026). Knowing every attainable outcome value also leaves the policy that achieves that outcome unspecified. During physical interaction, actions alter the observations on which subsequent decisions depend (Kaelbling et al., 1998; Klenske & Hennig, 2016). A model that supports control must preserve the relations connecting these decisions, information arrivals, and causal consequences. We examine the normative conditions imposed by these physical requirements before selecting a representation or a training algorithm (Farahmand et al., 2017; Grimm et al., 2020; Subramanian et al., 2022).

We consider a partially observed controlled process where hidden physical properties determine the transition response (Åström, 1965; Smallwood & Sondik, 1973). A world model provides a predictive substitute for this process. Our basic object is the response map, which assigns each causal feedback rule the probability distribution of the resulting complete interaction transcript (Yüksel & Saldi, 2017; Saldi & Yüksel, 2022). This response map contains both the interactions available to the agent and the decision rules that produce them. Crucially, capability and policy correspondence convey distinct information. Two systems can agree on the optimal achievable value for every outcome reward while assigning entirely different consequences to the same command. Reversing an actuator polarity preserves the convex hull of attainable outcome distributions while inverting the physical meaning

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.

of each input. To evaluate a policy selected inside a model, we compare its consequences in both systems under the same feedback rule, responding to each system's own observations. We prove that this model substitution discrepancy, combined with the shortfall in model evaluated score, tightly bounds downstream regret relative to the optimal physical policy in the declared class by a sharp factor of two.

Exact agreement with physical response laws on tested paths does not imply that every requested future is attainable. A requested trajectory may demand favorable initial conditions, require information before that information arrives, or rely on chance events after actuation. To determine which goals can be realized, we project an arbitrary full trajectory law onto the set of legal causal policies by relative entropy (Csiszár, 1975; Ziebart et al., 2010). This projection splits the target divergence into three chronological components: starting preparation, information required before acting, and physical response produced after acting. Conditioning demonstrations on successful outcomes can select initial states and post action disturbances simultaneously (Brandfonbrener et al., 2022; Yang et al., 2023b). Even with unlimited demonstration data, the resulting distribution can remain separated from every executable policy. When a state summary is sufficient for controlled responses (Littman et al., 2002; Singh et al., 2004; Subramanian et al., 2022), physical realization obeys an exact local criterion: the requested initial law must match the environment, and each requested next state conditional distribution must belong to the convex hull formed by the mixtures of available action responses at that history.

These realization constraints govern entire distributions rather than isolated summary metrics. For continuous Gaussian targets under additive disturbance after action, we solve the relative entropy projection analytically over all randomized controls in arbitrary dimension. Unrestricted mean actuation coexists with an insurmountable precision limit determined by disturbance entering after the control decision. Matching marginal variances along individual coordinates fails to guarantee feasibility, because target correlations can require variance along a principal direction below the physical noise floor. Furthermore, sensor timing alters this limit only through its relation to subsequent intervention (Blackwell, 1953; Todorov & Jordan, 2002). An informative sensor reading can reduce the outcome variance floor when it arrives before corrective action, whereas arriving after the final opportunity for correction leaves the physical dispersion unchanged.

This normative perspective explains an unexpected practical failure in embodied planning: expanding an agent's skill library (Sutton et al., 1999; Barreto et al., 2017). While adding skills expands the set of physically available responses, it simultaneously introduces new comparisons for a learned predictor.

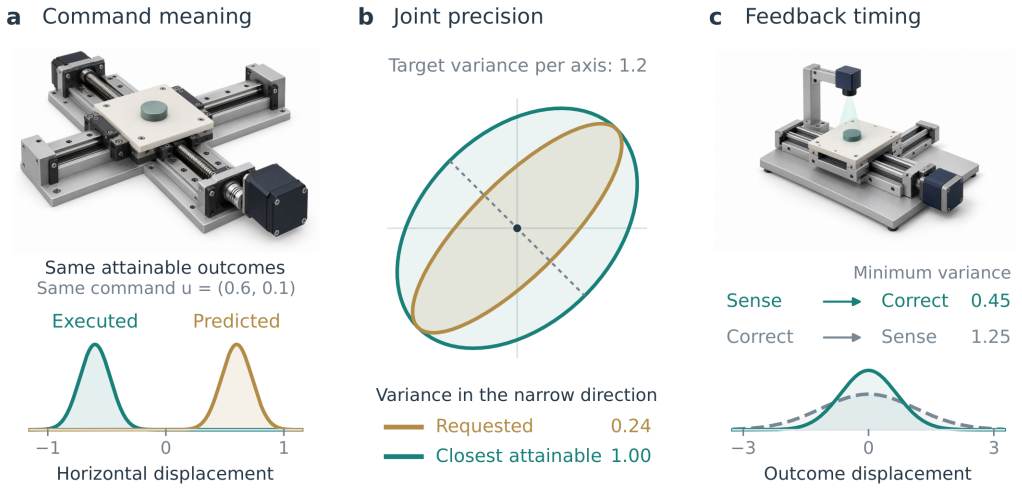


Figure 1: **Physical conditions for control.** a, Actuator reversal preserves attainable laws while changing a common command's consequences. b, The target has feasible coordinate variances and excessive joint precision under unit noise. c, Reading before correction lowers the variance floor; initial, sensor and later noise variances are 1, 0.25 and 0.25. Distributions are exact; conceptual apparatus illustrations were created with ImageGen. Appendix A.14 gives the constructions.

Policy selection depends on predicted advantages across candidate options (Smith & Winkler, 2006; van Hasselt, 2010). When relative prediction errors between newly introduced skills and incumbent skills exceed the predicted improvement, the selector switches to an inferior action. Expanding available skills can therefore degrade physical performance under a fixed model. Deployment can use paired physical evaluations of candidate options, error bounds across their comparisons, and latent state distinctions that preserve the relevant control continuations (Givan et al., 2003; Ferns et al., 2004; Subramanian et al., 2022).

We validate these theoretical predictions across exact finite processes, learned visual models, and continuous contact simulations. Finite processes provide exact discrepancies and confirm the sharpness of our planning bounds. Learned visual models demonstrate that large visual prediction errors from background distractors can coexist with zero change in downstream planning regret, whereas perturbing predicted contact responses can degrade control while leaving pixel predictions unchanged. In continuous frictional contact simulations, we isolate feedback timing by removing prediction error, demonstrate that targeted policy probes improve deployment over routine data collection, and evaluate skill expansion across 44,928 planning decisions. These findings connect prediction objectives, latent abstractions, and execution interfaces directly to the downstream control decisions a world model must support. Appendix D discusses connections to statistical decision theory, model equivalence, causal transport, and contemporary world-model systems; Appendix E develops a computational interpretation in biological cognition.

In summary, this work makes the following contributions:

1. **Normative Response Map Framework and Tight Substitution Bounds:** We formalize the response map from feedback policies to complete interaction laws and prove that model substitution error tightly bounds downstream physical regret with a sharp factor of two across all bounded task objectives, demonstrating that matching optimal outcome values alone leaves command correspondence undetermined.
2. **Causal Path Projection and Local Realization Criterion:** We prove the legal policy projection theorem, decomposing target divergence into initial preparation, pre action information, and post action physical response, and establish a local convex hull criterion showing that a target law is attainable if and only if its initial law is physical and each conditional transition belongs to the mixture of available action responses.
3. **Analytic Continuous Limits and Sensor Timing Mechanics:** For multidimensional Gaussian systems under additive noise, we derive the exact relative entropy projection over all randomized control laws, characterizing precision limits, correlation constraints, and the exact condition under which sensor timing reduces outcome variance.
4. **Mechanics of Skill Expansion and Abstraction:** We identify the error mechanism through which expanding skill libraries can degrade model selected performance, and establish criteria linking latent state distinctions to admissible control continuations.
5. **Multidomain Empirical Validation Across Systems:** We test the framework through exact finite processes, learned visual models, analytic Gaussian calculations, and continuous contact simulations, including 44,928 planning decisions, demonstrating the operational link between predictive models and downstream control.

2 PRESERVING THE CONSEQUENCES OF A DECISION

Executing a policy closes a feedback loop: actions change observations, which inform subsequent actions. A model supports that policy when it preserves the resulting interaction law.

We consider a partially observable Markov decision process (POMDP) (Åström, 1965; Kaelbling et al., 1998) with hidden state X_t, observation O_t, and action A_t. The available history is $H_t = (O_0, A_0, \ldots, A_{t-1}, O_t)$. A *causal policy* $\pi_t(a \mid H_t)$ selects actions using this history and private randomness independent of the world. Planning, feedback control, and adaptation during execution define such policies.

Model M and environment E share an action and observation interface. Running π for T steps in system S produces a law $\mathcal{R}_S^T(\pi)$ on $C_T = (O_0, A_0, \ldots, A_{T-1}, O_T)$. The *response map* $\pi \mapsto \mathcal{R}_S^T(\pi)$ retains the policy associated with each induced path law (Yüksel & Saldi, 2017). The same feedback

rule responds to each system's own observations, so its realized actions may differ. A skill interface includes returned information and duration (Appendix A.2).

For a declared policy class Π, two quantities distinguish model accuracy from physical attainability:

$$\Delta = \sup_{\pi\in\Pi} \mathrm{TV}\big(\mathcal{R}_M^T(\pi), \mathcal{R}_E^T(\pi)\big), \qquad \chi_E(q) = \inf_{\pi\in\Pi} \mathrm{TV}\big(q, \mathcal{R}_E^T(\pi)\big). \tag{1}$$

Total variation (TV) is the largest difference in event probability. Thus Δ measures model substitution error, and $\chi_E(q)$ is the smallest physical discrepancy from a requested law q. A specified projection onto task outcomes can serve as the evaluation record; the policy retains its declared observations.

Theorem 1 (Model substitution and physical performance)**.** *The discrepancy Δ equals the largest error in expected reward over policies in Π and rewards on the evaluation record taking values in $[0,1]$. A policy whose model reward is within η of the model optimum has physical regret at most $2\Delta+\eta$, relative to the best physical policy in Π. If π_M instead minimizes distance to q in the model within η, then*

$$\chi_E(q) \;\le\; \mathrm{TV}\big(q, \mathcal{R}_E^T(\pi_M)\big) \;\le\; \chi_E(q) + 2\Delta + \eta. \tag{2}$$

The model can overvalue the selected policy and undervalue the physical optimum. Both errors affect their ranking, and the factor two is sharp (Appendix A.3). This simulation-lemma argument (Kearns & Singh, 2002; Lobel & Parr, 2024) applies to complete interaction laws. Restricted task classes give corresponding discrepancies, as in value equivalence and approximate information states (Grimm et al., 2020; 2021; Subramanian et al., 2022).

On a finite outcome space, the optimal values for all rewards determine the closed convex hull of attainable laws. With all randomized causal policies, this is the attainable set itself. Actuator reversal preserves that set while exchanging the commands reaching its outcomes (Figure 1a). The response map retains this correspondence. For the same randomized class, an unattainable target has a bounded task whose requested value exceeds the physical optimum; the largest shortfall equals $\chi_E(q)$ (Proposition 8).

3 What a requested trajectory asks of the controller

A requested trajectory can remain unattainable even with an exact model. A successful grasp may involve favorable initial conditions, an action informed by hidden weight, or a fortunate disturbance after contact. Reproducing successful trials requires the relevant dependencies at each stage.

Let Q be a probability law on complete requested trajectories, including hidden states and actions. Write $\xi_t = (X_t, O_t)$ and $\bar{H}_t = (X_{0:t}, H_t)$. A physical law factors in temporal order:

$$P_E^\pi(\omega) = P_0(\xi_0) \prod_{t<T} \pi_t(a_t \mid H_t) K_t(\xi_{t+1} \mid \bar{H}_t, a_t). \tag{3}$$

The environment supplies P_0 and K_t; the controller chooses the action factors using available history. Factor Q in the same order, with initial law Q_0 and response conditionals L_t. Relative entropy $D_{\mathrm{KL}}(Q\|P)$ permits a chronological decomposition, measured in nats. It is infinite when Q requests a physically excluded event.

Theorem 2 (Projection onto legal behavior)**.** *For finite spaces and all randomized causal policies,*

$$\begin{aligned}\inf_\pi D_{\mathrm{KL}}(Q\|P_E^\pi) = \underbrace{D_{\mathrm{KL}}(Q_0\|P_0)}_{\textit{initial preparation}} &+ \underbrace{\sum_{t<T} I_Q(A_t; X_{0:t} \mid H_t)}_{\textit{information required before action}} \\ &+ \underbrace{\sum_{t<T} \mathbb{E}_Q D_{\mathrm{KL}}\big(L_t(\cdot \mid \bar{H}_t, A_t)\|K_t(\cdot \mid \bar{H}_t, A_t)\big)}_{\textit{physical response required after action}}.\end{aligned} \tag{4}$$

When this value is finite, the policy $\pi_t^Q(a \mid H_t) = Q(a \mid H_t)$ attains it.

The first term measures altered starting conditions. The second, conditional mutual information, measures dependence of requested actions on hidden history beyond available observations. The third

measures altered physical responses. The entropy chain rule and projection of action conditionals give this identity (Csiszár, 1975; Ziebart et al., 2010); Appendix A.4 provides the proof.

The terms locate where an intervention could change feasibility: preparation, sensing, or actuation. They sum for a fixed target and process; changing preparation or sensing can also alter later responses.

The minimizing policy reproduces the target's action frequencies conditional on available history. Execution still inherits physical initial conditions and disturbances. Conditioning demonstrations on success can select both, leaving a positive realization cost even with unlimited data. The decomposition identifies these selected dependencies in stochastic return conditioning (Brandfonbrener et al., 2022; Yang et al., 2023b); Appendix A.5 gives the associated information identity.

A goal specified through observations or outcomes permits several hidden trajectories. Its projection minimizes over all hidden completions with the prescribed observable law (Proposition 4). The next result expresses feasibility directly through observable responses.

4 Physical limits on outcome distributions

Let Z_t summarize observed history sufficiently to predict the next response under every action. Complete history qualifies; belief or predictive states can provide smaller descriptions. At $Z_t = z$, action a produces law $K_t^a(\cdot \mid z)$. Randomizing actions mixes their response laws, giving

$$\mathcal{K}_t(z) = \mathrm{conv}\{K_t^a(\cdot \mid z) : a \text{ is available at } z\}, \tag{5}$$

where conv consists of mixtures with nonnegative weights summing to one. A grasp with a chance of slipping contributes its full distribution of responses to this set.

Theorem 3 (Local criterion for realization)**.** *On finite spaces, suppose Z_t is sufficient for the controlled next-state response under every legal policy, and available actions are determined by Z_t. A target law on $Z_{0:T}$ is attainable by a randomized causal policy exactly when its initial law is physical and each next-state conditional, given a target history of positive probability, belongs to $\mathcal{K}_t(Z_t)$.*

A realizing policy uses the required mixture weights at each history. Minimizing conditional relative entropy over these sets gives the closest attainable law (Appendix A.6). Time and resources can enter the state. The same membership test determines when conditioned trajectories are executable, encompassing the physical assumptions of path-integral and linearly solvable control (Kappen, 2005; Todorov, 2006; Levine, 2018); see Appendix A.7.

Joint precision under additive noise. Consider an outcome $Y = U + W$. The agent chooses a possibly randomized displacement U, then an independent centered Gaussian disturbance W with positive definite covariance is added. Actuation has no amplitude limit. For a nonsingular Gaussian target q, express its covariance in units where the disturbance has identity covariance. Let λ_i be the resulting principal variances. In any dimension,

$$\inf_{\mathrm{Law}(U)} D_{\mathrm{KL}}(q \| \,\mathrm{Law}(U+W)) = \frac{1}{2} \sum_{i:\lambda_i<1} (\lambda_i - 1 - \log \lambda_i). \tag{6}$$

The closest response attains the requested mean and raises each principal variance below one to one. Proposition 10 proves optimality over all randomized controls, including non-Gaussian distributions. Arbitrary mean displacement thus coexists with a precise limit on reliability: control can add variation, while disturbance arriving afterward supplies a residual noise floor.

This is a joint constraint. Each coordinate can satisfy its marginal variance limit while correlation requires an infeasibly narrow direction (Figure 1b). Partial actuation also constrains dependence between controlled and uncontrolled coordinates (Corollary 11).

The timing of information. Suppose an initial Gaussian displacement V has a jointly Gaussian sensor reading S. If S arrives before an unrestricted correction, the controller can cancel $\mathbb{E}[V \mid S]$. With independent noise W added after correction, the remaining covariance is $\mathrm{Cov}(V \mid S) + \mathrm{Cov}(W)$. If the reading arrives after the final action, the floor is $\mathrm{Cov}(V) + \mathrm{Cov}(W)$ (Appendix A.13.2). The same sensor therefore changes achievable precision through the action opportunity that follows its reading (Figure 1c).

5 CONTROL DETERMINES THE REQUIRED PREDICTIONS

Adding skills can reduce the physical target error $\chi_E(q)$ while increasing substitution error Δ: new policies reach responses that earlier policies never used. Capability and the demands placed on its predictor grow together.

Which distinctions a state must retain. Suppose a representation merges two histories. If a common continuation produces future laws differing by TV δ, a predictor initialized only from that representation incurs error at least $\delta/2$ on one history (Appendix A.9). Mass matters when it changes a grip's response; texture can be discarded when it changes no relevant continuation. This connects response preservation to predictive states, bisimulation, and value equivalence (Littman et al., 2002; Ferns et al., 2004; Grimm et al., 2020; Subramanian et al., 2022).

A new tool can reveal a difference that earlier skills left irrelevant. Its use then requires recovering that difference from history or interaction. This supplies a criterion for information retained by feature models (Zhou et al., 2025; Assran et al., 2025).

Skills also convey information. An exploratory contact both moves an object and reveals evidence about friction (Klenske & Hennig, 2016). Identical visible endpoints can leave different information or time for correction. Composition requires a sufficient returned state and relevant duration (Sutton et al., 1999); interruption opportunities also belong to an action chunk's interface (Black et al., 2025).

Which prediction errors change a selection. For an incumbent skill a and a newly selected skill b, let c denote physical cost, $\hat{c}$ predicted cost, and $e = c - \hat{c}$ signed error. Their comparison obeys

$$\underbrace{c(a) - c(b)}_{\text{realized gain}} = \underbrace{\hat{c}(a) - \hat{c}(b)}_{\text{predicted gain}} - \underbrace{[e(b) - e(a)]}_{\text{error in the comparison}} . \tag{7}$$

The expansion worsens performance when the error in this comparison exceeds the predicted improvement. A bias shared by the candidates cancels. More generally, the regret of a predicted minimizer is bounded by the range of signed errors across its skill bank (Proposition 9). Paired executions in matched physical contexts measure these relative errors. Bounds valid simultaneously across candidate comparisons support accepting an option when its predicted advantage exceeds comparison uncertainty.

The required data follow from these decisions. A branch absent from training but reached with probability α at deployment leaves worst-case ambiguity of at least $\alpha/2$ (Appendix A.8). Probing consequential comparisons can therefore improve planning even when routine prediction scores change little. Appendices D.6 and F develop consequences for training, skill interfaces, and transfer across bodies.

6 EXPERIMENTS

The experiments follow the decisions required to use a world model for control. First, exact finite processes and learned visual models test whether a prediction preserves the consequences of the policy evaluated inside it. We then study one contact task through three interventions: when a target cue can change an action, which physical responses enter training, and which skills the planner may choose. Together these comparisons connect the theory's requirements to observable failures and to specific remedies. Appendix C.1 introduces every study; the linked methods give complete protocols.

6.1 PREDICTIONS AND THE POLICIES THEY SUPPORT

Finite processes allow us to calculate complete interaction laws. Reversing an actuator preserves attainable goal values while changing each command's consequences; translating the commands restores the response map (A). Across 180 model–environment pairs and twelve bounded objectives per pair, all 2,160 planning comparisons satisfy the transfer bound (B). In a preparation, sensing, and gripping task, reset, scanner, and tool interventions remove the corresponding terms of the trajectory projection cost (C). Success-conditioned demonstrations retain a positive realization cost even when their action conditionals are known exactly (D). Further studies vary which responses training logs identify (E) and which outcome distributions a feedback skill makes available (G).

We next train 36 visual models on eight-step grid navigation and pushing, varying architecture, data coverage, dataset size, and initialization (F). Separate image and operational prediction heads allow two paired interventions. Changing a background border that is masked before encoding increases pixel MSE by 0.0980 on average and leaves every tested deployment decision unchanged. Perturbing predicted contact responses preserves pixel predictions and increases mean regret by 0.310 on a unit-normalized utility scale. The distinction is computational: the planner selects actions through the operational predictions. These experiments identify how an error enters that decision (Appendix C.7).

6.2 When information can change a push

A circular pusher moves a rectangle toward one of two targets. A cue identifies the target during motion. Contact couples translation and rotation; contexts vary shape, mass, friction, pose, and actuator response. Commands run at 40 Hz. At each allowed update, an exact dynamics copy generates a sequence using the physical state and current cue. The pusher executes that sequence until the next update. Varying this commitment interval changes when new information can affect motion while preserving exact prediction.

We cross six commitment lengths with eleven cue times in 192 contexts, each paired with both targets (Figure 2). With the cue present initially, all commitment lengths yield the same trajectory and 73.7% success. At 0.875 s, one-step feedback gives 82.6% success, four-step chunks 85.7%, sixteen-step chunks 70.1%, and full-episode commitment zero. Conditions with the same first permitted update after the cue have identical success. That update occurs from different contact configurations, so performance under this fixed controller need not vary monotonically with commitment length.

Shuffling cues preserves their frequency and arrival time while reducing feedback success to 40.9%, a paired decrease of 41.7 percentage points (95% interval 39.1–44.3). Endpoint TV changes from 0.174 to 0.182, while joint target–endpoint TV rises from 0.174 to 0.591. Reaching the two target regions in the right proportions leaves the correspondence between each request and its outcome unspecified. Both interventions expose the same requirement: a cue affects control through the actions that can still use it (Appendix B.2).

6.3 Learning the responses compared during planning

The next study keeps feedback available and learns what its use will accomplish. The planner chooses among 27 two-second feedback skills that vary pushing distance, steering, and speed. A network predicts their final displacement, rotation, and peak force from public geometry, motion, calibration history, and skill arguments. Mass, friction, and actuator parameters remain hidden. Starting from the

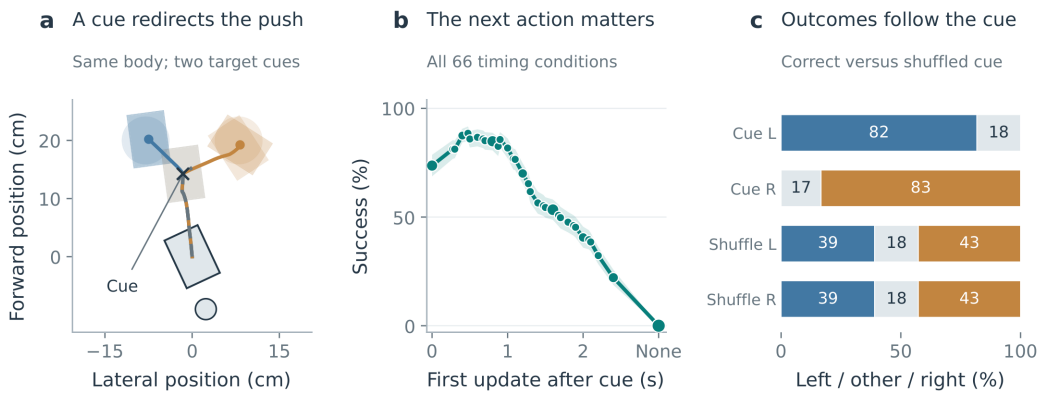


Figure 2: **Information helps while correction remains possible.** a, Recorded trajectories in context zero; the cue arrives at 0.875 s and target radii are 4 cm. b, All 66 timing conditions indexed by the first permitted update after the cue; marker area counts coincident conditions. Each uses 192 contexts and both targets. c, Endpoint fractions conditional on the requested target. Shuffled cues retain similar endpoint marginals and disrupt target correspondence. Prediction is exact throughout.

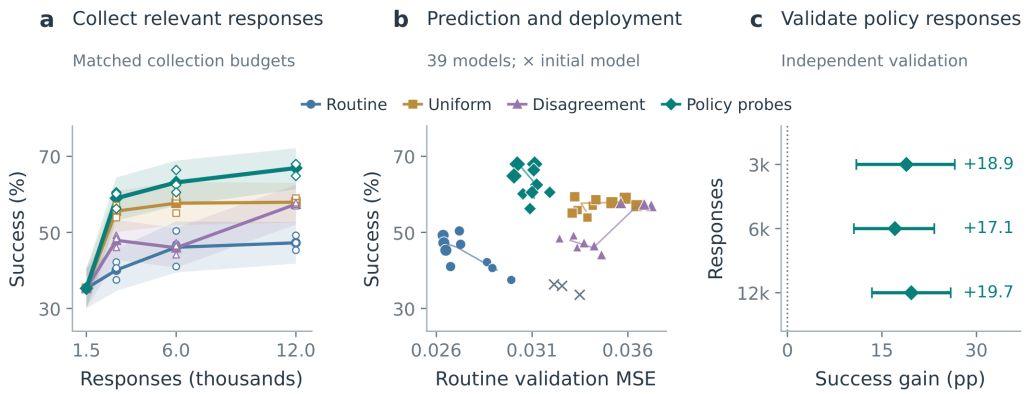


Figure 3: **Response coverage changes which model supports control.** a, Success versus response budget; small marks show three initializations. b, Routine validation MSE and success for all 39 models; marker area denotes budget. c, Held-out success gained by selecting models with independent deployment-probe validation. Bands and bars are paired 95% bootstrap intervals over 256 contexts and three initializations sharing each dataset.

same 1,500 responses, we acquire further labels through routine pushes, uniform sampling, initial ensemble disagreement, or probes covering deployment skills. Architecture and optimizer updates are fixed. Three initializations per condition give 39 models, evaluated on 256 new contexts and 128 contexts with shifted body parameters.

At 6,000 labels, deployment probes yield 63.2% success, routine collection 46.1%, and disagreement 46.0%. The paired gain over routine collection is 17.1 percentage points (95% interval 10.5–23.3). Doubling routine labels gives 47.3%. Routine validation nevertheless favors routine collection: MSE is 0.0271, compared with 0.0311 for probes. Independent validation on deployment responses selects models with better held-out control at every budget (Figure 3). The relevant data cover the responses between which the planner chooses. Under the body shift, uniform collection yields 59.9%, probes 58.9%, and routine collection 42.2%, showing how the useful coverage also depends on the deployment population. Appendix B.3 gives the collection distributions, objectives, models, and integration refinements.

6.4 Adding skills changes the required comparisons

We now hold the learned models and physical responses fixed and vary the available choices. The starting bank offers three speeds for a direct push at the requested distance. Each of the other 24 skills is added separately; the bank also expands in 128 random orders. Every change in selected performance therefore comes from a new comparison.

Position cost is final target distance divided by 30 cm, capped at one. From three to 27 skills, the mean minimum measured cost across the available skills falls from 0.0653 to 0.0269. The routine model's selected cost rises from 0.1339 to 0.1804; the probe model's rises from 0.0944 to 0.1280 (Figure 4). For routine collection, 22 of 24 individual additions increase mean realized cost despite nonnegative predicted gains; for probes, 19 do. The relative error in Equation 7 explains each reversal: an overly favorable comparison displaces a better incumbent.

The initial options determine how much expansion can help. Averaging over all 2,925 three-skill banks, expansion improves mean position cost for every collection rule on original bodies. A useful starting bank makes errors that displace its incumbent consequential. The full evaluation comprises 44,928 decisions: 39 models, 384 contexts, and three objectives, reusing 10,368 physical skill executions. All satisfy the finite-choice error bound. Appendix B.4 reports every addition, starting bank, objective, and body regime.

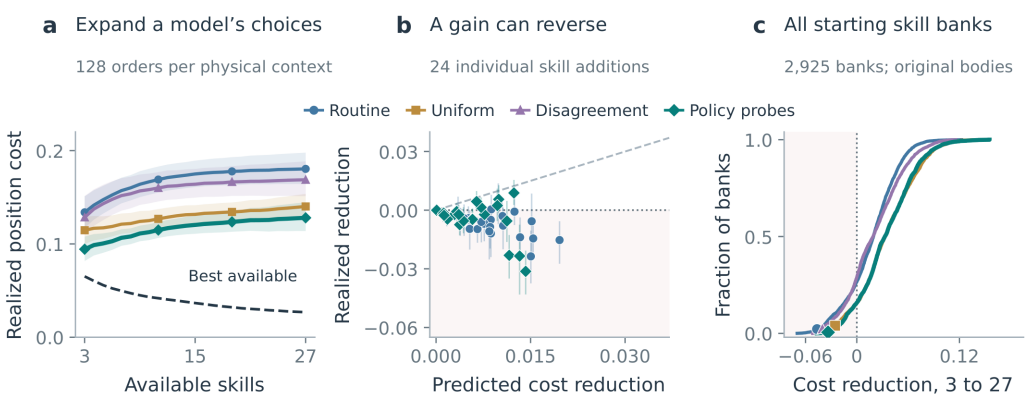


Figure 4: **New skills enlarge capability and change selection.** a, Position cost at 6,000 labels, averaging 128 expansion orders within context. b, Predicted and realized gains for all 24 additions. Bands and bars are paired 95% intervals. c, Effects across all 2,925 starting banks, each averaged over 256 contexts and three initializations. Symbols mark the bank in a–b. The best measured response in each context provides the capability reference.

7 Discussion

A world model supports control by preserving the consequences of the agent's feedback rules. The response map retains this correspondence; the attainable set describes what the agent can bring about. Together, they connect prediction to physical possibility. Models with different internal coordinates can support the same decisions when they preserve these relations. Their adequacy is defined by the interventions and tasks they are asked to support.

Projection of a requested trajectory identifies where it exceeds those possibilities. Success conditioning can select starting states, hidden information used to choose actions, and disturbances that follow them. Reproducing its action frequencies alone can leave the selected circumstances unattainable. The three projection terms locate the relevant changes in preparation, sensing, and actuation. Because these contributions follow from chronological factorization, they connect a statistical mismatch to the point in an interaction where a physical intervention can address it.

The Gaussian result makes this order visible in precision. Control can attain a requested mean while retaining variation that arrives after the final correction. Target correlations can demand more precision in a joint direction than the apparatus permits. A sensor lowers this floor through a later action. Biological feedback depends on the relevance of disturbances to the task (Todorov & Jordan, 2002; Franklin & Wolpert, 2008; Pruszynski et al., 2014). Moving the same informative cue across the last correction opportunity, then restoring a later correction, would test whether motor dispersion tracks the opportunity to use information while perceptual reports track the cue itself.

Available continuations also determine which distinctions a representation should retain. A new tool can make a previous tactile observation relevant by introducing a response to the property it revealed. A feedback skill can use that distinction internally and return a simpler state to the planner (Littman et al., 2002; Sutton et al., 1999; Subramanian et al., 2022). The information remains within the executing controller. Orbitofrontal representations of hidden state and prospective hippocampal activity provide settings for testing these predictions (Schuck et al., 2016; Pfeiffer & Foster, 2013; Mattar & Daw, 2018). Matched experience followed by different action opportunities would reveal whether remembered distinctions track their use in future decisions.

For learning, expanded capability changes the responses that data must resolve. New skills introduce comparisons whose relative prediction errors can determine the selected action. The finite, visual, and contact studies isolate these relations by changing response fidelity, feedback timing, and available choices. Learning the required comparisons online, with finite data and changing interfaces, would extend the normative conditions to adaptive world models.

REPRODUCIBILITY STATEMENT

Appendix A contains the assumptions, proofs, counterexamples, and numerical identities. Appendices B and C specify every experiment, including data collection, training, paired comparisons, and simulation resolution. The accompanying package contains all 75 trained neural checkpoints, raw result tables, cached physical responses, analysis and plotting code, editable figures, and build instructions. The skill expansion analyses reuse saved predictions and physical executions; their scripts retain every model and context. The numerical and visual exports can be regenerated independently of neural training.

STATEMENT ON THE USE OF AI TOOLS

AI tools were used to assist with writing and language polishing, retrieve information and discover related literature, and execute code. The authors take responsibility for the final manuscript and accompanying materials.

REFERENCES

Karl J. Åström. Optimal control of Markov processes with incomplete state information. *Journal of Mathematical Analysis and Applications*, 10:174–205, 1965. doi: 10.1016/0022-247X(65)90154-X.

Anurag Ajay, Yilun Du, Abhi Gupta, Joshua Tenenbaum, Tommi Jaakkola, and Pulkit Agrawal. Is conditional generative modeling all you need for decision-making? In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2211.15657.

Eloi Alonso, Adam Jelley, Vincent Micheli, Anssi Kanervisto, Amos Storkey, Tim Pearce, and François Fleuret. Diffusion for world modeling: Visual details matter in Atari. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2405.12399.

Dilip Arumugam and Benjamin Van Roy. Deciding what to model: Value-equivalent sampling for reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/3b18d368150474ac6fc9bb665d3eb3da-Abstract-Conference.html.

Mido Assran, Adrien Bardes, David Fan, et al. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025. URL https://arxiv.org/abs/2506.09985.

Julio Backhoff, Mathias Beiglböck, Yiqing Lin, and Anastasiia Zalashko. Causal transport in discrete time and applications. *SIAM Journal on Optimization*, 27(4):2528–2562, 2017. doi: 10.1137/16M1080197. URL https://arxiv.org/abs/1606.04062.

Julio Backhoff-Veraguas, Daniel Bartl, Mathias Beiglböck, and Manu Eder. Adapted Wasserstein distances and stability in mathematical finance. *Finance and Stochastics*, 24:601–632, 2020. doi: 10.1007/s00780-020-00426-3. URL https://arxiv.org/abs/1901.07450.

André Barreto, Will Dabney, Rémi Munos, Jonathan J. Hunt, Tom Schaul, Hado van Hasselt, and David Silver. Successor features for transfer in reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/350db081a661525235354dd3e19b8c05-Abstract.html.

Silvia Bernardi, Marcus K. Benna, Mattia Rigotti, Jérôme Munuera, Stefano Fusi, and C. Daniel Salzman. The geometry of abstraction in the hippocampus and prefrontal cortex. *Cell*, 183(4):954–967.e21, 2020. doi: 10.1016/j.cell.2020.09.031.

Dimitri P. Bertsekas and Steven E. Shreve. *Stochastic Optimal Control: The Discrete-Time Case*. Academic Press, 1978.

Kevin Black, Manuel Y. Galliker, and Sergey Levine. Real-time execution of action chunking flow policies. *arXiv preprint arXiv:2506.07339*, 2025. URL https://arxiv.org/abs/2506.07339.

David Blackwell. Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, 24(2):265–272, 1953. doi: 10.1214/aoms/1177729032.

David Brandfonbrener, Alberto Bietti, Jacob Buckman, Romain Laroche, and Joan Bruna. When does return-conditioned supervised learning work for offline reinforcement learning? In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2206.01079.

Ethan S. Bromberg-Martin and Okihide Hikosaka. Midbrain dopamine neurons signal preference for advance information about upcoming rewards. *Neuron*, 63(1):119–126, 2009. doi: 10.1016/j.neuron.2009.06.009.

Jake Bruce, Michael Dennis, Ashley Edwards, et al. Genie: Generative interactive environments. In *International Conference on Machine Learning*, 2024. URL https://arxiv.org/abs/2402.15391.

Lars Buesing, Johannes Bill, Bernhard Nessler, and Wolfgang Maass. Neural dynamics as sampling: A model for stochastic computation in recurrent networks of spiking neurons. *PLOS Computational Biology*, 7(11):e1002211, 2011. doi: 10.1371/journal.pcbi.1002211.

Lars Buesing, Theophane Weber, Yori Zwols, Sébastien Racanière, Arthur Guez, Jean-Baptiste Lespiau, and Nicolas Heess. Woulda, coulda, shoulda: Counterfactually-guided policy search. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=BJG0voC9YQ.

Wenxi Chen, Kaidi Zhang, Chi Lin, Zhiyuan Zhang, Yu She, Yuejiang Liu, Raymond A. Yeh, Shaoshuai Mou, and Yan Gu. DREAM-Chunk: Reactive action chunking with latent world model. *arXiv preprint arXiv:2606.18589*, 2026. URL https://arxiv.org/abs/2606.18589.

Yongxin Chen, Tryphon T. Georgiou, and Michele Pavon. Optimal steering of a linear stochastic system to a final probability distribution, part I. *IEEE Transactions on Automatic Control*, 61(5): 1158–1169, 2016. doi: 10.1109/TAC.2015.2457784. URL https://arxiv.org/abs/1408.2222.

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *arXiv preprint arXiv:2303.04137*, 2023. URL https://arxiv.org/abs/2303.04137.

Paul Cisek. Cortical mechanisms of action selection: The affordance competition hypothesis. *Philosophical Transactions of the Royal Society B*, 362:1585–1599, 2007. doi: 10.1098/rstb.2007.2054.

Imre Csiszár. I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, 3(1):146–158, 1975. doi: 10.1214/aop/1176996454.

Pim de Haan, Dinesh Jayaraman, and Sergey Levine. Causal confusion in imitation learning. In *Advances in Neural Information Processing Systems*, 2019. URL https://arxiv.org/abs/1905.11979.

Amir-Massoud Farahmand, André Barreto, and Daniel Nikovski. Value-aware loss function for model-based reinforcement learning. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1486–1494, 2017. URL https://proceedings.mlr.press/v54/farahmand17a.html.

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

David W. Franklin and Daniel M. Wolpert. Specificity of reflex adaptation for task-relevant variability. *Journal of Neuroscience*, 28(52):14165–14175, 2008. doi: 10.1523/JNEUROSCI.4406-08.2008.

Karl J. Friston, Jean Daunizeau, James Kilner, and Stefan J. Kiebel. Action and behavior: A free-energy formulation. *Biological Cybernetics*, 102(3):227–260, 2010. doi: 10.1007/s00422-010-0364-z.

Shenyuan Gao, William Liang, Kaiyuan Zheng, et al. DreamDojo: A generalist robot world model from large-scale human videos. *arXiv preprint arXiv:2602.06949*, 2026. URL https://arxiv.org/abs/2602.06949.

James J. Gibson. *The Ecological Approach to Visual Perception*. Houghton Mifflin, 1979.

Robert Givan, Thomas Dean, and Matthew Greig. Equivalence notions and model minimization in Markov decision processes. *Artificial Intelligence*, 147(1–2):163–223, 2003. doi: 10.1016/S0004-3702(02)00376-4.

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, 2020. URL https://arxiv.org/abs/2011.03506.

Christopher Grimm, André Barreto, Gregory Farquhar, David Silver, and Satinder Singh. Proper value equivalence. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/400e5e6a7ce0c754f281525fae75a873-Abstract.html.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018. URL https://arxiv.org/abs/1803.10122.

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

Danijar Hafner, Wilson Yan, and Timothy Lillicrap. Training agents inside of scalable world models. *arXiv preprint arXiv:2509.24527*, 2025b. URL https://arxiv.org/abs/2509.24527.

Nicklas Hansen, Hao Su, and Xiaolong Wang. TD-MPC2: Scalable, robust world models for continuous control. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.16828.

Christopher M. Harris and Daniel M. Wolpert. Signal-dependent noise determines motor planning. *Nature*, 394:780–784, 1998. doi: 10.1038/29528.

Jun Izawa and Reza Shadmehr. Learning from sensory and reward prediction errors during motor adaptation. *PLOS Computational Biology*, 7(3):e1002012, 2011. doi: 10.1371/journal.pcbi.1002012.

Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-based policy optimization. In *Advances in Neural Information Processing Systems*, 2019. URL https://arxiv.org/abs/1906.08253.

Michael Janner, Yilun Du, Joshua B. Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. In *International Conference on Machine Learning*, 2022. URL https://arxiv.org/abs/2205.09991.

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.

Hilbert J. Kappen. Path integrals and symmetry breaking for optimal control theory. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(11):P11011, 2005. doi: 10.1088/1742-5468/2005/11/P11011.

Tyler Kastner, Murat A. Erdogdu, and Amir-massoud Farahmand. Distributional model equivalence for risk-sensitive reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/b0cd0e8027309ea050951e758b70d60e-Abstract-Conference.html.

Matthew T. Kaufman, Mark M. Churchland, Stephen I. Ryu, and Krishna V. Shenoy. Cortical activity in the null space: Permitting preparation without movement. *Nature Neuroscience*, 17:440–448, 2014. doi: 10.1038/nn.3643.

Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine Learning*, 49(2–3):209–232, 2002. URL https://web.eecs.umich.edu/~baveja/journalpapers.html.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, et al. OpenVLA: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024. URL https://arxiv.org/abs/2406.09246.

Edgar D. Klenske and Philipp Hennig. Dual control for approximate Bayesian reinforcement learning. *Journal of Machine Learning Research*, 17(127):1–30, 2016. URL https://www.jmlr.org/papers/v17/15-162.html.

Alexander S. Klyubin, Daniel Polani, and Chrystopher L. Nehaniv. Empowerment: A universal agent-centric measure of control. In *2005 IEEE Congress on Evolutionary Computation*, volume 1, pp. 128–135, 2005. doi: 10.1109/CEC.2005.1554676.

Ashish Kumar, Zipeng Fu, Deepak Pathak, and Jitendra Malik. RMA: Rapid motor adaptation for legged robots. In *Robotics: Science and Systems*, 2021. URL https://arxiv.org/abs/2107.04034.

Lucien Le Cam. Sufficiency and approximate sufficiency. *The Annals of Mathematical Statistics*, 35 (4):1419–1455, 1964. doi: 10.1214/aoms/1177700372.

Sergey Levine. Reinforcement learning and control as probabilistic inference: Tutorial and review. *arXiv preprint arXiv:1805.00909*, 2018. URL https://arxiv.org/abs/1805.00909.

Lin Li, Qihang Zhang, Yiming Luo, Shuai Yang, Ruilin Wang, Fei Han, Mingrui Yu, Zelin Gao, Nan Xue, Xing Zhu, Yujun Shen, and Yinghao Xu. Causal world modeling for robot control. *arXiv preprint arXiv:2601.21998*, 2026a. URL https://arxiv.org/abs/2601.21998.

Xinye Li, Lingshuai Lin, Lei Wang, Liuzhou Zhang, Jialin Cui, Qingshan Li, Guanchu Wang, Qingbin Liu, Xi Chen, Jiang Bian, and Wai Lam. ForgeWM: Progressive causal training for few-step action-conditioned video world models. *arXiv preprint arXiv:2608.14022*, 2026b. URL https://arxiv.org/abs/2608.14022.

Michael L. Littman, Richard S. Sutton, and Satinder Singh. Predictive representations of state. In *Advances in Neural Information Processing Systems*, volume 14, 2002. URL https://papers.nips.cc/paper/1983-predictive-representations-of-state.

Sam Lobel and Ronald Parr. An optimal tightness bound for the simulation lemma. *Reinforcement Learning Journal*, 2:785–797, 2024. URL https://rlj.cs.umass.edu/2024/papers/Paper106.html.

Marcelo G. Mattar and Nathaniel D. Daw. Prioritized memory access explains planning and hippocampal replay. *Nature Neuroscience*, 21:1609–1617, 2018. doi: 10.1038/s41593-018-0232-z.

R. Chris Miall, Lars O. D. Christensen, Owen Cain, and James Stanley. Disruption of state estimation in the human lateral cerebellum. *PLOS Biology*, 5(11):e316, 2007. doi: 10.1371/journal.pbio.0050316.

Vincent Micheli, Eloi Alonso, and François Fleuret. Transformers are sample-efficient world models. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2209.00588.

Ida Momennejad, Evan M. Russek, Jin H. Cheong, Matthew M. Botvinick, Nathaniel D. Daw, and Samuel J. Gershman. The successor representation in human reinforcement learning. *Nature Human Behaviour*, 1:680–692, 2017. doi: 10.1038/s41562-017-0180-8.

NVIDIA. Cosmos world foundation model platform for physical AI. *arXiv preprint arXiv:2501.03575*, 2025. URL https://arxiv.org/abs/2501.03575.

NVIDIA. Cosmos 3: Omnimodal world models for physical AI. *arXiv preprint arXiv:2606.02800*, 2026. URL https://arxiv.org/abs/2606.02800.

NVIDIA, Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, et al. GR00T N1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025. URL https://arxiv.org/abs/2503.14734.

Rolf Pfeifer, Max Lungarella, and Fumiya Iida. Self-organization, embodiment, and biologically inspired robotics. *Science*, 318(5853):1088–1093, 2007. doi: 10.1126/science.1145803.

Brad E. Pfeiffer and David J. Foster. Hippocampal place-cell sequences depict future paths to remembered goals. *Nature*, 497:74–79, 2013. doi: 10.1038/nature12112.

Physical Intelligence, Kevin Black, Noah Brown, et al. $\pi_{0.5}$: A vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025. URL https://arxiv.org/abs/2504.16054.

J. Andrew Pruszynski, Mohsen Omrani, and Stephen H. Scott. Goal-dependent modulation of fast feedback responses in primary motor cortex. *Journal of Neuroscience*, 34(13):4608–4617, 2014. doi: 10.1523/JNEUROSCI.4520-13.2014.

Rajesh P. N. Rao and Dana H. Ballard. Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2:79–87, 1999. doi: 10.1038/4580.

Naci Saldi and Serdar Yüksel. Geometry of information structures, strategic measures and associated stochastic control topologies. *Probability Surveys*, 19:450–532, 2022. doi: 10.1214/20-PS356.

Jeffrey A. Saunders and David C. Knill. Visual feedback control of hand movements. *Journal of Neuroscience*, 24(13):3223–3234, 2004. doi: 10.1523/JNEUROSCI.4319-03.2004.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, et al. Mastering Atari, Go, chess and shogi by planning with a learned model. *Nature*, 588:604–609, 2020. doi: 10.1038/s41586-020-03051-4.

Nicolas W. Schuck, Ming Bo Cai, Robert C. Wilson, and Yael Niv. Human orbitofrontal cortex represents a cognitive map of state space. *Neuron*, 91(6):1402–1412, 2016. doi: 10.1016/j.neuron.2016.08.019.

Kohei Sendai, Maxime Alvarez, Tatsuya Matsushima, Yutaka Matsuo, and Yusuke Iwasawa. Leave no observation behind: Real-time correction for VLA action chunks. *arXiv preprint arXiv:2509.23224*, 2025. URL https://arxiv.org/abs/2509.23224.

Satinder Singh, Michael R. James, and Matthew R. Rudary. Predictive state representations: A new theory for modeling dynamical systems. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, pp. 512–519, 2004. URL https://arxiv.org/abs/1207.4167.

Richard D. Smallwood and Edward J. Sondik. The optimal control of partially observable Markov processes over a finite horizon. *Operations Research*, 21(5):1071–1088, 1973. doi: 10.1287/opre.21.5.1071.

James E. Smith and Robert L. Winkler. The optimizer's curse: Skepticism and postdecision surprise in decision analysis. *Management Science*, 52(3):311–322, 2006. doi: 10.1287/mnsc.1050.0451. URL https://jimsmith.host.dartmouth.edu/wp-content/uploads/2022/04/The_Optimizers_Curse.pdf.

Marc A. Sommer and Robert H. Wurtz. Influence of the thalamus on spatial visual processing in frontal cortex. *Nature*, 444:374–377, 2006. doi: 10.1038/nature05279.

Kimberly L. Stachenfeld, Matthew M. Botvinick, and Samuel J. Gershman. The hippocampus as a predictive map. *Nature Neuroscience*, 20:1643–1653, 2017. doi: 10.1038/nn.4650.

Jayakumar Subramanian, Amit Sinha, Raihan Seraj, and Aditya Mahajan. Approximate information state for approximate planning and reinforcement learning in partially observed systems. *Journal of Machine Learning Research*, 23(12):1–83, 2022. URL https://www.jmlr.org/papers/v23/20-1165.html.

Haofeng Sun, Jiangbo Pei, Fei Kang, Zexiang Liu, Yaokun Li, Boyi Jiang, Hua Xue, Cindy Zhou, Wei Li, Yichen Wei, Mengyin An, Fanliang Zhao, Biao Jiang, Zile Wang, Yang Liu, and Yangguang Li. Riemann-1.0: An embodied world action model for physical AI. *arXiv preprint arXiv:2608.27033*, 2026. URL https://arxiv.org/abs/2608.27033.

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.

Emanuel Todorov. Linearly-solvable Markov decision problems. In *Advances in Neural Information Processing Systems*, volume 19, 2006. URL https://papers.nips.cc/paper/3002-linearly-solvable-markov-decision-problems.

Emanuel Todorov and Michael I. Jordan. Optimal feedback control as a theory of motor coordination. *Nature Neuroscience*, 5:1226–1235, 2002. doi: 10.1038/nn963.

Ya-weng Tseng, Jörn Diedrichsen, John W. Krakauer, Reza Shadmehr, and Amy J. Bastian. Sensory prediction errors drive cerebellum-dependent adaptation of reaching. *Journal of Neurophysiology*, 98(1):54–62, 2007. doi: 10.1152/jn.00266.2007.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, et al. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. URL https://arxiv.org/abs/1812.01717.

Hado van Hasselt. Double Q-learning. In *Advances in Neural Information Processing Systems*, volume 23, pp. 2613–2621, 2010. URL https://papers.neurips.cc/paper_files/paper/2010/hash/091d584fced301b442654dd8c23b3fc9-Abstract.html.

Claas Voelcker, Anastasiia Pedan, Arash Ahmadian, Romina Abachi, Igor Gilitschenski, and Amir-massoud Farahmand. Calibrated value-aware model learning with probabilistic environment models. *arXiv preprint arXiv:2505.22772*, 2025. URL https://arxiv.org/abs/2505.22772.

Jan C. Willems. The behavioral approach to open and interconnected systems. *IEEE Control Systems Magazine*, 27(6):46–99, 2007. doi: 10.1109/MCS.2007.906923. URL https://homes.esat.kuleuven.be/~sistawww/smc/jwillems/Articles/JournalArticles/2007.1.pdf.

Robert C. Wilson, Andra Geana, John M. White, Elliot A. Ludvig, and Jonathan D. Cohen. Humans use directed and random exploration to solve the explore–exploit dilemma. *Journal of Experimental Psychology: General*, 143(6):2074–2081, 2014. doi: 10.1037/a0038199.

Daniel M. Wolpert and Mitsuo Kawato. Multiple paired forward and inverse models for motor control. *Neural Networks*, 11(7–8):1317–1329, 1998. doi: 10.1016/S0893-6080(98)00066-5.

Daniel M. Wolpert, Zoubin Ghahramani, and Michael I. Jordan. An internal model for sensorimotor integration. *Science*, 269(5232):1880–1882, 1995. doi: 10.1126/science.7569931.

Mengjiao Yang, Yilun Du, Kamyar Ghasemipour, et al. Learning interactive real-world simulators. *arXiv preprint arXiv:2310.06114*, 2023a. URL https://arxiv.org/abs/2310.06114.

Mengjiao Yang, Dale Schuurmans, Pieter Abbeel, and Ofir Nachum. Dichotomy of control: Separating what you can control from what you cannot. In *International Conference on Learning Representations*, 2023b. URL https://arxiv.org/abs/2210.13435.

Yuxue Yang, Shuyao Shang, Jiahe Wang, Zitong Zhou, Liang Tan, Junhan Zeng, Ruizhi Li, Junyan Li, Yu Liu, Xiao Yang, Yong Li, Jun Zhu, Hongsheng Li, Tieniu Tan, Lue Fan, and Zhaoxiang Zhang. WorldExam: Benchmarking world models from apparent appearance to inherent reactivity. *arXiv preprint arXiv:2608.02603*, 2026. URL https://arxiv.org/abs/2608.02603.

Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, et al. World action models are zero-shot policies. *arXiv preprint arXiv:2602.15922*, 2026. URL https://arxiv.org/abs/2602.15922.

Yang Yu, Shiyuan Zhang, Yifei Sheng, Haoxiang Ren, and Haoxin Lin. How should world models be evaluated for embodied decision-making? a decision-making-centric position. *arXiv preprint arXiv:2606.15032*, 2026. URL https://arxiv.org/abs/2606.15032.

Serdar Yüksel and Naci Saldi. Convex analysis in decentralized stochastic control, strategic measures, and optimal solutions. *SIAM Journal on Control and Optimization*, 55(1):1–28, 2017. doi: 10.1137/15M1049129.

Amy Zhang, Rowan McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. Learning invariant representations for reinforcement learning without reconstruction. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=-2FCwDKRREu.

Jiahan Zhang, Muqing Jiang, Nanru Dai, Taiming Lu, Arda Uzunoglu, Shunchi Zhang, Yana Wei, Jiahao Wang, Vishal M. Patel, Paul Pu Liang, Daniel Khashabi, Cheng Peng, Rama Chellappa, Tianmin Shu, Alan Yuille, Yilun Du, and Jieneng Chen. World-in-world: World models in a closed-loop world. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=yDmb7xAfeb.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. DINO-WM: World models on pre-trained visual features enable zero-shot planning. In *International Conference on Machine Learning*, 2025. URL https://arxiv.org/abs/2411.04983.

Brian D. Ziebart, J. Andrew Bagnell, and Anind K. Dey. Modeling interaction via the principle of maximum causal entropy. In *International Conference on Machine Learning*, 2010. URL https://publications.ri.cmu.edu/modeling-interaction-via-the-principle-of-maximum-causal-entropy/.

APPENDIX CONTENTS

Each entry links to its derivation or experimental protocol. Study identifiers A–I are shared by the manuscript and accompanying code.

THEORY AND PROOFS

Trajectories, observations, and hidden completions 19
Causal interfaces and execution 20
Substitution bounds and attainable outcome values 21
Projection onto legal policies 22
Success conditioning and information in actions 23
Local realization and dynamic projection 24
Conditioned paths and physical response constraints 25
Coverage and causal identification 26
State representations and possible continuations 27
Physical implementation of a posterior marginal 28
Bounded tasks witnessing unattainability 28
Selection errors when a skill library expands 29
Gaussian projection over all randomized controls 30
Partial actuation and correlation constraints 31
Sensor timing and residual uncertainty 31
Analytic constructions in Figure 1 33

EXPERIMENTS AND REPRODUCTION

Continuous contact: dynamics and shared protocol 35
H. Information timing and action commitment 37
I. Response collection and deployment 41
Skill expansion and all initial skill banks 49
Finite processes: study and reporting guide 55
A. Command calibration 56
B. Model substitution and planning transfer 58
C. Preparation, information, and physical response 61
D. Learning from successful trajectories 63
E. Coverage and hidden action selection 65
F. Learned visual models 68
G. Available response mixtures 71
Illustrated guide to the experimental questions 87

INTERPRETATION AND MODEL DESIGN

Related theory and contemporary world models 74
Computational interpretation in biological cognition 78
Prediction, representation, and execution interfaces 81

A MATHEMATICAL DETAILS

The derivations follow the order of the physical questions in the main text. We first specify what a controller observes and what a task evaluates. We then compare a common controller across systems, project requested trajectories onto legal behavior, and eliminate the action variables to characterize attainable outcome laws. The final continuous calculation gives the exact Gaussian projection over all randomized controls. Each result keeps the observation interface and the available decisions explicit.

A.1 FORMAL SETTING AND THE CHOICE OF TRAJECTORY

The chronological and local realization results use finite sets and a finite horizon. Appendix A.13 treats the continuous additive case. In the finite setting, probability kernels can be treated as conditional probability tables. Conditional laws at events of probability zero may be assigned arbitrary values. We use natural logarithms, with $0\log(0/p) = 0$ and $q\log(q/0) = +\infty$ for $q > 0$. Total variation is $\mathrm{TV}(P, Q) = \frac{1}{2}\sum_{\omega} |P(\omega) - Q(\omega)|$. Relative entropy is $D_{\mathrm{KL}}(Q\|P) = \sum_{\omega} Q(\omega)\log[Q(\omega)/P(\omega)]$. These conventions also specify the numerical calculations.

The physical path is $\omega = (x_0, o_0, a_0, x_1, o_1, \ldots, a_{T-1}, x_T, o_T)$. Its public part is $C_T = (o_0, a_0, \ldots, a_{T-1}, o_T)$. A policy may use all of $H_t = (o_0, a_0, \ldots, o_t)$ and independent randomization. It may implement a recurrent controller, a planner, or an algorithm that updates its model online. A random seed chosen before interaction is independent of the physical initial state; an internal memory update is a causal function of observations, actions, and private randomness. The resulting process can be written using behavioral kernels $\pi_t(a \mid H_t)$. No unrecorded observation is available to this policy.

A POMDP has the factorization

$$P_E^{\pi}(\omega) = p_0(x_0)g_0(o_0 \mid x_0)\prod_{t<T} \pi_t(a_t \mid H_t)p_t(x_{t+1} \mid x_t, a_t)g_{t+1}(o_{t+1} \mid x_{t+1}). \tag{8}$$

The more general kernel $K_t(\xi_{t+1} \mid \bar{H}_t, a_t)$ in the main text includes hidden memory, nonstationarity, and history-dependent physical responses. The response law on public histories can be defined by marginalizing hidden states. Under legal policies it has an interventional meaning at histories with positive probability. A shared observation and action interface is required when two response maps are compared; the two systems need no common hidden-state representation.

A task can be specified on the full public transcript, on a measurable projection of that transcript, or, in a simulator with access to physical variables, on an explicitly enlarged record. The choice affects the scientific question. A force signal used only for evaluation can be included in the full record while remaining absent from the policy's information. In that case the policy is still conditioned on H_t, and the physical record has a finer history than the controller. The entropy identities below retain this distinction. In particular, an evaluator's access to a hidden variable does not give the controller access to that variable.

A.1.1 OBSERVABLE TARGETS AND HIDDEN COMPLETIONS

The projection in Theorem 2 specifies a complete hidden path. The projection in Theorem 3 specifies an operational process. Their relationship can be stated exactly. Let $\rho : \Omega \to \mathcal{C}$ forget the hidden coordinates and let q be a law on $\mathcal{C}$. Write $P_C^{\pi} = \rho_{\#}P_E^{\pi}$.

Proposition 4 (Completion of an observable target)**.** *In finite spaces,*

$$\inf_{\pi} D_{\mathrm{KL}}(q\|P_C^{\pi}) = \inf_{Q:\rho_{\#}Q=q} \inf_{\pi} D_{\mathrm{KL}}(Q\|P_E^{\pi}). \tag{9}$$

Consequently, an observable target is attainable exactly when it has a hidden completion that is a legal physical path law.

Proof. For fixed π with finite marginal divergence, disintegration gives

$$D_{\mathrm{KL}}(Q\|P_E^{\pi}) = D_{\mathrm{KL}}(q\|P_C^{\pi}) + \sum_{c} q(c)D_{\mathrm{KL}}\big(Q(\cdot \mid c)\|P_E^{\pi}(\cdot \mid c)\big).$$

The conditional term is minimized at zero by taking the physical conditional completion $Q(\cdot \mid c) = P_E^\pi(\cdot \mid c)$. If $q(c) > 0 = P_C^\pi(c)$, both sides of this fixed-policy minimization are infinite. Taking the two infima in either order proves the equality. The set of finite-horizon legal policies is a finite product of compact simplices, and its map into path laws is continuous. Its image is compact. Thus a zero infimum in either attainability statement has a realizing policy. □

An observable goal leaves the hidden trajectory free. The minimization over completions uses that freedom: a requested endpoint may be attainable through a hidden history different from the one selected by a demonstration. A feasibility claim therefore specifies the record whose distribution is requested, whether it is the hidden path, the public interaction, or the endpoint.

A.2 Causal interfaces and consistency across plans

An abstract controller may issue commands U_t and receive abstract observations V_t, while the physical interface exposes A_s, O_s. A causal interface Γ is a fixed transducer with private memory and independent randomness. It maps its current abstract command and available physical history to physical actions, and it returns an abstract observation using the physical observations obtained so far. Its stopping rule for a macro command is a stopping time. The transducer has no access to future observations or unavailable physical state.

Composition of the abstract policy with this transducer gives a legal physical policy $\Gamma[\pi]$. The relevant model discrepancy is

$$\Delta_{\Gamma,\Pi,T} = \sup_{\pi\in\Pi} \mathrm{TV}\big(\mathrm{Law}_M^\pi(C_T^{\mathrm{abs}}), \mathrm{Law}_E^{\Gamma[\pi]}(C_T^{\mathrm{abs}})\big).$$

The two laws are on the same abstract transcript. A different physical command alphabet is harmless once this translation is fixed. A translation optimized independently for every task changes the quantified object and has to be included in the chosen policy class. An unknown translation learned during deployment is part of the policy; its probing actions, information, and time count in the interaction.

If macro duration, energy, or intermediate failures matter to the task, the interface reports those quantities and the evaluation includes them. Comparing one primitive time step to one macro completion without specifying duration compares different resource budgets. A skill can improve the high-level interface by packaging an available primitive behavior; this observation alone does not establish a gain over the complete primitive policy class.

A.2.1 A causal realization of open-loop predictions

For every deterministic action plan $\mathbf{a} = (a_0, \dots, a_{T-1})$, suppose a generator supplies a probability law $Q^{\mathbf{a}}$ on $O_{0:T}$. A causal realization is an initial law g_0 and kernels $g_{t+1}(o_{t+1} \mid o_{0:t}, a_{0:t})$ whose product equals every $Q^{\mathbf{a}}$.

Proposition 5 (Prefix consistency)**.** *Such a realization exists if and only if, for every t and any two plans agreeing on $a_{0:t-1}$, their marginal laws on $O_{0:t}$ agree. The realization determines every adaptive interaction law on its reachable histories.*

Proof. A causal product makes the marginal on $O_{0:t}$ depend only on $a_{0:t-1}$, proving necessity. For sufficiency, let $q_t(o_{0:t} \mid a_{0:t-1})$ be the common prefix marginal. Define $g_0 = q_0$ and, whenever the denominator is positive,

$$g_{t+1}(o_{t+1} \mid o_{0:t}, a_{0:t}) = \frac{q_{t+1}(o_{0:t+1} \mid a_{0:t})}{q_t(o_{0:t} \mid a_{0:t-1})}.$$

Consistency and marginalization make this a probability kernel. Choose arbitrary probability rows at zero denominators. Multiplication telescopes to q_T, giving the desired open-loop laws. If a denominator is zero, every plan with that prefix assigns zero probability to that observation prefix. An adaptive policy reaching the same prefix has selected the same past actions, so this event is also unreachable under that policy. Arbitrary off-support rows therefore cannot change an adaptive interaction law. □

The statement concerns families of distributions. A single sampled future need not remain unchanged when a random seed is resampled. The required invariance is the law of the generated past under a fixed action prefix.

A.3 Substitution, planning transfer, and taskwise capability

Proof of Theorem 1. For finite probability measures, TV duality gives

$$\Delta = \sup_{\pi \in \Pi} \sup_{0 \le F \le 1} |\mathbb{E}_M^\pi F - \mathbb{E}_E^\pi F|. \tag{10}$$

For a fixed pair of laws, one maximizing test is the indicator of the set where $P(\omega) > Q(\omega)$. Taking the supremum over π gives (10). With equal hypotheses, a terminal decision rule based on the transcript has Bayes accuracy $(1 + \mathrm{TV}(P_M^\pi, P_E^\pi))/2$. Every legal adaptive test is a policy followed by such a decision rule. Optimizing gives $(1 + \Delta)/2$; a supremum is used if the policy class is not compact.

For reward transfer, put $V_S(\pi) = \mathbb{E}_S^\pi F$. Let $\pi_E^\star$ be ε-optimal in the environment and let π_M be η-optimal in the model. Then

$$\begin{aligned}\sup_\pi V_E(\pi) - V_E(\pi_M) &\le \varepsilon + V_E(\pi_E^\star) - V_M(\pi_E^\star) + V_M(\pi_E^\star) - V_M(\pi_M) \\ &\quad + V_M(\pi_M) - V_E(\pi_M) \\ &\le \varepsilon + 2\Delta + \eta.\end{aligned}$$

Let $\varepsilon \downarrow 0$. A reward with range of width B has bound $2B\Delta + \eta$ when η is measured in the reward's original units.

For the target bound, the reverse triangle inequality gives $|\,\mathrm{TV}(q, \mathcal{R}_E(\pi)) - \mathrm{TV}(q, \mathcal{R}_M(\pi))| \le \Delta$. It also gives $|\chi_E(q) - \chi_M(q)| \le \Delta$. Therefore

$$\mathrm{TV}(q, \mathcal{R}_E(\pi_M)) \le \mathrm{TV}(q, \mathcal{R}_M(\pi_M)) + \Delta \le \chi_M(q) + \eta + \Delta \le \chi_E(q) + \eta + 2\Delta.$$

The lower inequality follows from the definition of χ_E. The same proof applies after a common projection of the transcript because total variation contracts under measurable maps. □

A.3.1 Sharpness and interpretation

For $0 < \delta \le 1/2$, consider one step and two actions. In the real environment their Bernoulli success probabilities are $1/2 + \delta$ and $1/2 - \delta$. The model assigns probability $1/2$ to both. Its feedback error is δ. Selecting the second model-optimal action gives regret 2δ. For target $q = \delta_{\mathrm{success}}$, the same example has physical floor $1/2 - \delta$ and deployed distance $1/2 + \delta$, also attaining the target bound. These examples show that a uniform improvement of the factor two requires further assumptions about the planner or model errors. The bound can always be capped by one for normalized rewards or TV distances.

The testing equality gives Δ an operational meaning. A policy determines which interaction takes place; a terminal test determines which difference in its consequences is measured. Any such pair provides a lower bound on Δ. An upper bound over the full policy class additionally requires known kernels, coverage conditions, regularity, or a finite class with statistical guarantees. Appendix A.8 gives the connection with prediction errors measured under a data policy.

A.3.2 What independently optimized task scores preserve

Let $Y = \rho(C_T)$ be the outcome used to score a task and define $\mathcal{S}_S = \{\mathrm{Law}_S^\pi(Y) : \pi \in \Pi\}$. On finite Y,

$$\sup_{P \in \mathcal{S}_M} \mathbb{E}_P F = \sup_{P \in \mathcal{S}_E} \mathbb{E}_P F \quad \text{for every } F$$

holds if and only if the closed convex hulls of $\mathcal{S}_M$ and $\mathcal{S}_E$ agree. The forward implication follows from separation: a point outside one closed convex hull admits a linear functional whose value exceeds the support function of that hull. The reverse implication follows because a linear functional has the same supremum on a set and its closed convex hull.

This equivalence preserves the set of outcome distributions after independent optimization. In the actuator reversal experiment, reversing every command is a bijection between policies realizing each final-location law. The two systems therefore have the same optimal value for every final-location reward. For a given unchanged command, their consequences differ. Including action coordinates in Y records this correspondence and distinguishes the attainable sets. The choice of evaluation record specifies exactly which of these properties a score measures.

A.3.3 Task classes and increasing capability

For a stated collection of tests $\mathcal{F}$, one can replace TV by $d_{\mathcal{F}}(P,Q) = \sup_{F\in\mathcal{F}} |\mathbb{E}_P F - \mathbb{E}_Q F|$. The reward transfer bound holds for every $F \in \mathcal{F}$. Its meaning is restricted to that class. This allows an application to ignore variations of appearance that have no role in its rewards, safety signals, or future decisions. The chosen test class defines the intended scope of model use. Value equivalence preserves the planning quantities associated with a reward specification (Grimm et al., 2020; 2021); distributional model equivalence preserves return laws for risk-sensitive objectives (Kastner et al., 2023). Taking all bounded tests of the full transcript additionally retains the temporal association among observations, actions, and outcomes. A finite collection of empirical probes establishes agreement for the tested quantities; a uniform upper bound on $d_{\mathcal{F}}$ requires control of the whole declared class.

For fixed systems, interface, and horizon, $\Pi_1 \subseteq \Pi_2$ implies

$$\Delta_{\Pi_1,T} \le \Delta_{\Pi_2,T}, \qquad \chi_{E,\Pi_2}(q) \le \chi_{E,\Pi_1}(q).$$

The first inequality follows by maximizing over a larger set, the second by minimizing over it. More available policies can improve physical attainability and simultaneously expose model errors that earlier policies could not reach. A world model validated for a smaller set of skills therefore needs additional tests when its controller acquires new skills.

A.4 The projection onto legal policies

A policy can change the action factors in a trajectory law. Its information determines which histories those factors may distinguish. The initial distribution and the response after each action remain fixed by the environment. This separation makes the projection explicit. Factor an arbitrary full-path law chronologically as

$$Q(\omega) = Q_0(\xi_0) \prod_{t<T} q_t(a_t \mid \bar{H}_t) L_t(\xi_{t+1} \mid \bar{H}_t, a_t).$$

Write $\pi_t^Q(a \mid h) = Q(A_t = a \mid H_t = h)$. Define

$$B_E(Q) = D_{\mathrm{KL}}(Q_0 \| P_0) + \sum_{t<T} I_Q(A_t; X_{0:t} \mid H_t) + \sum_{t<T} \mathbb{E}_Q D_{\mathrm{KL}}(L_t \| K_t).$$

Proposition 6 (Full policy projection identity)**.** *For every legal π,*

$$D_{\mathrm{KL}}(Q \| P_E^\pi) = B_E(Q) + \sum_{t<T} \mathbb{E}_{H_t\sim Q} D_{\mathrm{KL}}\big(\pi_t^Q(\cdot \mid H_t) \| \pi_t(\cdot \mid H_t)\big). \tag{11}$$

The equality is interpreted in the extended nonnegative reals. In particular, Theorem 2 follows by setting $\pi = \pi^Q$.

Proof. Expand the trajectory likelihood ratio in chronological order. The initial and physical response factors contribute fixed terms; only the action factors contain the policy being optimized. The chain rule gives

$$D_{\mathrm{KL}}(Q \| P_E^\pi) = D_{\mathrm{KL}}(Q_0 \| P_0) + \sum_{t<T} \mathbb{E}_Q \log \frac{q_t(A_t \mid \bar{H}_t)}{\pi_t(A_t \mid H_t)} + \sum_{t<T} \mathbb{E}_Q D_{\mathrm{KL}}(L_t \| K_t).$$

At each action factor, insert the target conditional $\pi_t^Q(A_t \mid H_t)$, which uses exactly the information available to a legal policy. This separates two requirements: dependence of the requested action

on hidden history, and the discrepancy between two legal action distributions. The first part has expectation

$$\mathbb{E}_Q \log \frac{Q(A_t \mid X_{0:t}, H_t)}{Q(A_t \mid H_t)} = I_Q(A_t; X_{0:t} \mid H_t).$$

The second has expectation $\mathbb{E}_{H_t \sim Q} D_{\mathrm{KL}}(\pi_t^Q \| \pi_t)$. Each resulting term is nonnegative. If a physical response or initial event has zero probability and positive target probability, both the original divergence and the corresponding fixed term are infinite. If a policy excludes a target action, its final conditional divergence is infinite. Thus no subtraction of infinities is needed. □

When the projection value is finite, every minimizing policy agrees with π^Q on histories of positive target probability. Choices on null target histories are arbitrary. The minimum uses the full class of randomized policies, which permits a separate action distribution at every available history. A deterministic or parametrized controller can impose additional restrictions. For such a class, Equation (11) gives the exact excess cost of each policy. A resource budget can be represented in the physical history and the feasible action sets, with the target restricted to legal actions.

The excess policy term averages over the requested histories. Fitting the target action conditionals therefore solves the action part of the problem. During execution, those actions still encounter the physical initial and response laws. The complete target is realized precisely when all three terms in $B_E(Q)$ vanish: its initial law is physical, its actions depend only on available information, and its response conditionals agree with the environment on target histories. These requirements describe preparation, information, and response within the same chronological process.

A.4.1 An exact likelihood-ratio criterion

Let μ have positive probability for every action at every available history. A full-path law Q is attainable if and only if it is absolutely continuous with respect to P_E^μ and its likelihood ratio has the form

$$\frac{Q(\omega)}{P_E^\mu(\omega)} = \prod_{t<T} \frac{\pi_t(a_t \mid H_t)}{\mu_t(a_t \mid H_t)} \quad P_E^\mu\text{-almost surely} \tag{12}$$

for legal normalized kernels π_t. If $Q = P_E^\pi$, cancel the common physical factors to obtain the criterion. Conversely, multiplying (12) by the baseline law gives the factorization of P_E^π. Absolute continuity follows for every legal policy because the baseline has full action support and shares the same physical kernels. The normalization of each replacement kernel matters; an arbitrary product of positive history functions need not define legal action probabilities.

A.4.2 Additional information on a fixed physical record

Suppose an extra variable S_t already belongs to the full past record and is revealed before action A_t. Extend the policy information from H_t to (H_t, S_t) while keeping the full target and physical response fixed. Then

$$I_Q(A_t; X_{0:t} \mid H_t) = I_Q(A_t; S_t \mid H_t) + I_Q(A_t; X_{0:t} \mid H_t, S_t)$$

when S_t is a function of the hidden and public past. The reduction in this action's information cost is $I_Q(A_t; S_t \mid H_t)$. The resulting policy class contains the old class, so the attainable full-path KL floor cannot increase. Installing a new noisy sensor also changes the physical record and its observation channel. Such a comparison specifies the target on the enlarged record. Experiment C compares targets aligned with that record and a target whose original description is held fixed.

A.5 Conditioned futures and the information in action

Fix the initial public history $H_0 = h_0$. Every law and information quantity in this section is conditional on that event. Let $P = P_E^\mu$ for a legal behavior policy μ, and let a finite label G be generated from the full path, possibly by an additional random labeling channel. Let $Q_g = P(\cdot \mid G = g)$ for labels of positive probability.

For fixed g, (11) with $\pi = \mu$ yields

$$\inf_\pi D_{\mathrm{KL}}(Q_g \| P_E^\pi) = D_{\mathrm{KL}}(Q_g \| P) - \sum_{t<T} \mathbb{E}_{H_t \sim Q_g} D_{\mathrm{KL}}\big(Q_g(A_t \mid H_t) \| \mu_t(A_t \mid H_t)\big).$$

Here all displayed terms are finite whenever the conditional label has positive probability on a finite baseline support. Average over G. The first term becomes $I(G;\Omega)$. Since $P(A_t \mid H_t) = \mu_t(A_t \mid H_t)$, each subtracted term becomes $I(G; A_t \mid H_t)$. This gives the identity

$$\mathbb{E}_G \inf_{\pi} D_{\mathrm{KL}}(Q_G \| P_E^{\pi}) = I(G;\Omega) - \sum_{t<T} I(G; A_t \mid H_t). \tag{13}$$

An alternative reading comes from chronological chain rules. Since the baseline action ignores the hidden history given H_t, the averaged decomposition is

$$\begin{aligned}\mathbb{E}_G \inf_{\pi} D_{\mathrm{KL}}(Q_G \| P_E^{\pi}) = I(G;\xi_0) &+ \sum_{t<T} I(A_t; X_{0:t} \mid H_t, G) \\ &+ \sum_{t<T} I(G; \xi_{t+1} \mid \bar{H}_t, A_t).\end{aligned} \tag{14}$$

The initial observation is fixed, so the first term concerns remaining initial hidden information. The middle term is a dependence induced by conditioning on the future. The final terms concern evidence about responses sampled after actions. To verify the equality, apply the mutual-information chain rule to the chronological path and use

$$I(G; A_t \mid \bar{H}_t) - I(G; A_t \mid H_t) = I(A_t; X_{0:t} \mid H_t, G) - I(A_t; X_{0:t} \mid H_t),$$

whose last term is zero under the legal baseline.

A.5.1 Privileged demonstrations

Let the demonstration law P use the correct initial law and physical responses but allow actions to depend on $\bar{H}_t$. Its closest legal policy is $\bar{\mu}_t = P(A_t \mid H_t)$, with baseline cost $b(P) = \sum_t I_P(A_t; X_{0:t} \mid H_t)$. Averaging the log action ratios between P and $P_E^{\bar{\mu}}$ under the conditionals $Q_g = P(\cdot \mid g)$ gives the identity

$$\mathbb{E}_G \inf_{\pi} D_{\mathrm{KL}}(Q_G \| P_E^{\pi}) = b(P) + I(G;\Omega) - \sum_{t<T} I(G; A_t \mid H_t). \tag{15}$$

One proof begins with $\mathbb{E}_G D_{\mathrm{KL}}(Q_G \| P_E^{\bar{\mu}}) = I(G;\Omega) + D_{\mathrm{KL}}(P \| P_E^{\bar{\mu}})$ and then subtracts the averaged action projection costs, which are the conditional mutual informations on the right. The term $b(P)$ measures the action dependence already present before conditioning on the label. The remaining terms describe how selecting a future label changes the target. To compare information access for one requested behavior, hold that target law fixed while changing the information available to its controller.

These identities quantify trajectory divergence in nats. Task success is measured in probability. For any particular pair of laws, Pinsker's inequality bounds their event discrepancies by their relative entropy. Experiment D reports both quantities, retaining their distinct interpretations.

A.6 Eliminating actions and projecting transition laws

Let $Z_t = \zeta_t(H_t)$ take values in a finite space. Suppose the nonempty set of available actions $\mathcal{A}_t(Z_t)$ is determined by this summary, and, for every legal policy and available action, the controlled sufficiency condition holds:

$$P(Z_{t+1} = z' \mid H_t, A_t = a) = K_t^a(z' \mid Z_t).$$

The initial law $P_{Z,0}$ is fixed. Let $\mathcal{K}_t(z) = \{\sum_{a \in \mathcal{A}_t(z)} \lambda_a K_t^a(\cdot \mid z) : \lambda \in \mathcal{P}(\mathcal{A}_t(z))\}$. It is a compact convex polytope. The equality of conditional responses is required on physically reachable history-action pairs; rows elsewhere may be chosen arbitrarily. Starting from the physical initial law, any target satisfying the membership criterion reaches only such pairs under the realizing policy. A target may retain dependence on its entire summary history even when the physical response depends only on the current summary. Write $L_t^Q(\cdot \mid Z_{0:t})$ for this target conditional. Randomized policies may use that entire history to select an action mixture. The quantitative form of Theorem 3 is

$$\inf_{\pi} D_{\mathrm{KL}}(Q \| P_Z^{\pi}) = D_{\mathrm{KL}}(Q_0 \| P_{Z,0}) + \sum_{t<T} \mathbb{E}_Q \min_{\kappa \in \mathcal{K}_t(Z_t)} D_{\mathrm{KL}}\Big(L_t^Q(\cdot \mid Z_{0:t}) \| \kappa\Big). \tag{16}$$

Proof of Theorem 3. For any policy, condition on $Z_{0:t} = z_{0:t}$. Iterated conditioning gives

$$P_Z^\pi(Z_{t+1} \in \cdot \mid z_{0:t}) = \sum_a P^\pi(A_t = a \mid z_{0:t}) K_t^a(\cdot \mid z_t) \in \mathcal{K}_t(z_t).$$

The equality holds even if the policy uses additional distinctions in H_t: those distinctions affect the mixing weights, and controlled sufficiency fixes the response for each action. The entropy chain rule therefore lower bounds $D_{\mathrm{KL}}(Q\|P_Z^\pi)$ by the right side of (16).

Conversely, at each target history choose a minimizing kernel $\kappa_t^\star \in \mathcal{K}_t(z_t)$. Lower semicontinuity on a compact simplex guarantees a minimum in the extended reals. Choose mixing weights $\lambda_t^\star(\cdot \mid z_{0:t})$ that represent this kernel. The legal policy $\pi_t(a \mid H_t) = \lambda_t^\star(a \mid \zeta_0(H_0), \ldots, \zeta_t(H_t))$ realizes all selected conditionals. At histories absent from Q, choose arbitrary weights. Its divergence equals the sum of the minima by the chain rule. If a selected target prefix is impossible under the physical initial law or response support, both values are infinite. This proves the projection formula.

If the initial laws agree and every target conditional lies in the corresponding control set, choose weights representing the target conditionals exactly. Induction over time proves equality of the path laws. Conversely, equality of path laws forces equality of the initial law and equality of every conditional on positive target prefixes. Each physical conditional belongs to its control set, proving necessity. □

The stagewise optimization is possible because a policy may choose a distinct action mixture at every available history. A constrained policy class may couple those choices. An energy budget can be included in Z_t if its remaining value and dynamics are available; its feasible actions then depend on the augmented state. For a learned latent state, controlled sufficiency is the condition that connects the formula to the physical process. Approximate summaries, including finite quantizations of a belief state, introduce an additional prediction error.

A.6.1 Terminal equivalence and composition of skills

A macro action is defined by a primitive policy and a termination rule. To compose macro models, their output state must retain information sufficient for the next macro's response. An endpoint image alone can fail this requirement. Consider two macros ending at the same visible position and image. The first leaves a hidden latch at zero and the second leaves it at one. A subsequent common action opens a door exactly when the latch equals one. Their immediate endpoint laws agree, but the two-macro transcripts have total variation one. The lost variable is relevant through a later intervention.

A valid coarse model can instead report the posterior over the latch, a predictive state that retains its effect, or the full public history needed to infer it. More generally, suppose two models of each macro have conditional transcript TV at most ε_j uniformly over the coarse histories reached before macro j, including duration and termination information. A sequential maximal coupling gives total transcript TV at most $\sum_j \varepsilon_j$, capped at one. Couple the two next-macro records conditional on agreement so far; disagreement at stage j has probability at most ε_j. A union bound completes the proof. The same argument works with expectations of conditional error under the coupling's agreed histories. The condition concerns predictions at the histories where a subsequent skill will be invoked. Appendix F.3 states the composition result with its full returned record.

A.7 Conditioning, stochastic branches, and local action directions

Corollary 7 (Realizing every positive tilt)**.** *Let $p \in \mathcal{P}(\{1, \ldots, m\})$ have full support, and let $\mathcal{K}$ be a closed subset of that simplex. If $p_j w_j / \sum_\ell p_\ell w_\ell \in \mathcal{K}$ for every positive vector w, then $\mathcal{K}$ is the full simplex. If $\mathcal{K}$ is the convex hull of finitely many primitive probability kernels, for each j some primitive kernel equals δ_j.*

Proof. For every strictly positive target q, choose $w_j = q_j/p_j$; the resulting tilt is q. Thus $\mathcal{K}$ contains the simplex interior and, by closure, its boundary. Since it is contained in the simplex, equality follows. If $\delta_j = \sum_a \lambda_a K^a$, then for every $i \neq j$, $0 = \sum_a \lambda_a K_i^a$. Nonnegativity implies $K_i^a = 0$ for every action with positive weight. Each such kernel therefore has all mass at j. At least one weight is positive. □

For continuous or infinitely supported outcomes, additional topology and compactness assumptions are needed to turn density of positive tilts into exact boundary attainability. The finite statement covers the branchwise criterion used here.

A.7.1 THE PHYSICAL ASSUMPTION IN INFERENCE FORMULATIONS

A trajectory posterior proportional to a passive law times task evidence generally changes its response conditionals. On a Markov state path with terminal weight $w(X_T)$, the conditional transition has the Doob form

$$L_t(x' \mid x) = K_t(x' \mid x)\frac{h_{t+1}(x')}{h_t(x)}, \qquad h_t(x) = \mathbb{E}[w(X_T) \mid X_t = x],$$

where $h_t(x) > 0$. It is a stochastic kernel because $h_t = K_t h_{t+1}$. Executing this posterior requires these transformed kernels to be physically available. In linearly solvable control, the controlled transition itself is a decision variable with a KL cost and support restrictions (Todorov, 2006). In path-integral control, tractability relies on particular relations between noise and control channels (Kappen, 2005). These models supply explicit conditions under which inference and control agree. For an arbitrary robot action set, Theorem 3 gives the additional membership test. A variational control method can preserve the physical dynamics while optimizing only action factors; then its feasible family already enforces the distinction (Levine, 2018).

A.7.2 LOCAL GEOMETRY OF CONTROLLABLE PROBABILITY CHANGES

Consider a smooth family of full-support response distributions p_u on a finite outcome set, with u_0 in the interior of an open action domain. Set $p_0 = p_{u_0}$ and let $s_i(y) = \partial_{u_i} \log p_u(y)|_{u_0}$ be its score functions. Normalization gives $\mathbb{E}_{p_0} s_i = 0$. Let a desired small change be $q_\epsilon(y) = p_0(y)[1 + \epsilon f(y)]$, where $\mathbb{E}_{p_0} f = 0$ and ϵ is small enough for positivity.

For any bounded direction v, Taylor expansion of the probability vectors and logarithm gives

$$D_{\mathrm{KL}}(q_\epsilon \| p_{u_0+\epsilon v}) = \frac{\epsilon^2}{2}\mathbb{E}_{p_0}\left[f - s^\top v\right]^2 + o(\epsilon^2). \tag{17}$$

The remainder is uniform for v in a fixed compact set under twice continuous differentiability. To verify the coefficient, write $p_{u_0+\epsilon v} = p_0(1 + \epsilon g + \epsilon^2 r + o(\epsilon^2))$ with $g = s^\top v$. Expanding both logarithms and multiplying by q_ϵ leaves the second-order coefficient $\frac{1}{2}\mathbb{E}(f-g)^2 - \mathbb{E}r$; normalization makes $\mathbb{E}r = 0$.

With Fisher matrix $J = \mathbb{E}_{p_0} ss^\top$ and $b = \mathbb{E}_{p_0} sf$, the quadratic minimum is

$$\frac{\epsilon^2}{2}\big(\mathbb{E}_{p_0} f^2 - b^\top J^\dagger b\big).$$

Indeed, a vector in the null space of J has zero score almost surely and is orthogonal to b, so b lies in the range of J. The minimizing score is the orthogonal projection of f onto the span of the action scores; $J^\dagger b$ supplies one minimizing direction. This expression describes the leading cost among regular local actions near u_0. Global attainability also depends on responses available elsewhere in the action family.

At an action constraint boundary, feasible first-order directions form a tangent cone, and the minimization is over that cone. For a discrete action mixture at a simplex boundary, new action weights likewise have one-sided directions. Replacing these cones by an unconstrained vector space can incorrectly declare a probability change controllable. An irreducible component under the known local response family concerns physical attainability. Further data can identify that family more accurately; enlarging the family requires a different available intervention.

A.8 PREDICTION, COVERAGE, AND CAUSAL IDENTIFICATION

Here $K_{S,t}(\cdot \mid h, a)$ denotes the next-observation law in system S given the public history and chosen action. Let $\nu^\pi_{E,t}(h,a) = P^\pi_E(H_t = h, A_t = a)$ be its deployment occupancy, and let $D_0 = D_{\mathrm{KL}}(P_E(O_0)\|P_M(O_0))$ be the initial observation divergence. The same legal policy is used

in both systems, so its action factors cancel in the path likelihood ratio. The entropy chain rule gives

$$D_{\mathrm{KL}}(P_E^\pi \| P_M^\pi) = D_0 + \sum_{t<T} \sum_{h,a} \nu_{E,t}^\pi(h,a) D_{\mathrm{KL}}(K_{E,t}(\cdot \mid h,a) \| K_{M,t}(\cdot \mid h,a)). \tag{18}$$

The identity holds in extended nonnegative reals; finite divergence requires model support on the real deployment paths. Pinsker gives TV at most the square root of one half of this expression.

Suppose, for every policy in Π and time t, $\nu_{E,t}^\pi(h,a) \le C_t \nu_{E,t}^\mu(h,a)$. Put $e_t = \mathbb{E}_{\nu_{E,t}^\mu} D_{\mathrm{KL}}(K_{E,t} \| K_{M,t})$. Nonnegativity of the local divergences gives the uniform certificate

$$\Delta_{\Pi,T} \le \min\left\{1, \sqrt{\frac{D_0 + \sum_{t<T} C_t e_t}{2}}\right\}. \tag{19}$$

The occupancy ratio records how much more often deployment visits a history and action than the data policy does. For a single deployment policy, its own ratios suffice. The uniform bound uses ratios valid for every policy in the class. When e_t and C_t are estimated, their statistical uncertainty enters the corresponding high-probability bound. This expresses the additional information needed to turn a prediction average into a guarantee about control.

A.8.1 An unobserved branch

Assume a deployment policy reaches a history-action event B with probability α, while the data policy assigns it probability zero. Construct two environments that have the same kernels up to B and everywhere visited by the data policy. After B, let one emit a public symbol zero and the other emit a distinct symbol one. Their complete data laws agree for any sample size. Under the deployment policy their transcript laws have TV exactly α if they otherwise agree. For any model produced from those data,

$$\max_{i \in \{0,1\}} \mathrm{TV}(P_{E_i}^\pi, P_M^\pi) \ge \alpha/2$$

by the triangle inequality. A randomized learner obeys the same lower bound in the maximum expected error because its output distribution is identical under the two data-generating environments. Coverage assumptions remove precisely this freedom to change an unseen response.

A.8.2 When an action label identifies an intervention

Suppose A_t is sampled from $\mu_t(\cdot \mid H_t)$ using randomization independent of the hidden physical state conditional on H_t. Then $A_t \perp X_{0:t} \mid H_t$. Hence, wherever $\mu_t(a \mid h) > 0$,

$$\begin{aligned} P(O_{t+1} \mid H_t = h, A_t = a) &= \sum_{x_{0:t}} P(O_{t+1} \mid x_{0:t}, h, a) P(x_{0:t} \mid h) \\ &= P(O_{t+1} \mid H_t = h, \mathrm{do}(A_t = a)). \end{aligned}$$

The second expression sets the action without changing the distribution of the preceding hidden state. If action selection also depends on an unrecorded variable U_t associated with the physical state, the observed conditional weights become $P(x_{0:t} \mid h, a)$ and can differ from $P(x_{0:t} \mid h)$. Positive action probabilities ensure that each command is observed. Identifying its physical response also requires accounting for the information used to choose it. Experiment E varies this information condition and response coverage separately.

A.9 A representation's required distinctions

Fix a decision time and two reachable histories h, h' assigned the same representation z. Let σ be a common continuation policy that starts from that representation and uses subsequent feedback. Compare the future public transcripts beginning after this decision; the differing past histories are excluded from the tested record. Let their real conditional laws be $P_h^\sigma, P_{h'}^\sigma$. A model initialized only by z predicts the same future law M_z^σ in both cases. If $\mathrm{TV}(P_h^\sigma, P_{h'}^\sigma) = \delta$, then

$$\delta \le \mathrm{TV}(P_h^\sigma, M_z^\sigma) + \mathrm{TV}(M_z^\sigma, P_{h'}^\sigma),$$

so one conditional prediction error is at least $\delta/2$. Allowing the planner to retain further information from h removes the premise that its initial state is only z; this information then belongs to the effective representation.

Define two histories to be equivalent when every common admissible continuation policy induces the same future law. The resulting quotient retains exactly the distinctions required by that policy class and interface. Predictive state representations use future tests to describe state, while bisimulation imposes recursive conditions preserving controlled behavior (Littman et al., 2002; Singh et al., 2004; Givan et al., 2003; Ferns et al., 2004). The response-map formulation applies this idea to the whole interaction and makes the policy class explicit. A state that is sufficient for a single passive policy can merge histories separated by another intervention. Conversely, a state may discard substantial visual variation while preserving every relevant controlled future.

This is an operational criterion, independent of whether the latent coordinates resemble physical coordinates. It leaves room for multiple representations of the same controlled process. It also supplies an empirical question: which intervention distinguishes two histories that a learned representation merges?

A.10 A CONSTRUCTIVE HIDDEN-STATE RESPONSE IMPLEMENTING A POSTERIOR MARGINAL

Consider a binary physical state $X \in \{0, 1\}$ with known public prior probability $b \in (0, 1)$ for state one. An observation likelihood assigns positive weights w_0, w_1. The corresponding Bayesian posterior probability is

$$r = \frac{bw_1}{(1-b)w_0 + bw_1}.$$

Suppose the actuator exposes two families of stochastic commands. Command U_u flips zero to one with probability u and leaves one unchanged. Command D_d flips one to zero with probability d and leaves zero unchanged. The mechanism's response depends on the hidden current state, as ordinary physical transitions do. Choosing its parameter uses only public quantities.

If $r \geq b$, use U_u with $u = (r - b)/(1 - b)$. Its final probability of one is $b + (1 - b)u = r$. If $r < b$, use D_d with $d = (b - r)/b$. Its final probability is $b(1 - d) = r$. In both cases the chosen parameter lies in $[0, 1]$. Thus a single legal command implements the posterior marginal without observing the realized hidden state.

A continuous-time version exposes unidirectional rates for a fixed duration τ. The upward flip probability is $1 - e^{-k\tau}$, so an interior upward target uses $k = -\tau^{-1} \log[(1 - r)/(1 - b)]$. The downward target uses $k = -\tau^{-1} \log(r/b)$. Deterministic boundary targets require unbounded rate or an additional deterministic reset. Rate bounds restrict the attainable interval and can therefore be tested using Theorem 3.

The constructed law keeps the original distribution of X_0 and changes X_1 by a legal transition. Conditioning the original static path on the evidence changes the distribution of X_0. These two full-path laws differ even though their final marginals agree. Applied to a neural sampling variable, the construction describes how a physical stochastic mechanism can implement a change in represented belief. Applied to an external object, it describes an intervention that changes the object. Specifying which variable is represented prevents an inference update from being confused with control over external uncertainty.

A.11 A TASK THAT WITNESSES AN UNATTAINABLE REQUEST

Let $\mathcal{S}_E$ be the set of response laws on a fixed finite evaluation record. With all randomized causal policies, this set is compact and convex. Compactness follows from continuity of the finite-horizon map from a product of policy simplices to path probabilities. Convexity follows by choosing between two complete policies with an independent coin before interaction. The resulting controller is legal and has the corresponding mixture law. Equivalently, its behavioral action kernels can be obtained by conditioning the private coin on its public action and observation history. Projection to a smaller evaluation record preserves both properties.

Proposition 8 (An observable witness of infeasibility)**.** *For a compact convex set $\mathcal{S}_E$ of attainable laws and any requested law q,*

$$\inf_{P\in\mathcal{S}_E} \mathrm{TV}(q,P) = \max_{0\le F\le 1} \left\{ \mathbb{E}_q F - \sup_{P\in\mathcal{S}_E} \mathbb{E}_P F \right\}. \tag{20}$$

Thus a positive attainable target error equals the largest shortfall on a bounded observable task.

Proof. For two probability vectors, $\mathrm{TV}(q,P) = \max_{F\in[0,1]^{|\mathcal{C}|}} \langle q-P, F\rangle$. The bilinear objective is continuous on two compact convex sets. Finite-dimensional minimax therefore interchanges the minimum over P and the maximum over F. For fixed F, minimizing $\langle q-P, F\rangle$ subtracts the support function of $\mathcal{S}_E$, giving (20). Compactness gives attainment. □

This expresses convex separation in the units of a control task: the maximizing F identifies an observable demand whose requested value exceeds what the environment can attain. Closure under initial randomization is the condition that makes the attainable set convex. For a restricted policy family without that property, the right side gives distance to its closed convex hull. The substitution bound and chronological projection apply directly to the original policy classes stated in those results.

A.12 Selection when a control interface expands

Fix a context and a bounded task. Let $c(a)$ be the physical cost of each registered skill and $\hat{c}(a)$ its predicted cost. In a stochastic interpretation these can be conditional expected costs; in the contact analysis they are the realized and predicted normalized costs in a fixed physical context. Define the signed prediction error $e(a) = c(a) - \hat{c}(a)$. Let a minimize predicted cost in an initial bank, and let b minimize predicted cost in an expanded bank containing it. Retain the incumbent on ties. The predicted gain is $g = \hat{c}(a) - \hat{c}(b) \ge 0$.

Proposition 9 (The error in a skill comparison)**.** *The realized gain and the predicted gain obey*

$$c(a) - c(b) = g - \big(e(b) - e(a)\big). \tag{21}$$

Consequently, an expansion worsens the selected cost exactly when its relative prediction error $e(b) - e(a)$ exceeds g. For a model minimizer $\hat{a}$ in any fixed bank,

$$c(\hat{a}) - \min_a c(a) \le \max_a e(a) - \min_a e(a) \le 2 \max_a |e(a)|. \tag{22}$$

Proof. Substituting $c = \hat{c} + e$ gives (21). For a physical minimizer $a^\star$, model optimality gives $\hat{c}(\hat{a}) \le \hat{c}(a^\star)$. Hence $c(\hat{a}) - c(a^\star) \le e(\hat{a}) - e(a^\star)$, bounded by the range of e. The second inequality bounds the range by twice its largest absolute value. □

A bias common to all skill costs cancels in the comparison. The relevant error is the change in bias across competing options. Paired interventions on the same context measure this difference while keeping target difficulty and body parameters fixed. The response discrepancy Δ controls every bounded test of the stated record. For a fixed task and skill bank, Equation (22) identifies the particular component of prediction error that can alter selection.

The identity also gives a conservative rule for accepting an additional option. Suppose valid bounds satisfy $|e(a)| \le u(a)$ for the incumbent and $|e(b)| \le u(b)$ for the candidate. Replacing a only when

$$\hat{c}(b) + u(b) \le \hat{c}(a) - u(a)$$

ensures $c(b) \le c(a)$. On any event where all declared bounds hold simultaneously, every accepted comparison has this property, including candidates chosen adaptively from that bank. The statistical requirement is a simultaneous bound over the candidate comparisons under the intended context distribution. An uncertainty estimator, including one based on ensemble disagreement, supports this rule when it provides that bound.

A.13 AN EXACT CONTINUOUS PROJECTION: DESIRED RELIABILITY UNDER ADDITIVE NOISE

The local response criterion also yields an exact continuous result. Consider one decision at a fixed observed history. An agent chooses a possibly randomized displacement $U \in \mathbb{R}^d$, after which an independent disturbance $W \sim \mathcal{N}(0, \Sigma)$ is added. The outcome is $Y = U + W$, with $\Sigma \succ 0$. There is no amplitude restriction on U. Thus every mean displacement is available. The timing of W still restricts the concentration of the outcome law. This is the conditional form of a fully actuated linear system when the current state is observed and the input matrix has full row rank.

Proposition 10 (Gaussian target under causal additive actuation)**.** *Let the requested outcome law be $q = \mathcal{N}(m, S)$ with $S \succ 0$, and write*

$$\Sigma^{-1/2} S \Sigma^{-1/2} = V \operatorname{diag}(\lambda_1, \dots, \lambda_d) V^\top, \qquad d_i = \max\{1, \lambda_i\}.$$

Among all distributions of U independent of W, a law minimizing $D_{\mathrm{KL}}(q \| \operatorname{Law}(U + W))$ is

$$p^* = \mathcal{N}\Big(m, \Sigma^{1/2} V \operatorname{diag}(d_i) V^\top \Sigma^{1/2}\Big).$$

Its exact cost is

$$\inf_{\operatorname{Law}(U)} D_{\mathrm{KL}}(q \| \operatorname{Law}(U + W)) = \frac{1}{2} \sum_{\lambda_i < 1} \big(\lambda_i - 1 - \log \lambda_i\big). \tag{23}$$

The target is attainable if and only if $S \succeq \Sigma$.

Proof. Apply the invertible change of variables $y \mapsto V^\top \Sigma^{-1/2}(y - m)$. Relative entropy is invariant under this change. The target becomes a centered Gaussian with diagonal covariance $\Lambda = \operatorname{diag}(\lambda_i)$, and the disturbance covariance becomes the identity. Put $D = \operatorname{diag}(d_i)$. The proposed response $p^* = \mathcal{N}(0, D)$ is available by selecting a Gaussian control with covariance $D - I$ and zero mean; a zero variance control coordinate is deterministic.

To prove optimality over all control laws, compare this response with each deterministic shift of the noise. Averaging those comparisons will cover every randomized control, including non-Gaussian mixtures. Let ϕ_C denote the centered Gaussian density with covariance C. Direct integration in each coordinate gives

$$\mathbb{E}_{Y \sim \mathcal{N}(0,\Lambda)} \frac{\phi_I(Y - u)}{\phi_D(Y)} = \exp\left[-\frac{1}{2} \sum_{i:\lambda_i < 1} (1 - \lambda_i) u_i^2\right] \le 1. \tag{24}$$

For a coordinate with $\lambda_i \ge 1$, $d_i = \lambda_i$ and the integral is one. For a coordinate with $\lambda_i < 1$, $d_i = 1$, and the Gaussian moment generating function gives the displayed exponential factor. This proves the equality for all $u \in \mathbb{R}^d$.

Any feasible response has density $p_\nu(y) = \int \phi_I(y - u)\nu(du)$ for a probability law ν on controls. Integrating (24) against ν, justified by nonnegativity, yields $\mathbb{E}_q[p_\nu / p^*] \le 1$. This bounds the average likelihood improvement of every mixture relative to the proposed response. Concavity of the logarithm then gives

$$\mathbb{E}_q \log(p_\nu / p^*) \le \log \mathbb{E}_q[p_\nu / p^*] \le 0.$$

Consequently $D_{\mathrm{KL}}(q \| p_\nu) \ge D_{\mathrm{KL}}(q \| p^*)$, with the inequality interpreted in the extended sense when necessary. Evaluating Gaussian relative entropy produces (23).

If $S \succeq \Sigma$, a control $U \sim \mathcal{N}(m, S - \Sigma)$ realizes the target. Conversely, if Y has the target's finite covariance, then $U = Y - W$ has finite second moments. Independence gives $\operatorname{Cov}(Y) = \operatorname{Cov}(U) + \Sigma \succeq \Sigma$. This proves the characterization. □

The eigenvalues compare requested variation with physical noise in the geometry set by Σ. Directions with $\lambda_i \ge 1$ can be produced by adding independent control variation. Directions with $\lambda_i < 1$ request a reduction of noise arriving after the decision. Their costs add. The proof covers all mixing laws, so this limit persists under arbitrary control randomization. The requested mean remains freely adjustable. Mean reachability and attainable precision therefore impose different requirements on the same actuator.

Later feedback can correct earlier disturbances when subsequent actuation permits it. At the final decision, a remaining independent disturbance imposes the same conditional restriction. Physical compliance, sensing, and the timing of correction change this statement through their effect on the response channel. The additive model applies to disturbance independent of the control at that decision. Figure 7 shows the analytic cost and numerical mixture optimization in one dimension.

A.13.1 Partial actuation and correlations between outcomes

The same projection can be solved when actions move only a subspace. Consider $Y = BU + W$, where $W \sim \mathcal{N}(0, \Sigma)$, $\Sigma \succ 0$, and U is an arbitrary randomized control independent of W. Whiten by $\Sigma^{-1/2}$ and choose orthogonal coordinates in which the image of $\Sigma^{-1/2}B$ is the first k coordinates. In these coordinates write the Gaussian target mean and covariance as

$$m = \begin{pmatrix} m_c \\ m_u \end{pmatrix}, \qquad S = \begin{pmatrix} S_{cc} & S_{cu} \\ S_{uc} & S_{uu} \end{pmatrix} \succ 0.$$

The subscripts designate controlled and uncontrolled coordinates; $k = \operatorname{rank}(B)$ and the uncontrolled dimension is $d-k$. Let D_c share the eigenvectors of S_{cc} and replace each eigenvalue λ_i by $\max(1, \lambda_i)$. Define the conditional covariance $S_{u|c} = S_{uu} - S_{uc}S_{cc}^{-1}S_{cu}$.

Corollary 11 (Gaussian target with partial actuation)**.** *The closest attainable law to the transformed target is*

$$p^* = \mathcal{N}\left(\begin{pmatrix} m_c \\ 0 \end{pmatrix}, \begin{pmatrix} D_c & 0 \\ 0 & I_{d-k} \end{pmatrix} \right).$$

The minimum relative entropy over all control laws equals

$$\frac{1}{2} \sum_{i:\lambda_i < 1} (\lambda_i - 1 - \log \lambda_i) + \frac{1}{2} \left(\|m_u\|^2 + \operatorname{tr} S_{uu} - (d-k) - \log\det S_{u|c} \right). \tag{25}$$

For $k = 0$ or $k = d$, empty blocks and sums are omitted.

Proof. After whitening and rotating, every attainable density factors as $p(y_c, y_u) = p_c(y_c)\phi_I(y_u)$, where p_c is a mixture of translated unit Gaussians. The uncontrolled noise is independent of both the controlled noise and the chosen control. The relative entropy chain rule gives

$$D_{\mathrm{KL}}(q\|p) = D_{\mathrm{KL}}(q_c\|p_c) + \mathbb{E}_{q_c} D_{\mathrm{KL}}(q(\cdot \mid Y_c)\|\phi_I).$$

Proposition 10 minimizes the first term over all mixtures and yields the first sum. The conditional target in the second term has covariance $S_{u|c}$ and mean $m_u + S_{uc}S_{cc}^{-1}(Y_c - m_c)$. Averaging its Gaussian relative entropy yields the second term because

$$\operatorname{tr} S_{u|c} + \mathbb{E}\|m_u + S_{uc}S_{cc}^{-1}(Y_c - m_c)\|^2 = \operatorname{tr} S_{uu} + \|m_u\|^2.$$

The proposed p^* is attainable by a Gaussian control on the controlled subspace. This proves global optimality, including against non-Gaussian controls. □

A target can therefore fail through its requested correlations even when its uncontrolled marginal matches the disturbance. With $m_u = 0$, $S_{uu} = I$ and $S_{cu} \neq 0$, the second term is positive. Correlating a controlled coordinate with an independent disturbance that arrives afterward would require access to that disturbance at the time of action. For example, in two dimensions with one controlled coordinate and target covariance $\left(\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}\right)$, both marginal targets are individually attainable but the joint projection cost is $-\frac{1}{2}\log(1-\rho^2)$. This gives a continuous instance of the difference between matching marginal outcomes and matching their required dependence.

A.13.2 Sensing before the action changes the precision limit

Let the uncontrolled initial displacement V and sensor reading S be jointly Gaussian. The agent observes S, then chooses a fully actuated displacement U, and finally receives independent disturbance $W \sim \mathcal{N}(0, \Sigma_W)$, with $\Sigma_W \succ 0$. The outcome is $Y = V + U + W$. The action may use S and independent private randomness. A Gaussian conditional decomposition gives

$$V = \mathbb{E}[V \mid S] + R, \qquad R \perp S, \qquad \operatorname{Cov}(R) = \Sigma_{V|S}.$$

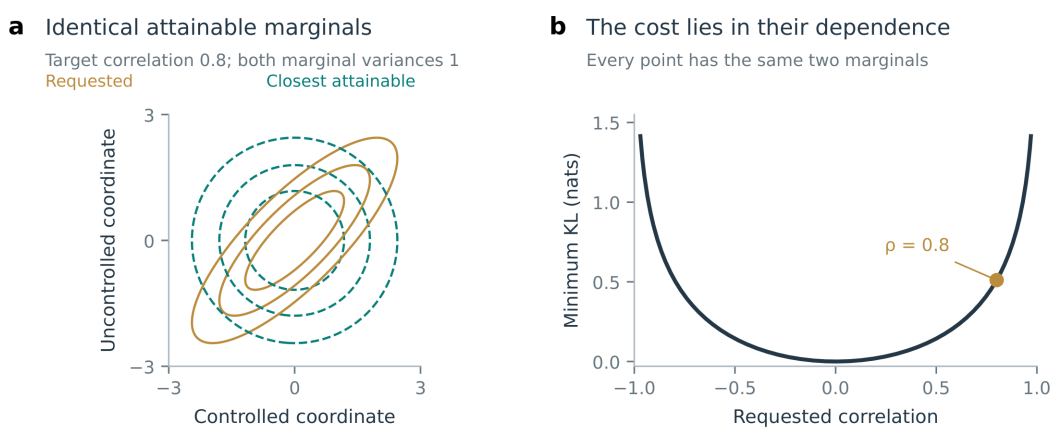


Figure 5: **Individually feasible marginals can require an infeasible dependence.** a, Target correlation 0.8 and its closest attainable joint law when only the first coordinate can be actuated before independent unit Gaussian noise. Contours enclose 50%, 80%, and 95% probability. Both marginal targets are unit Gaussians. b, The exact minimum KL as requested correlation changes, from Corollary 11. The highlighted target costs 0.5108 nats. These are analytic distributions.

Consequently every attainable outcome law is a convolution of an arbitrary control law with a Gaussian of covariance

$$\Sigma_{\text{remain}} = \Sigma_{V|S} + \Sigma_W. \tag{26}$$

To see both inclusions, write $Y = (U + \mathbb{E}[V \mid S]) + (R + W)$. The two terms are independent. Conversely, for any independently chosen F, the legal action $U = F - \mathbb{E}[V \mid S]$ gives $Y = F + R + W$. Proposition 10 therefore applies with Σ_{remain} as the disturbance covariance. A Gaussian target covariance Σ_q is attainable exactly when $\Sigma_q \succeq \Sigma_{\text{remain}}$.

If the same sensor reading arrives after the only action, the corresponding floor is $\mathrm{Cov}(V) + \Sigma_W$. The sensor may then explain the realized error, but its information cannot change that decision. The comparison holds for an outcome target, known joint Gaussian statistics, unrestricted displacement, and a single corrective action. These assumptions make the timing effect explicit: early sensing removes the predictable component of the initial displacement, and the later independent disturbance remains. Actuation limits and sensing costs can be included through their corresponding response sets.

A.13.3 Independent numerical calculations

Numerical calculation. The independent convex optimization in Figure 7 uses 201 equally spaced control displacements from -10 to 10, unit noise variance, and target variance ratios $0.125, 0.2, 0.25, 0.4, 0.5, 0.7, 0.85, 1, 1.5, 2, 3, 4$. A 96-node Gauss–Hermite rule evaluates the target expectation. SLSQP minimizes the resulting convex objective over nonnegative weights summing to one, with analytic gradient, tolerance 10^{-11}, and at most 600 iterations. The largest difference from the analytic optimum is 2.20×10^{-11}. A separate quadrature calculation of (24) uses seven variance ratios and 31 shifts, with maximum residual 1.78×10^{-15}. The script also compares primal and dual TV linear programs in 150 randomly generated finite systems, with maximum residual 2.55×10^{-14}, and evaluates the relative-error identity in 10,000 random finite decisions. These calculations supplement the proofs; their finite grids do not replace the optimization over all laws in Proposition 10.

The partial-actuation formula is evaluated independently in 256 Gaussian systems with dimensions from two to eight and actuation rank from one to $d - 1$. For each system, direct Gaussian KL is compared with Equation (25), and 64 alternative Gaussian controls test the proposed minimum. The largest identity residual is below 10^{-12}. The proof supplies the guarantee over non-Gaussian controls. The script `code/check_gaussian_channels.py` records the matrices' dimensions, component costs, identity residuals, and alternative-control gaps in `results/theory_extensions/`. Figure 5 and Figure 6 are analytic evaluations of the same results.

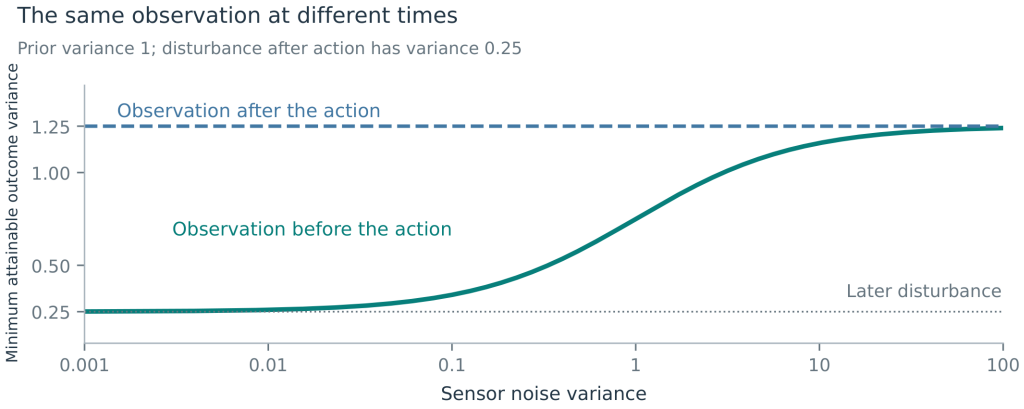


Figure 6: **An observation changes achievable precision through its timing.** The initial displacement has variance one. A noisy Gaussian reading is available before or after the only action, which is followed by independent noise of variance 0.25. Early sensing permits cancellation of the conditional mean. The two curves are the exact remaining variance in Equation (26); their difference measures the benefit of observing before acting.

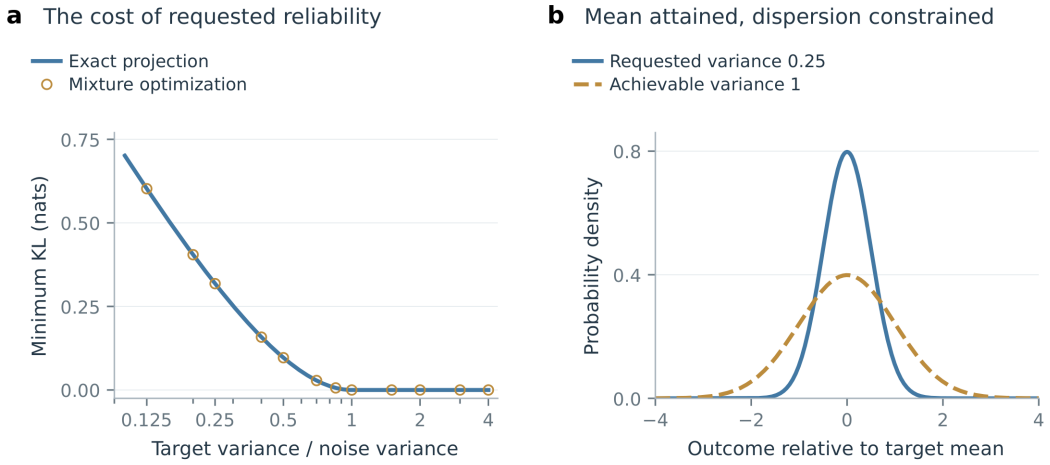


Figure 7: **Achievable reliability with unrestricted mean actuation.** a, Exact relative entropy projection cost for a scalar Gaussian target, as a function of target variance divided by post-action noise variance. The solid curve is Proposition 10; circles are a separate nonnegative mixture optimization on a discretized action grid. b, Requested and closest achievable outcome densities when target variance is one-quarter of noise variance. The mean is attained exactly. The remaining discrepancy concerns dispersion. These panels are analytic and numerical calculations with unit noise, without sampled physical trials or uncertainty intervals.

A.14 Analytic constructions in Figure 1

Figure 1 connects three questions: which command produces an outcome, which joint outcome distributions are attainable, and when a sensor can improve their precision. The examples below follow directly from the response map and Gaussian projection. Figure 8 shows the corresponding joint distributions, projection cost surface, and sensor-quality curve. Every plotted quantity is obtained from the analytic laws specified here.

A common attainable set with different command meanings. Let $D = \{u \in \mathbb{R}^2 : \|u\| \le 1\}$ be the action set, and let $W \sim \mathcal{N}(0, \sigma^2 I_2)$ be independent of the chosen control, with $\sigma = 0.13$.

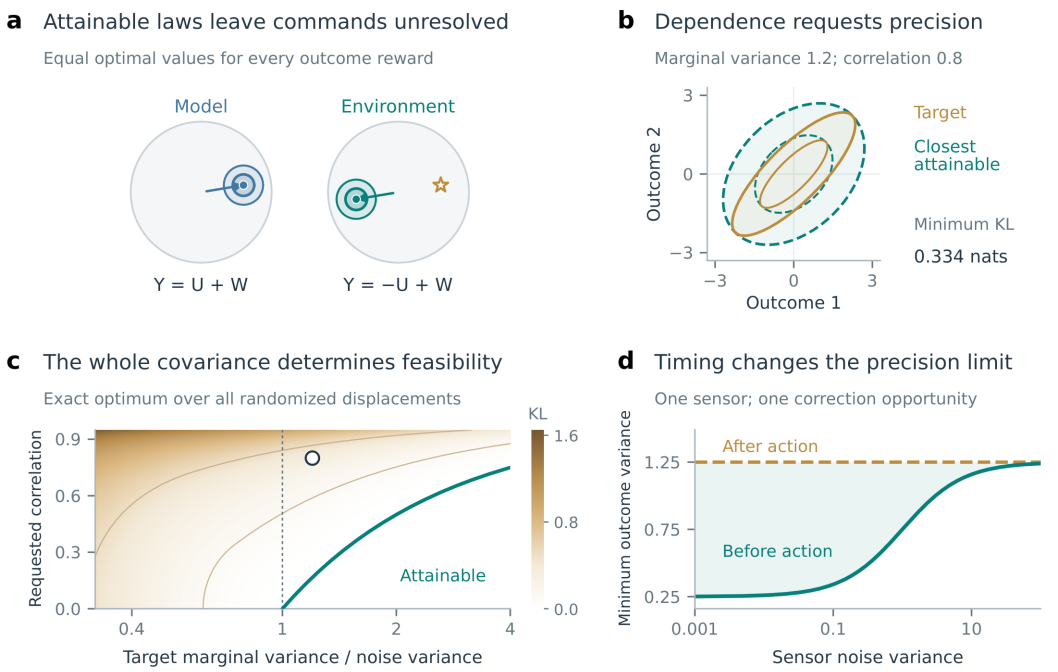


Figure 8: **Command correspondence and distributional limits.** a, Equal sets of attainable outcome laws can assign different consequences to the same command. b, A correlated Gaussian target and its closest attainable law, shown by 50% and 90% probability contours. c, Exact projection cost across target marginal variances and correlations; the marked target is the one in b. d, The minimum outcome variance when the same sensor reading arrives before or after corrective action. All quantities are analytic.

Consider the two one-decision systems

$$Y_M = U + W, \qquad Y_E = -U + W, \qquad U \in D.$$

For every randomized model control ν, its reflection, the law of $-U$ for $U \sim \nu$, is a legal environment control producing the same outcome distribution. Reflection is a bijection of probability laws on D. The two systems therefore have identical sets of attainable outcome laws, including all mixtures of their Gaussian responses. It follows that

$$\sup_{\nu:\,\nu(D)=1} \mathbb{E}_{M,\nu} F(Y) = \sup_{\nu:\,\nu(D)=1} \mathbb{E}_{E,\nu} F(Y)$$

for every bounded measurable outcome reward F. Under the *same* command $U = u$, the means are u and $-u$. Figure 1a uses $u = (0.6, 0.1)$ and displays the first-coordinate response densities, with means 0.6 and -0.6 and common standard deviation 0.13. Figure 8a shows the 50% and 90% joint probability contours; its star marks the model's predicted mean. The background discs are sets of available means, with unit radius; the Gaussian outcome distributions extend beyond them. A translator sending u to $-u$ restores the correspondence. This continuous construction illustrates the independently optimized value characterization in Appendix A.3.

Joint precision with individually attainable marginals. For Figure 1b and Figure 8b,c, let $U \in \mathbb{R}^2$ be unrestricted and chosen before $W \sim \mathcal{N}(0, I_2)$. The requested law is centered Gaussian with covariance

$$S(r, \rho) = r \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \qquad r > 0, \quad |\rho| < 1.$$

Its principal variances are $\lambda_- = r(1 - |\rho|)$ and $\lambda_+ = r(1 + |\rho|)$. For positive correlation, the narrower direction measures the difference between the two coordinates. A narrow difference distribution requires the coordinates to move together more precisely than their separate variances

reveal. Proposition 10 gives the exact projection onto all responses $\mathrm{Law}(U+W)$:

$$C(r,\rho) = \frac{1}{2} \sum_{s\in\{-,+\}:\lambda_s<1} (\lambda_s - 1 - \log\lambda_s).$$

The target is attainable precisely when $r(1-|\rho|) \geq 1$. Each marginal alone is attainable when $r \geq 1$. Consequently the region

$$1 \leq r < \frac{1}{1-|\rho|}$$

contains targets with feasible marginal distributions and infeasible joint distributions. Correlation specifies additional precision along one principal direction.

The marked target takes $r = 1.2$ and $\rho = 0.8$, giving target variances 0.24 and 2.16 in the principal directions. Figure 1b shows the 90% probability contours of the requested law and closest response. Figure 8b also includes their 50% contours. The closest response has principal variances 1 and 2.16, or covariance

$$S^* = \begin{pmatrix} 1.58 & 0.58 \\ 0.58 & 1.58 \end{pmatrix}.$$

The exact minimum is $C(1.2, 0.8) = 0.3335581778$ nats. It is achieved by a centered Gaussian control with covariance $S^* - I_2$, which is positive semidefinite and has rank one. Every non-Gaussian control obeys the same lower bound. A contour enclosing probability p is parameterized by

$$x(\theta) = \sqrt{-2\log(1-p)}\; S^{1/2} \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix}, \qquad 0 \leq \theta \leq 2\pi,$$

using S^* for the attainable contour. The supplementary panels retain 361 points per contour; the main figure uses 721. Figure 8c evaluates C on 301 logarithmically spaced values of $r \in [0.32, 4]$ and 241 uniformly spaced values of $\rho \in [0, 0.95]$. Its solid boundary is $\rho = 1 - 1/r$ for $r \geq 1$; the dashed line marks the marginal threshold $r = 1$. The thin contours have costs 0.1, 0.5, and 1 nat. The main figure displays the requested and attainable joint distributions at the marked target. All 72,541 grid values accompany the supplementary surface. A separate 97×89 grid compares the spectral expression with direct Gaussian KL, with maximum absolute discrepancy below 10^{-12}.

The same observation channel on either side of an action. In Figure 1c and Figure 8d, the initial displacement is $V \sim \mathcal{N}(0,1)$, the sensor reads $S = V + N$ with independent $N \sim \mathcal{N}(0,\nu)$, and the outcome is $Y = V + U + W$ with independent $W \sim \mathcal{N}(0, 1/4)$. If the reading precedes the action, choose $U = -\mathbb{E}[V \mid S]$. The minimum outcome variance over all randomized corrective actions is

$$v_{\mathrm{before}}(\nu) = \frac{1}{4} + \frac{\nu}{1+\nu}.$$

If the reading arrives after the only action, U has no information about V and the minimum is

$$v_{\mathrm{after}}(\nu) = \frac{5}{4}.$$

These are instances of the conditional Gaussian decomposition in Appendix A.13.2. Figure 1c fixes $\nu = 1/4$ and displays the centered Gaussian outcome densities attaining the two minima: $v_{\mathrm{before}} = 0.45$ and $v_{\mathrm{after}} = 1.25$. Its event sequences show which reading is available at the corrective action. Figure 8d includes 301 logarithmically spaced sensor variances from 0.001 to 100; shading shows the difference between the two exact floors. The main density curves use 1,001 evenly spaced outcome coordinates in $[-3.25, 3.25]$; the command-response curves use 801 coordinates in $[-1.15, 1.15]$. Sensing quality and noise after correction are held fixed in each timing comparison. The change in achievable precision comes from which observation the corrective action can use.

B Continuous contact experiments

B.1 Continuous contact dynamics and measurement

The continuous studies examine three relations in one task: how target information reaches an action, how a model learns the consequences of its available skills, and how it selects among those skills. A

circular pusher moves a rectangular object toward a target. Contact can translate or rotate the object, move to another face, and separate. The consequences of a command depend on shape, mass, friction, pose, and actuator response. Each sampled physical context is reused across the conditions being compared, allowing these relations to be varied separately.

Study H supplies an exact dynamics model and changes the schedule through which a new target cue affects commands. Study I keeps feedback available during execution, hides physical parameters from the learned model, and changes which skill responses enter training. The expansion analysis then holds those models and recorded responses fixed while adding choices. The simulator, controller, and measurement definitions below are shared by these studies.

B.1.1 State, actions, and the observed body

The ten state coordinates are the pusher position and velocity $(p, v_p) \in \mathbb{R}^4$, object position and velocity $(x, v) \in \mathbb{R}^4$, orientation θ, and angular velocity ω. A body has six episode parameters: mass m, ground friction μ, half-width a, half-height b, command rotation ϕ, and actuator gain g. The rectangle's planar moment of inertia is $I = m(a^2 + b^2)/3$. Parameters remain fixed within an episode. Table 1 gives all distributions. Draws are independent unless a paired design explicitly reuses a context.

Table 1: **Continuous simulator and body distributions.** Distances, masses, times, and forces use SI units. "Original" denotes the distribution used to collect training responses. All rows apply to H and I except the shifted-body column, used in I.

Quantity	Original setting	Shifted bodies
Mass m	$U(0.18, 0.36)$ kg	$U(0.36, 0.50)$ kg
Ground friction μ	$U(0.10, 0.30)$	$U(0.30, 0.45)$
Half-width a	$U(0.045, 0.065)$ m	Same distribution
Half-height b	$U(0.030, 0.045)$ m	Same distribution
Command rotation ϕ	$U(-0.25, 0.25)$ rad	Random sign $\times U(0.25, 0.45)$
Actuator gain g	$U(0.85, 1.15)$	Same distribution
Object center x_0	$U(-0.015, 0.015)^2$ m	Same distribution
Object angle θ_0	$U(-0.45, 0.45)$ rad	Same distribution
Initial pusher relative to object	$(-a - 0.045,\ U(-0.025, 0.025))$ m	Same distribution
Initial velocities	Zero	Zero
Control interval	0.025 s, 40 Hz	Same
Physics substeps per control	40	Same
Pusher radius	0.018 m	Same
Nominal servo speed	0.30 m/s	Same
Servo time constant	0.04 s	Same
Normal stiffness	450 N/m	Same
Normal damping	6 N s/m	Same
Contact friction	0.35	Same
Gravity	9.81 m/s^2	Same

The command $u \in [-1, 1]^2$ is clipped coordinatewise and then divided by $\max\{1, \|u\|\}$. The desired pusher velocity is $0.30gR(\phi)u$, where R is a planar rotation. At each physics substep of duration $\delta = 0.025/40$ s, the servo updates its velocity by forward relaxation toward this desired velocity with time constant 0.04 s, then integrates its position. The Cartesian servo imposes pusher motion; its prescribed velocity determines the contact forces acting on the object.

In I, the model observes eleven instantaneous features: $p - x$, v_p, v, $\sin\theta$, $\cos\theta$, ω, a, and b. A calibration history contributes two more features. Mass, ground friction, command rotation, and gain are withheld as inputs. Public geometry therefore specifies shape and kinematics while leaving physical response partially uncertain. H grants its physical twin the actual episode parameters to isolate the information restriction from estimation error.

B.1.2 Contact and ground friction

We transform $p - x$ into the rectangle frame and clamp it to $[-a, a] \times [-b, b]$ to obtain the nearest contact point. Outside the rectangle, the difference from that point gives the outward unit normal

and distance. Inside, the nearest face defines the outward normal and a negative signed distance. Ties use the first coordinate. Let n denote the world-frame outward normal, r the contact-point lever arm about the object center, and d the signed distance. Contact penetration is $(0.018 - d)_+$. With contact-point relative normal velocity

$$v_n = \big(v_p - v - \omega(-r_y, r_x)\big) \cdot n,$$

the normal force magnitude is zero outside contact and

$$f_n = \big[450(0.018 - d) - 6v_n\big]_+$$

inside contact. Force $-f_n n$ and its moment update object linear and angular velocities.

After the normal update, let $t = (-n_y, n_x)$ and let v_t be the updated relative tangential velocity. The bounded tangential impulse is

$$j_t = \text{clip}\left(\frac{v_t}{m^{-1} + (r \times t)^2/I},\ -0.35 f_n \delta,\ 0.35 f_n \delta\right).$$

The object receives $j_t t$ and angular impulse $(r \times t) j_t$. This update uses the effective mass at the contact point and bounds friction by the normal impulse.

Ground friction reduces linear speed by $\mu\, 9.81\delta$, stopping at zero. Torsional friction reduces $|\omega|$ by $\mu m\, 9.81\, (0.45\sqrt{a^2 + b^2})\delta/I$, also stopping at zero. These stopping updates avoid reversing motion at very small velocities. The object position and orientation are integrated from the updated velocities; angles are wrapped to $[-\pi, \pi)$. Peak contact force is the maximum normal force across physics substeps and then across the skill. Tangential and ground forces are not included in this peak.

B.1.3 The shared feedback skill

The controller first steers the pusher behind the object and then pushes toward a specified goal. Its arguments are goal displacement, speed, and a steering offset. Let e be the unit vector from object center to the goal, rotated by the steering offset, and $e_\perp = (-e_y, e_x)$. The pusher's relative forward and lateral positions are $(p - x) \cdot e$ and $(p - x) \cdot e_\perp$. The rotated rectangle support along e is

$$s = a\left|(R(\theta)^\top e)_1\right| + b\left|(R(\theta)^\top e)_2\right|.$$

A rear waypoint is $x - (s + 0.018 + 0.012)e$. The approach command points to this waypoint, normalized with a 0.02 m denominator floor. The pushing command is $e - (\ell/0.045)e_\perp$, where ℓ is lateral displacement.

Two smooth weights interpolate between these commands. The lateral weight is the smoothstep $z^2(3 - 2z)$ applied to $z = \text{clip}((|\ell| - 0.007)/0.018, 0, 1)$. The forward weight applies the same smoothstep to $\text{clip}(((p - x) \cdot e + 0.9s)/(0.5s), 0, 1)$. The final approach weight is one minus the product of the complements. The command is scaled by the skill's speed and by $\text{clip}((\|\text{goal} - x\| - 0.006)/0.02, 0, 1)$ near the target. H additionally compensates for the known command rotation. I uses the uncorrected public controller for every collection rule and test skill.

The controller and simulator are separate functions. All comparisons within a study use the same functions and constants. H changes when a generated sequence receives a cue; I changes which response labels train its predictor.

B.2 H: information timing and cue correspondence

A target cue can arrive after a push has begun. Its value for control depends on whether it can influence commands before the opportunity for correction has passed. To study that dependence, every condition receives the same exact physical model and steering rule. At a decision boundary the model generates a sequence of commands; the physical system follows it until the next boundary. A cue received during that interval first affects the following sequence. We vary the boundary spacing and cue time while pairing the two target branches within each physical context. A second comparison shuffles the cue assignments, preserving their aggregate frequency and arrival time while changing their correspondence with the requested target.

B.2.1 DESIGN AND COMPLETE EXECUTION RECORD

We sample 192 original-distribution physical contexts with seed 81030 and pair each with target branches $B = -1, +1$. Relative to the initial object center, the selected target is $(0.20, 0.08B)$ m. Before the cue, the goal supplied to the controller is $(0.16, 0)$ m. An episode lasts 120 public steps, or 3 s. Success means final object-center distance below 0.04 m from the branch's target. The two success disks are disjoint.

Cue times, in public steps, are $0, 11, 19, 27, 35, 43, 51, 59, 67, 75, 83$. Chunk lengths are $1, 4, 8, 16, 32, 120$. At each chunk boundary, the physical twin resets to the observed state and latches the most recent cue. It generates actions using its predicted states and this latched goal. The physical system executes the same actions. The cue can change the next sequence only at a boundary. This procedure produces a planned sequence whose commands vary with predicted contact; it does not repeat one command throughout the chunk. When a cue arrives exactly at a boundary it is available for that boundary's decision.

Conditional on a body and executed commands, the physical dynamics are deterministic and shared by twin and environment. Therefore the state predictions remain exact, even when a chunk retains an obsolete goal. With the cue initially available, all unshuffled chunk lengths agree at every recorded state and action; the maximum state difference is zero in the saved arrays. The measured difference is therefore caused by the schedule through which subsequent observations affect actions.

For the correspondence intervention, seed 81031 assigns a balanced permutation of cue values across the 192 contexts, independently of their true branches. Both branches of a context receive the same assigned cue. Thus the assigned cue has the same aggregate frequency and timing as the true cue, and exactly one of the two branch episodes in each context receives the correct assignment. The physical controller uses that assigned cue with one-step feedback. All seven conditions share initial states and parameters. The full sweep has $192 \times 2 \times 11 \times 7 = 29{,}568$ episodes. These are repeated measurements on 192 independent physical contexts. Complete state, action, and force trajectories are saved at cue steps 0, 35, and 67; all conditions retain their episode outcomes.

B.2.2 CONDITIONAL AND MARGINAL MEASUREMENTS

We classify an endpoint into $O = -1$ when it is inside the left target disk, $O = +1$ inside the right disk, and $O = 0$ otherwise. The desired endpoint marginal has masses $(1/2, 0, 1/2)$. The desired joint law assigns probability $1/2$ to each of $(B, O) = (-1, -1)$ and $(+1, +1)$. Since the empirical branch marginal is exactly balanced, the TV distance to this joint target is exactly one minus correct-target success. The marginal TV retains neither the cue nor its correspondence with the outcome.

At step 35, the feedback condition has 82.55% success and marginal/joint TV 0.1745. Shuffling yields 40.89% success, marginal TV 0.1823, and joint TV 0.5911. At step 67, feedback and shuffling have success 50.78% and 25.26%, respectively. Their marginal TVs remain close, 0.4896 and 0.4948. At both cue times, the endpoint marginal answers how often the object reaches each region. The joint law also records whether the reached region is the one requested in that episode. This additional relation accounts for the larger discrepancy under shuffled cues.

The 16-step condition succeeds in 70.05% of step-35 episodes; 32-step chunks succeed in 53.39%. Four-step chunks succeed in 85.68%, exceeding the one-step controller in this particular timing condition. Changing timing changes the contact configuration in which a fixed steering rule acts. A larger class of information-dependent policies can always ignore additional observations, but the fixed controller evaluated here does not solve the optimum over that class. The numerical curves consequently need not be monotone.

Figure 10 relates the timing surface to the recorded commands and contact forces. For a cue revealed at step r and a commitment length k, the next permitted command update is at step $k\lceil r/k \rceil$. The 66 correct-cue conditions therefore produce 33 distinct first-update times. In these recorded executions, conditions with the same first-update time have identical mean success. The controller, initial states, and exact physical twin are shared across conditions. The first permitted update determines which contact configuration is available when the cue first changes the command sequence. This explains why re-indexing the timing sweep collapses conditions that looked different when described only by

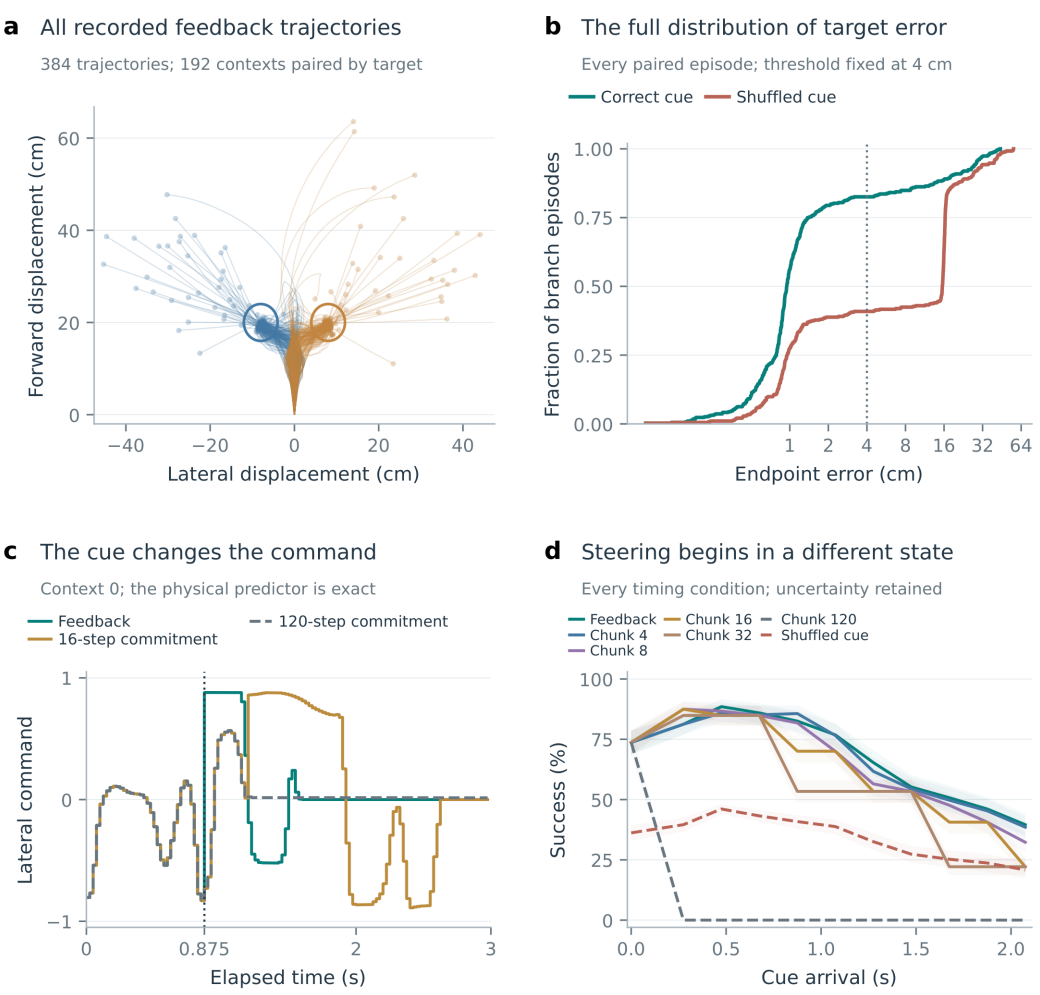


Figure 9: **Feedback timing across the physical population.** a, All 384 recorded feedback trajectories at a cue time of 0.875 s, with endpoint marks and target disks. Both branches of every context are retained, including the longest failure trajectories. b, Complete endpoint-error distributions for correct and shuffled cues. The vertical reference is the 4 cm success tolerance. c, Recorded normalized lateral commands in the first context, comparing feedback with two commitment lengths; the vertical reference marks cue arrival. d, Every timing condition, with 95% intervals from paired resampling of 192 contexts. Timing changes the contact state in which the fixed steering rule acts.

cue arrival and chunk length. The body-parameter display locates every paired cue effect within the full physical population.

B.2.3 The information restriction behind the comparison

The following specialization of the causal restriction clarifies what the intervention tests. Let B be independent of the initial physical state, private policy randomization, and physical transition noise. Suppose the public information used by every action that can affect the terminal outcome is independent of B. This condition includes a completely committed sequence chosen before an independently drawn cue. Let G_b be pairwise disjoint success regions. The terminal outcome X is then independent of B, so

$$\mathbb{P}(X \in G_B) = \sum_b \mathbb{P}(B = b)\mathbb{P}(X \in G_b) \le \max_b \mathbb{P}(B = b) \sum_b \mathbb{P}(X \in G_b) \le \max_b \mathbb{P}(B = b).$$

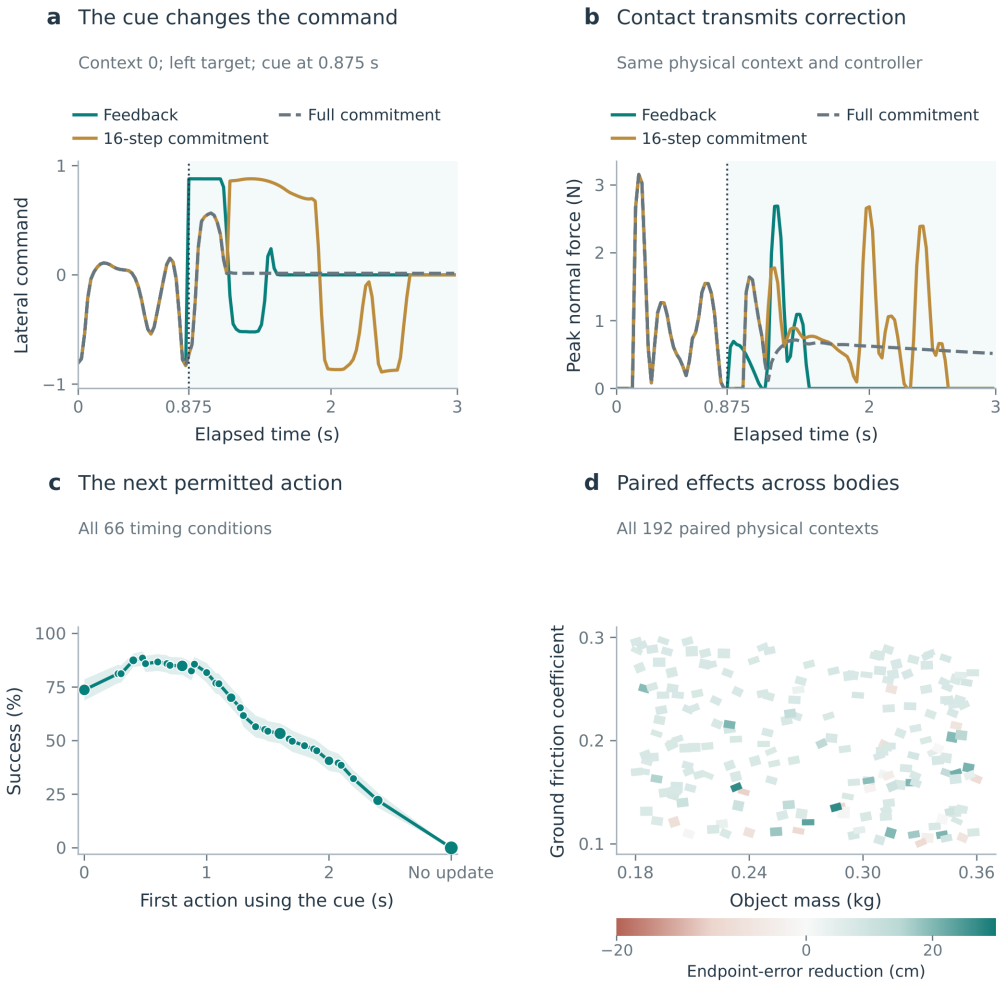


Figure 10: **A cue changes control through the next permitted action.** a,b, Recorded lateral commands and contact forces for held-out context zero and the left target. The shaded period begins at cue arrival, 0.875 s. The three conditions use the same exact physical twin. c, All 66 correct-cue timing conditions, indexed by the first permitted command update after the cue. An update at the episode boundary is labeled No update. Marker area records the number of coincident conditions. Bands retain the original paired 95% context intervals; each group has identical success in these executions. d, All 192 physical contexts. Each glyph's aspect and orientation describe its initial body geometry. Color shows the reduction in endpoint error from correct cues relative to shuffled cues, averaging both target branches within a context. The parameter map is descriptive. All panels reuse recorded executions.

For two balanced branches the upper bound is one-half. Consequently TV to the desired joint cue–outcome law is at least one-half. Additional failures of reaching the disks can increase the discrepancy. If both disks are attainable, a cue-independent randomized controller can reproduce their equal mixture while attaining only one-half correct correspondence. This explains why a correct marginal provides insufficient evidence for the joint target in this design.

The restriction applies to all policies with the stated information access, independently of the accuracy or capacity of their physical predictor. For intermediate cue times, the declared controller acts from the contact state reached at its next permitted update. Figure 9 measures these outcomes directly.

B.3 I: LEARNING AND SELECTING RESPONSES OF FEEDBACK SKILLS

The planner now chooses a complete pushing skill using its predicted result. A skill runs for two seconds and continues responding to public observations throughout execution. The learned model predicts its final displacement, rotation, and peak force. Thus the model evaluates the consequence of a feedback rule whose internal actions depend on the motion it encounters.

We compare four ways of collecting additional response labels, starting from the same initial data. Architecture, optimizer updates, and test contexts remain fixed. The comparison asks whether data concentrated on familiar pushes or data covering deployment choices better support selection. Two independent validation collections then examine how the distribution used to measure prediction error affects which trained model is chosen for deployment.

B.3.1 PUBLIC CALIBRATION AND MODEL INPUTS

Each context first executes twelve calibration steps: four with command $(0.4, 0)$, four with its negative, and four with zero command. The pusher velocity measured after the first four steps, divided by 0.30 m/s, supplies two public history features. The resulting physical state is the initial state of the skill. This short history can reveal properties of the actuator without supplying its parameters as labels. The response model receives the eleven instantaneous features described above, these two history features, requested goal displacement (two coordinates), speed, and steering offset: seventeen inputs in total.

A skill then executes the shared feedback controller for 80 steps, or 2 s. Its five response labels are object displacement divided by 0.25 m, $\sin(\Delta\theta)$, $\cos(\Delta\theta) - 1$, and peak normal force divided by 5 N. The learning objective is mean squared error over these five normalized coordinates. These are temporally abstract physical responses to a specified feedback policy. The three declared costs are functions of these five response coordinates, so their predictions determine the planner's ranking of the skills. Intermediate observations affect execution through the feedback controller.

B.3.2 COLLECTION DISTRIBUTIONS AND ACQUISITION BUDGETS

All methods begin with 1,500 routine responses, seed 90201. A requested displacement is $d(\cos\alpha, \sin\alpha)$; Table 2 specifies the collection distributions. All response bodies are drawn from the original parameter ranges in Table 1. The policy-probe distribution uses knowledge of the intended skill class, including its direction range. It does not use held-out physical outcomes.

Table 2: **Distributions for collecting skill responses.** Each listed variable is sampled uniformly and independently. Ensemble disagreement acquires labels from the broad pool according to the variance of its initial models.

Collection	Direction α (rad)	Distance d (m)	Speed	Steering (rad)
Routine	$[-0.18, 0.18]$	$[0.16, 0.24]$	$[0.60, 1.00]$	$[-0.12, 0.12]$
Policy probes	$[-1.15, 1.15]$	$[0.12, 0.30]$	$[0.45, 1.00]$	$[-0.32, 0.32]$
Broad pool	$[-\pi, \pi]$	$[0.08, 0.36]$	$[0.20, 1.00]$	$[-0.70, 0.70]$

The routine and policy-probe candidate pools each contain 10,500 responses, with seeds 90211 and 90221. The broad pool contains 21,000 candidates, seed 90231. Uniform collection takes a random-order prefix of this pool. Ensemble disagreement trains three initial models on the common 1,500 responses, predicts all broad-pool candidates, computes sample variance over the three predictions separately for each normalized output, and averages those variances over the five outputs. Candidates are ranked by decreasing variance with stable tie-breaking. All requested labels are acquired in one batch from this initial ranking. The initial ensemble therefore determines the entire acquisition order.

For total budgets $3{,}000$, $6{,}000$, and $12{,}000$, each rule adds respectively $1{,}500$, $4{,}500$, and $10{,}500$ responses to the shared data. These budgets count labelled skill outcomes. The simulator caches candidate outcomes before evaluation, but the acquisition function reads only public inputs and ensemble predictions. Held-out candidate labels do not enter its ranking. Public candidate contexts, their calibration features, and the two validation collections are available as shared experimental resources. Their collection costs are additional to the labelled-skill budget. The budget comparison measures allocation of response labels under this common candidate-inspection protocol.

Two independent collections, each containing 2,000 labelled responses, evaluate predictions: routine validation uses seed 90241 and policy-probe validation uses seed 90251. They are separate from the per-model fitting split. In the model-selection comparison, each rule uses the corresponding validation set to choose among the same four trained collection candidates at a given budget and initialization seed. Test contexts and test outcomes are never used in this choice.

B.3.3 Architecture and fitting

The response network is a fully connected $17 \to 128 \to 128 \to 128 \to 5$ network with SiLU activations between linear layers, containing 35,973 trainable parameters. Input means and standard deviations are fitted on each model's training split, with standard deviations floored at 0.02. Outputs use the fixed physical normalization above. A seeded permutation reserves 10% of the labelled responses, with a minimum of 100, for internal validation. The rest trains the model.

Each model receives 2,200 AdamW updates, batch size 256, learning rate 10^{-3}, weight decay 10^{-5}, and gradient-norm clipping at 5. Internal validation is evaluated every 100 updates; the checkpoint with lowest internal validation MSE is retained. Batch indices are sampled with replacement from the training split. Every collection condition uses initialization seeds 11, 23, and 47. The common initial ensemble has three models; the four rules, three subsequent budgets, and three seeds add 36, giving 39 trained models. All checkpoints, recorded training and validation measurements, input normalizations, and response predictions are included.

B.3.4 Registered skills, tasks, and held-out evaluation

We draw 256 test contexts from the original body distribution with seed 90500, and 128 from the shifted distribution with seed 90520. A test target has direction uniform in $[-1.05, 1.05]$ rad and distance uniform in $[0.16, 0.24]$ m, relative to the object after calibration. Each context registers the 27 combinations of requested distance multiplier $\{0.8, 1.0, 1.2\}$, steering offset $\{-0.25, 0, 0.25\}$, and speed $\{0.5, 0.75, 1.0\}$. The predictor sees each skill's arguments but not its realized outcome.

Let d_f be final target distance, ψ the absolute wrapped object-angle change, and F the peak normal force. All three realized costs lie in $[0, 1]$:

$$\begin{aligned} c_{\text{position}} &= \min\{d_f/0.30, 1\}, \\ c_{\text{pose}} &= 0.7c_{\text{position}} + 0.3\min\{\psi/(\pi/2), 1\}, \\ c_{\text{force}} &= 0.75c_{\text{position}} + 0.25\min\{F/5, 1\}. \end{aligned}$$

The model evaluates the same expressions from its predicted response. Angle change is recovered with $\operatorname{atan2}(\widehat{\sin\Delta\theta}, 1 + \widehat{\cos\Delta\theta} - 1)$. Force predictions are clipped to the unit interval in the normalized force term. The selected skill minimizes the predicted cost, with the lowest registry index breaking ties. Its actual simulated outcome provides the deployment measurement. Reported success uses the position objective and requires $d_f < 0.04$ m. Position, pose, and force costs are reported separately so that a change in orientation or contact force remains visible even when the position tolerance is met.

Each of the 384 held-out contexts supplies 27 executed skills, giving $384 \times 27 = 10{,}368$ distinct physical responses. The same bank is evaluated by 39 trained models under three objectives, giving $39 \times 384 \times 3 = 44{,}928$ model-based selections. The physical context is the unit shared by these comparisons. A model, an objective, or an expansion of the skill bank creates another selection from that context's recorded responses.

For each context and objective, the least measured cost among the 27 responses gives the best recorded response. This reference answers how much performance is available in that particular bank, using the outcomes already observed in the simulator. Comparing selected cost with it isolates the loss associated with choosing a skill. The reference includes the effects of the hidden physical parameters through their realized outcomes. The theorem's optimum is defined separately over legal policies with the stated information access.

B.3.5 Population results and their interpretation

At 6,000 labels, mean success on original bodies is 46.09% for routine collection, 57.68% for uniform collection, 45.96% for disagreement, and 63.15% for policy probes. All means average over the same

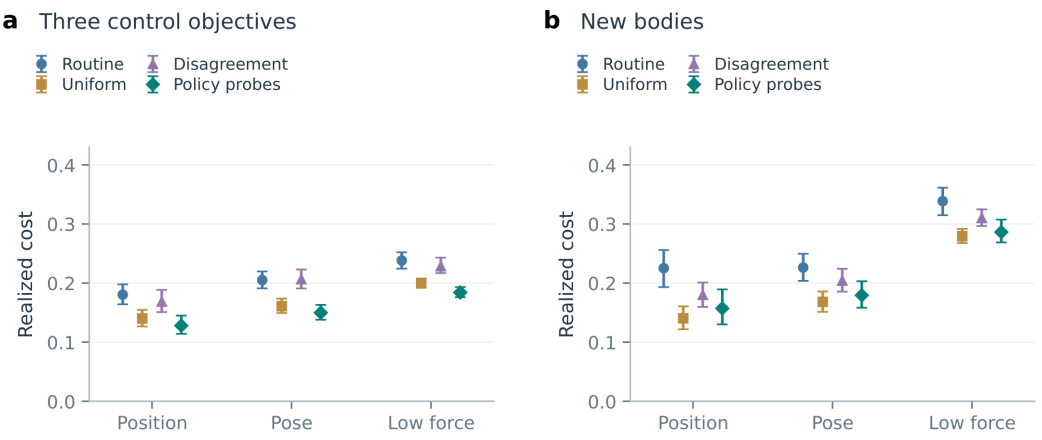


Figure 11: **Objective and body variation at 6,000 response labels.** Position, pose, and force-sensitive costs on original bodies (a) and shifted bodies (b). Lower cost is better. Each model chooses among the same 27 skills using the selected objective. Points and bars show means and 95% bootstrap intervals over contexts and three initialization seeds. There are 256 original and 128 shifted contexts.

256 contexts and three initialization seeds. Policy probes improve success over routine collection by 17.06 percentage points, with a paired 95% interval of 10.55–23.31. Their advantage over uniform collection is 5.47 points, with interval −0.91–12.11. The advantage over routine collection is larger and more precisely resolved than the difference from uniform coverage.

At 12,000 labels, the corresponding successes are 47.27%, 57.94%, 57.42%, and 66.93%. Policy probes at 6,000 labels exceed routine collection at 12,000 labels by 15.89 points (95% interval 9.38–22.40). Routine validation MSE favors routine collection at every budget and seed, whereas independent probe validation favors policy probes. Selecting the candidate model using each of these validation distributions yields gains of 18.88, 17.06, and 19.66 points at budgets 3,000, 6,000, and 12,000, respectively. The model-selection result follows from independent validation responses, with the final comparison measured on untouched test contexts.

Under shifted body parameters and 6,000 labels, uniform collection yields 59.90% success, policy probes 58.85%, routine collection 42.19%, and disagreement 42.71%. Uniform sampling also gives smaller mean position cost than policy probes in this regime. Broader support can benefit a changed body even when more focused collection helps the original deployment class. The useful collection distribution follows both the skills to be compared and the physical contexts in which they will run. Focused probes provide more observations near the intended choices on original bodies; uniform coverage supplies responses across a wider range of directions and arguments. The shifted-body result shows why these two forms of coverage can exchange their ordering.

B.3.6 Expanding the skill interface with fixed models

We restrict the same physical response bank to three nested interfaces. The smallest contains the three speeds with requested distance multiplier one and steering zero. The intermediate interface adds all three distance multipliers, giving nine skills with zero steering. The full interface adds the three steering offsets, giving 27 skills. Every learned model is evaluated on all three interfaces, all test contexts, and all three objectives. Model weights, response predictions, and physical outcomes are held fixed. These comparisons reuse the complete bank of 10,368 physical executions.

On original bodies at 6,000 labels, mean least realized position cost is 0.0653, 0.0602, and 0.0269 for the three interfaces. Routine-model selected costs are 0.1339, 0.1711, and 0.1804. Policy-probe selected costs are 0.0944, 0.1344, and 0.1280. Thus expansion improves the set of registered physical responses while exposing errors that can worsen selection. The full 27-skill decisions reproduce the primary evaluation exactly. Figure 14 relates this expansion to the largest same-skill prediction discrepancy and to independent validation for choosing among models.

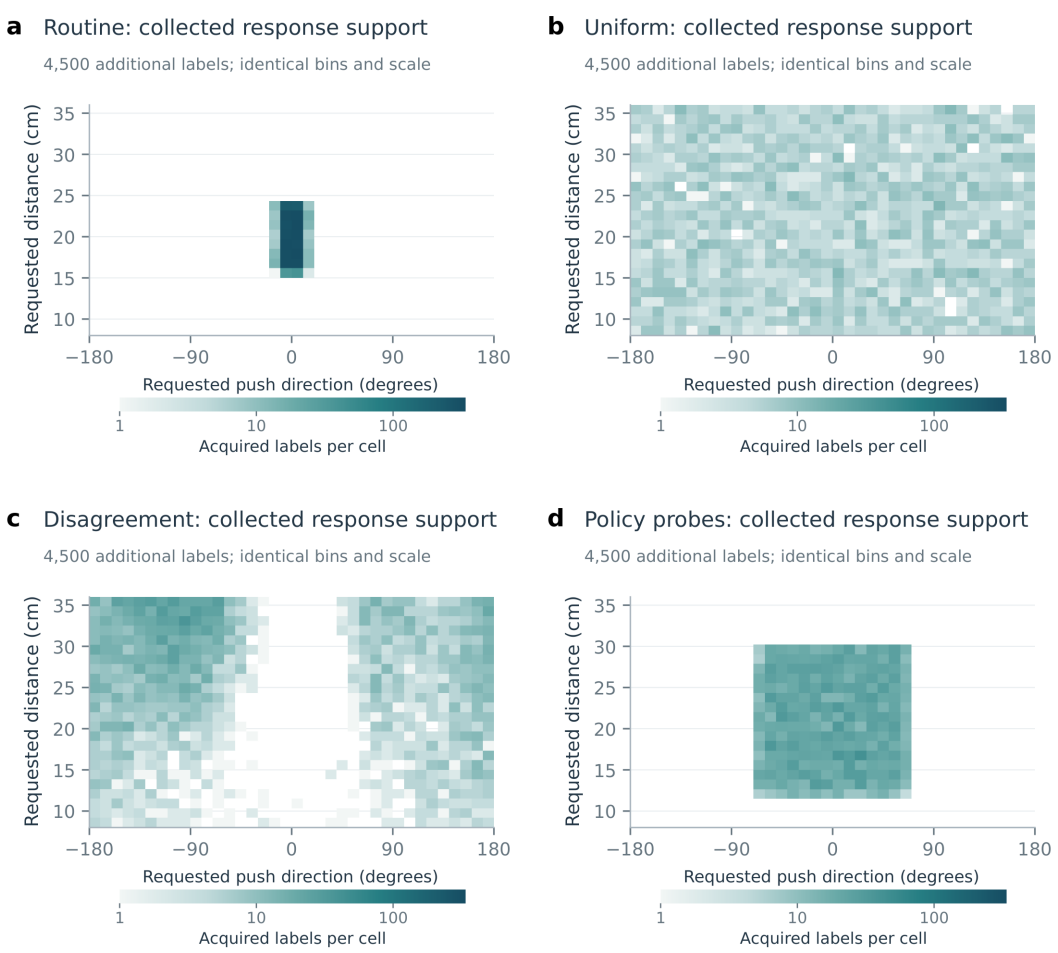


Figure 12: **Where additional response labels are collected.** Each panel contains all 4,500 additional labels at a total budget of 6,000, for routine, uniform, disagreement and policy-probe collection. The methods also share 1,500 initial routine responses. Requested direction and distance are shown on identical axes and one logarithmic count scale. The complete spatial distributions show which responses receive supervision. Counts are acquired response labels; the three initialization seeds share these collections.

The least realized registered cost decreases by set inclusion. Selected cost also depends on which option the model prefers. The available-response curve measures the opportunities created by additional skills. The selected-response curve also includes the model's comparisons among those opportunities. A new skill can improve the first curve and receive an excessively favorable prediction that worsens the second. Appendix B.4 resolves the change into individual additions and their predicted advantages.

B.3.7 Prediction errors and the selected response

For a fixed context, let c_j and $\hat{c}_j$ be the realized and predicted costs of each of the 27 skills, and let $\hat{j}$ minimize $\hat{c}$. Direct comparison with a least-cost registered response gives

$$c_{\hat{j}} - \min_j c_j \leq 2 \max_j |c_j - \hat{c}_j|.$$

The proof adds and subtracts $\hat{c}_{\hat{j}}$ and $\hat{c}_{j^*}$, and uses $\hat{c}_{\hat{j}} \leq \hat{c}_{j^*}$. The saved predictions satisfy this inequality in all 44,928 decisions, within 10^{-7} numerical tolerance. Figure 15 shows the finite-choice envelope and every paired context effect. The empirical error is evaluated over the registered choices

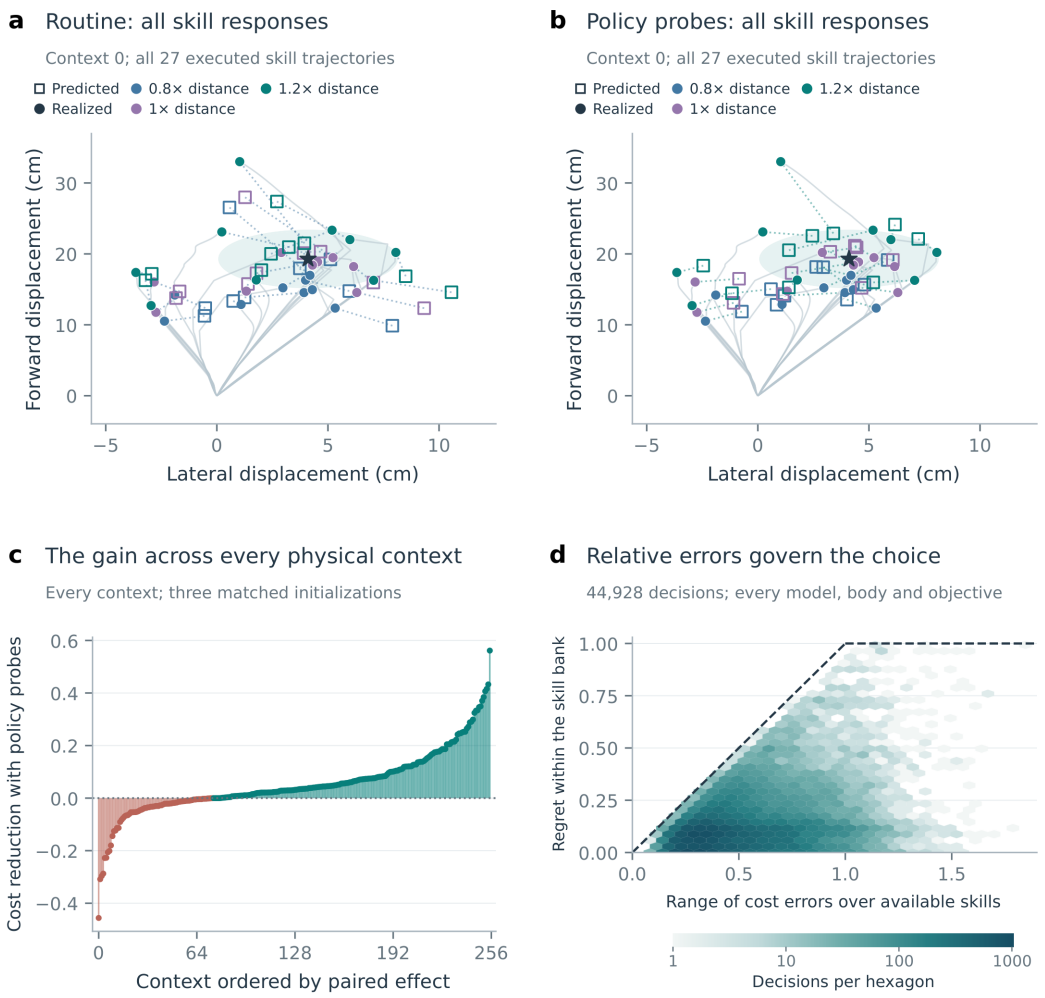


Figure 13: **Prediction errors within and across physical contexts.** a,b, All 27 realized skill trajectories in held-out context zero, together with predicted and realized endpoints for routine and policy-probe models at 6,000 labels and initialization seed 11. Dotted segments join each prediction to the same skill's realized endpoint; both panels use identical limits. c, Every original-body context's paired cost reduction with policy probes, averaging the same three model seeds and retaining negative effects. d, All 44,928 planning decisions. Hexagons count decisions. The dashed envelope is the finite-choice bound given by the range of signed cost errors. These decisions reuse the registered physical response bank.

for the specified scalar cost. The general transfer theorem supplies the corresponding statement for a declared class of feedback policies and bounded trajectory objectives.

Figure 16 resolves the comparisons behind skill selection. It displays all candidate costs in a single context and every context–skill comparison across the held-out physical population. Each error compares two skills in the same physical context: the candidate and the best recorded response. A common cost bias cancels in this difference. The display therefore exposes the prediction errors that can change a ranking even when average prediction error appears small. Context order is shared across collection rules; individual initialization values remain in the source table.

B.3.8 Numerical sensitivity and statistical units

The simulator includes three elementary tests: an unmoving state remains at rest to 2.3×10^{-16}, the discrete free-space servo matches its analytic recurrence to 8.4×10^{-17}, and ground friction



Figure 14: **Choosing a model and expanding its control interface.** a, The gain in held-out success from selecting among trained candidates using independent policy-probe validation rather than routine validation, with paired 95% bootstrap intervals. Both rules inspect the same candidate models and exclude test outcomes. b, The largest same-skill cost prediction error across nested sets of 3, 9, and 27 skills, averaged over 256 original-body contexts and three seeds at 6,000 labels. Bars are 95% context-and-seed bootstrap intervals. Every model and physical context is held fixed across the interfaces.



Figure 15: **Individual context effects and the registered decision envelope.** a, Routine position cost minus policy-probe position cost at 6,000 labels, averaged over three matched initialization seeds within each of 256 original-body contexts. Every context is shown, ordered by its effect. Positive values favor policy probes. b, Hexagonal density of all 4,608 position-task decisions at 6,000 labels across twelve models and both body regimes; darker cells contain more decisions. Excess cost over the best recorded skill is plotted against twice the largest same-skill cost prediction error, clipped at one. The dashed diagonal is the finite-choice bound.

dissipates the kinetic energy of a released object to zero. These tests check specified numerical operations.

A separate resolution study follows the same controller in 128 contexts from each body regime with 40, 80, and 160 physics substeps. Relative to 160 substeps, median endpoint differences at 40 are 1.61 mm on original bodies and 1.05 mm on shifted bodies. The corresponding means are 27.35 and

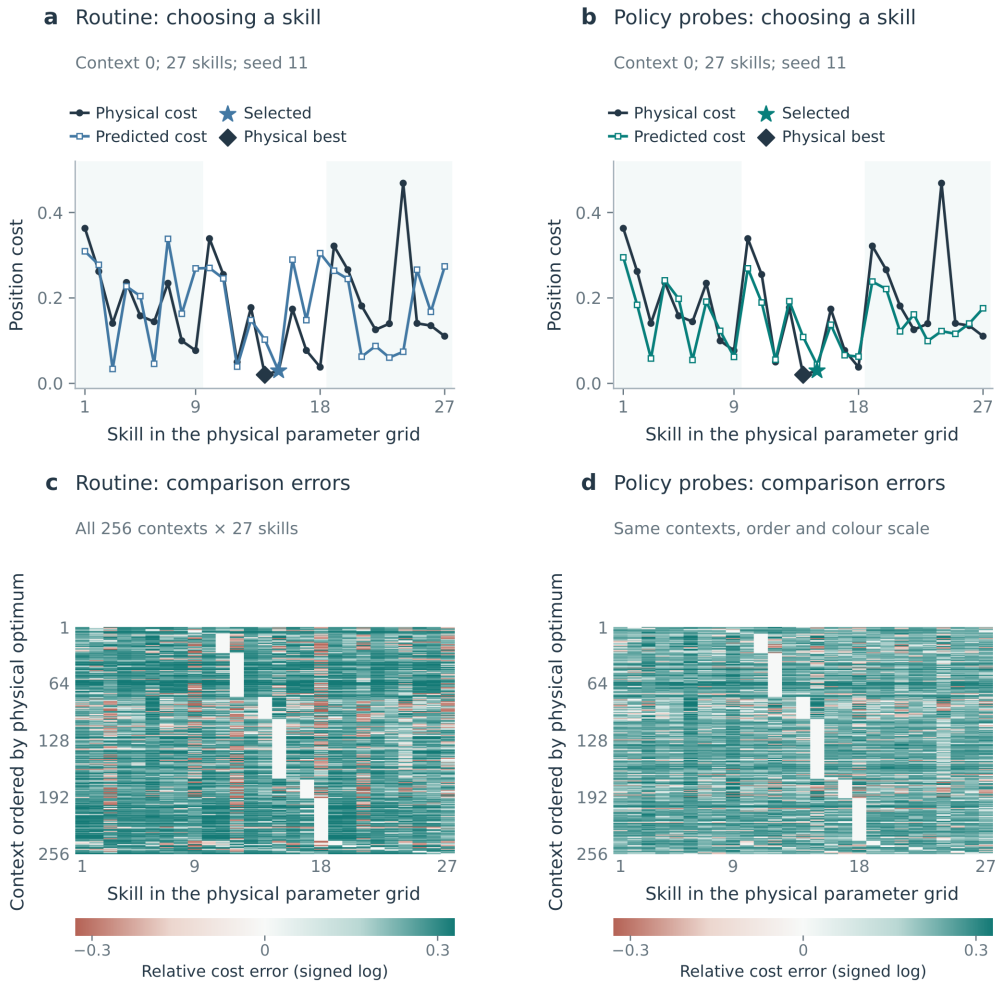


Figure 16: **A model chooses through comparisons between physical responses.** a,b, Physical and predicted position costs for all 27 skills in held-out context zero and initialization seed 11, at 6,000 labels. Stars mark model-selected skills; diamonds mark the best physical responses. The background separates the three distance groups in the physical skill grid. c,d, All 256 × 27 context–skill comparisons, averaged across three matched model seeds. Context order and color limits are shared. The color scale is signed logarithmic and linear within ±0.02. Each error equals the actual cost difference from the physically best skill minus the predicted difference from that same skill. Positive values indicate an overly favorable predicted comparison. The full source table retains all individual seed values.

18.76 mm, with maxima 402.25 and 481.94 mm. These long tails reflect contact-sensitive trajectories; median agreement alone does not characterize the solver's effect on performance.

We therefore re-execute the selected skills from every 6,000-label model at all three resolutions, keeping model weights, chosen skills, initial post-calibration states, and targets fixed. On original bodies, the policy-probe versus routine successes are 63.15% versus 46.09% at 40 substeps, 63.15% versus 46.35% at 80, and 63.41% versus 45.96% at 160. On shifted bodies they are 58.85% versus 42.19%, 58.59% versus 42.97%, and 58.07% versus 42.97%. All other collection conditions and every re-executed response remain in the results. This checks the stability of the collection comparison under integration refinement within the defined contact model.

For H, the bootstrap resamples physical contexts and retains both cue branches and all compared conditions. For I, it resamples physical contexts and the three initialization seeds independently, using the same resampling weights for paired methods. Primary summary intervals use 4,000 bootstrap

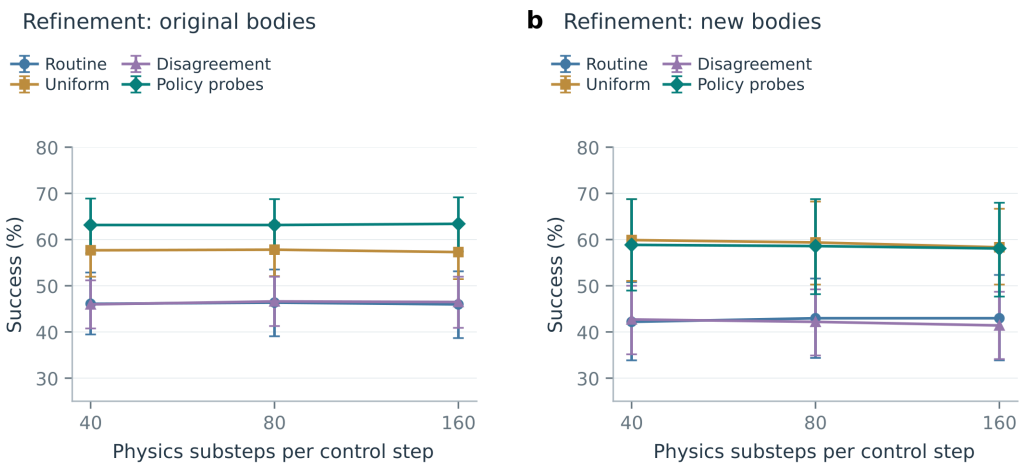


Figure 17: **Re-executing selected skills at finer contact resolution.** Success on original (a) and shifted (b) bodies at 40, 80, and 160 physics substeps per public step. All twelve 6,000-label models retain their weights and previously chosen skills. Points and bars show means and 95% bootstrap intervals over contexts and initialization seeds. These comparisons retain the full context population at each resolution.

draws; primary paired comparisons use 8,000. The expanded-interface analysis uses 2,000 and 4,000, respectively. All use seed 17032026. Intervals are percentile 95% intervals for the stated comparisons, with no adjustment for multiplicity. The three initialization seeds share one collected dataset per condition. Consequently, the intervals describe variation across physical contexts and network initialization conditional on those collections. Expansion orders and alternative starting banks reuse this same physical response bank; they supply matched comparisons within a context.

All training response labels, fitting splits and normalizations, independent validation measurements, acquired candidate indices, full held-out skill trajectories, model predictions, per-decision outcomes, and plotting source data are released. The original and shifted test sets remain separate throughout the analysis. The model-selection rule reads validation losses before evaluating selected candidates on those test sets.

B.3.9 Code, arrays, and figure reconstruction

The simulator and feedback controller are implemented in `code/continuous/contact.py`. Study H runs through `run_information.py`; its `information/episodes.csv` records every context, branch, reveal time, commitment length, target error, success, outcome class, and peak force. Its `trajectories.npz` additionally records complete state, action, and force sequences for cue steps 0, 35, and 67, alongside initial states and body parameters.

Study I runs through `run_learning.py`. The `learning/` directory contains the shared base data, all collection pools, both independent validation sets, both complete test banks, 39 checkpoints, training curves, validation metrics, and predictions. Test-bank arrays retain the 17-dimensional input, five-dimensional response, physical parameters, calibrated state, target, skill index, and full physical trajectory. `deployment.csv` contains one row per model, context, and objective, including the chosen skill, realized cost, success, force, orientation change, and registered error quantities.

`analyze.py` constructs the paired summaries; `analyze_interfaces.py` constructs the nested-bank and validation-selection comparisons. `check_physics.py` and `check_decisions.py` produce the integration-resolution results. Paths here are relative to `results/continuous/` for outputs and `code/continuous/` for scripts unless a full relative path is given. Every plotted result has a CSV source table under `figure_source/data/`; figure scripts read the saved observations and predictions.

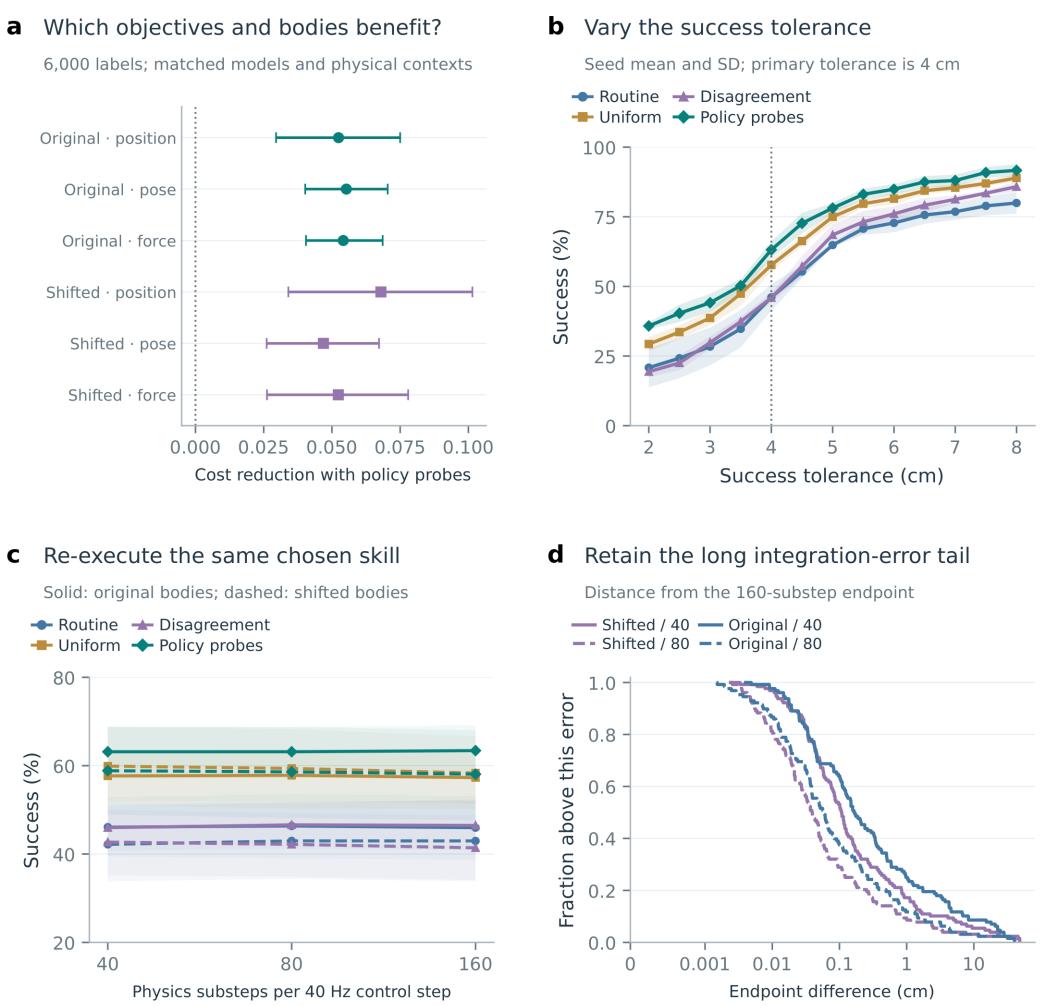


Figure 18: **Objective, tolerance, body distribution and numerical resolution.** a, All paired policy-probe versus routine cost effects across three objectives and both body regimes; bars are 95% context-and-initialization intervals. b, Complete success-tolerance curves on original bodies, with one initialization-seed SD. c, Fixed selected skills re-executed at three integration resolutions; bands are paired 95% intervals, with solid and dashed curves for original and shifted bodies. d, Complete endpoint-difference distributions relative to 160 physics substeps. Solid and dashed curves denote 40 and 80 substeps. The long error tail is retained. These numerical comparisons concern the specified contact simulator.

B.4 Skill additions, order variation, and the starting interface

B.4.1 Question and matched comparison

Adding a skill creates another physical opportunity and another comparison for the predictor. We study both effects using the complete response bank from experiment I. The target, physical context, executed trajectories, and trained model are held fixed. The planner receives a larger subset of the recorded skills and selects again. Any change in its realized cost can then be traced to the newly available comparison.

The single-addition and expansion-order analyses evaluate all 39 models, both body regimes, and all three objectives. An exhaustive analysis of starting banks uses the position objective and the twelve 6,000-label models. All comparisons reuse the same physical executions and fitted predictions.

The bank is the Cartesian product of distance multipliers $\{0.8, 1, 1.2\}$, steering offsets $\{-0.25, 0, 0.25\}$ radians, and speeds $\{0.5, 0.75, 1\}$. Let B_0 contain the three speeds at multiplier one and zero steering. For each of the other 24 skills, we compare selection from B_0 with selection from $B_0 \cup \{j\}$. Ties retain the incumbent during an expansion and use the stored skill order for an independently evaluated bank. An incumbent remains available after each addition. The full-bank predictions contain no ties at the minimum in these evaluations.

For each context and model, the predicted gain is the incumbent's predicted cost minus the selected candidate's predicted cost. The realized gain uses the costs measured from their saved physical responses. Positive gain means the addition improves control. The predicted gain is nonnegative by selection. The best realized gain compares the minimum realized cost available in each bank and is also nonnegative. The realized selected gain can have either sign.

We store the selected-switch indicator and both signed prediction residuals. The identity in Proposition 9 holds to 1.20×10^{-7} across every individual addition in both regimes and all objectives. This identity identifies how a predicted improvement becomes a physical loss. If the newly selected option is predicted too favorably relative to the incumbent, the error in their comparison can exceed its predicted gain. The realized gain then changes sign. The saved prediction residuals locate that reversal for each addition and context.

B.4.2 Expansion order and all alternative starting banks

We independently shuffle the 24 added skills 128 times, starting from B_0 in every order. At each bank size from 3 through 27, we compute the selected cost, predicted selected cost, best realized cost, largest absolute cost error, and probability of changing the selected skill. Orders are averaged within each physical context and model before estimating uncertainty. The endpoint is identical across orders because each contains the same full bank. Random order describes alternative paths through the same expansion, with uncertainty evaluated across physical contexts and initialization seeds.

The designated starting bank corresponds to three speeds for a direct push. To determine the dependence on this choice, we also enumerate all $\binom{27}{3} = 2{,}925$ possible three-skill banks. For the position objective, each context and each 6,000-label model is evaluated on every one of these banks. We compare its average selected cost over the 2,925 banks with its selected cost from the complete bank. We also retain average best realized cost, the fraction of banks for which expansion is harmful, and the fraction for which it is helpful. This exhaustive comparison describes the scope of the effect seen from B_0. We also retain each bank's mean effect across contexts and initializations. The cumulative distributions in Figures 4 and 21 show every one of these bank means, with the designated bank marked on the same scale. Each point in that empirical distribution is a starting bank's mean effect across the same contexts and initializations. The horizontal spread describes how the effect depends on the initial options. Bootstrap intervals answer the separate question of variation across sampled contexts and model initializations.

B.4.3 Results and their interpretation

On original bodies with the position objective and 6,000 labels, 22 of the 24 single additions to B_0 have negative mean realized gain under routine models. The corresponding counts are 11 for uniform collection, 21 for disagreement, and 19 for policy probes. Every added skill and every physical context is included. The counts summarize the signs of the mean effects; paired intervals for every addition appear in Figure 22.

Averaging across all additions, routine models predict a cost reduction of 0.00889 and incur an increase of 0.00712. Probe models predict a reduction of 0.00578 and incur an increase of 0.00423. The signed error difference accounts for the reversal exactly. Thus the effect can occur after adding a single option, before any long sequence of expansions. Figure 19 includes the individual additions for every collection rule and the paired changes across the starting-interface population.

Across all alternative three-skill banks, expansion improves mean position cost for all collection rules on original bodies (Table 3). Their initial costs are larger on average than those of the designated direct-push bank. Additional choices can then overcome poor initial options despite imperfect prediction. On shifted bodies, the routine comparison is close to zero and its interval spans both signs. These results reveal two contributions to expansion. Across the population of starting banks

Table 3: **Expansion from every possible three-skill bank.** Mean selected position cost with 6,000-label models. The three-skill column averages all 2,925 banks within each context before averaging contexts and seeds. The change is full-bank cost minus three-skill cost, so negative values favor expansion. Brackets are paired percentile 95% intervals.

Body regime	Collection	Three skills	27 skills	Change [95% interval]
Original	Routine	0.1975	0.1804	−0.0171 [−0.0303, −0.0026]
	Uniform	0.1749	0.1401	−0.0347 [−0.0462, −0.0225]
	Disagreement	0.1888	0.1687	−0.0201 [−0.0336, −0.0047]
	Policy probes	0.1625	0.1280	−0.0345 [−0.0470, −0.0211]
Shifted	Routine	0.2265	0.2251	−0.0014 [−0.0222, 0.0190]
	Uniform	0.1865	0.1403	−0.0461 [−0.0614, −0.0288]
	Disagreement	0.2100	0.1800	−0.0301 [−0.0473, −0.0129]
	Policy probes	0.1836	0.1572	−0.0264 [−0.0482, −0.0032]

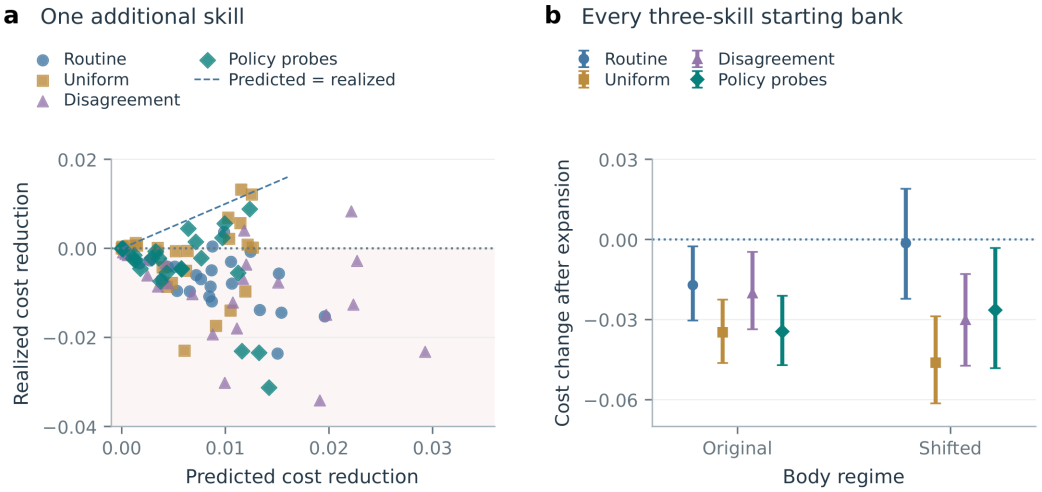


Figure 19: **The predicted comparison and the starting interface.** a, All 24 single additions for each of the four collection rules, on original bodies at 6,000 labels. Coordinates are mean predicted and realized cost reductions. Each rule shares 256 contexts and three initialization seeds. b, The full-bank cost minus the mean cost over all 2,925 initial banks, computed within each context, for original and shifted bodies. Negative values favor expansion. Bars are paired 95% intervals over contexts and initialization seeds.

on original bodies, gains from additional responses outweigh selection errors on average. A useful starting bank can still lose performance when an overly favorable prediction of an alternative displaces a good incumbent. The direct-push bank lies in this second regime. Its location in the distribution of all starting banks explains the deterioration observed in the main comparison.

B.4.4 Relative error and finite-choice regret

For the complete bank, we additionally compare regret with the range of signed cost errors, $\max_j(c_j - \hat{c}_j) - \min_j(c_j - \hat{c}_j)$. All 44,928 decisions satisfy the bound in Proposition 9 within 3×10^{-7}. The range removes common cost bias and can give a smaller bound than twice the largest absolute error. It is calculated from the fully observed registered response bank. For future deployment, applying the bound requires a simultaneous estimate of these comparison errors over the proposed skill bank.

B.4.5 Randomness, uncertainty, and complete outputs

The expansion-order generator uses seed 20260905. Single-addition and order summaries use 2,000 bootstrap draws; the exhaustive-starting-bank summary uses 4,000. The bootstrap independently

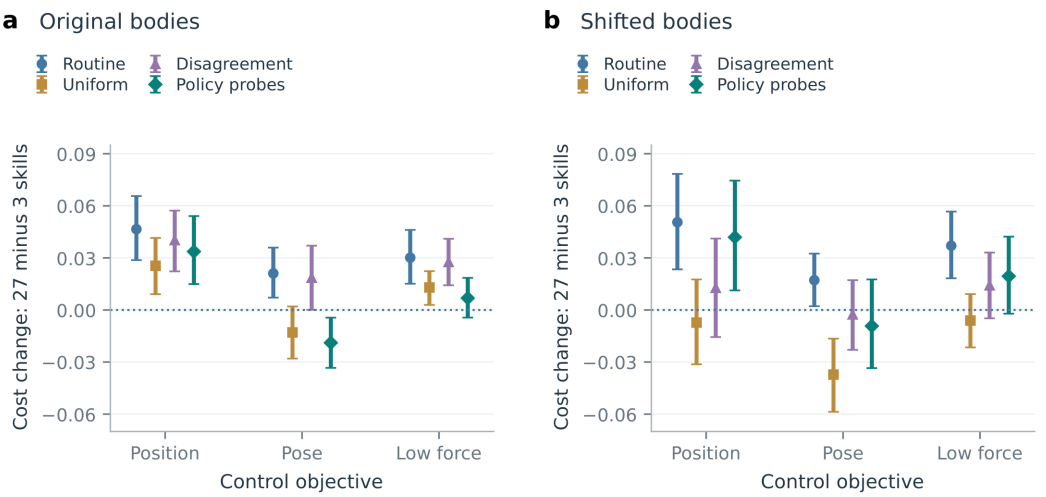


Figure 20: **Expansion across objectives and bodies.** Changes in selected cost from the designated three-skill bank to the full 27-skill bank at 6,000 labels. a, Original bodies. b, Shifted bodies. Each collection rule is evaluated for position, pose, and force-sensitive cost. Positive values indicate deterioration. Bars are paired 95% intervals over contexts and initialization seeds; no expansion order changes these endpoint comparisons.

resamples physical contexts and the three initialization seeds, and preserves matched comparisons. Initialization seeds share the collected data, as in experiment I. Intervals are conditional on those collections and on the registered physical test bank. All means include every context. The order study uses 128 random permutations, while the starting-bank study is an exact enumeration.

The single-addition outputs retain every model and all 24 additions in both body regimes and all three objectives. The order analysis evaluates 25 bank sizes along 128 permutations and stores the mean over orders within each model and context, together with the explicit permutations, random seed, and summary tables. The exhaustive-bank outputs cover all 2,925 starting banks for the position objective at 6,000 labels in both body regimes. Per-model and pooled summaries retain the comparison errors and their statistical units. The analyses are implemented in `code/continuous/analyze_expansion.py`, `summarize_expansion_scope.py`, and `analyze_visual_structure.py`.

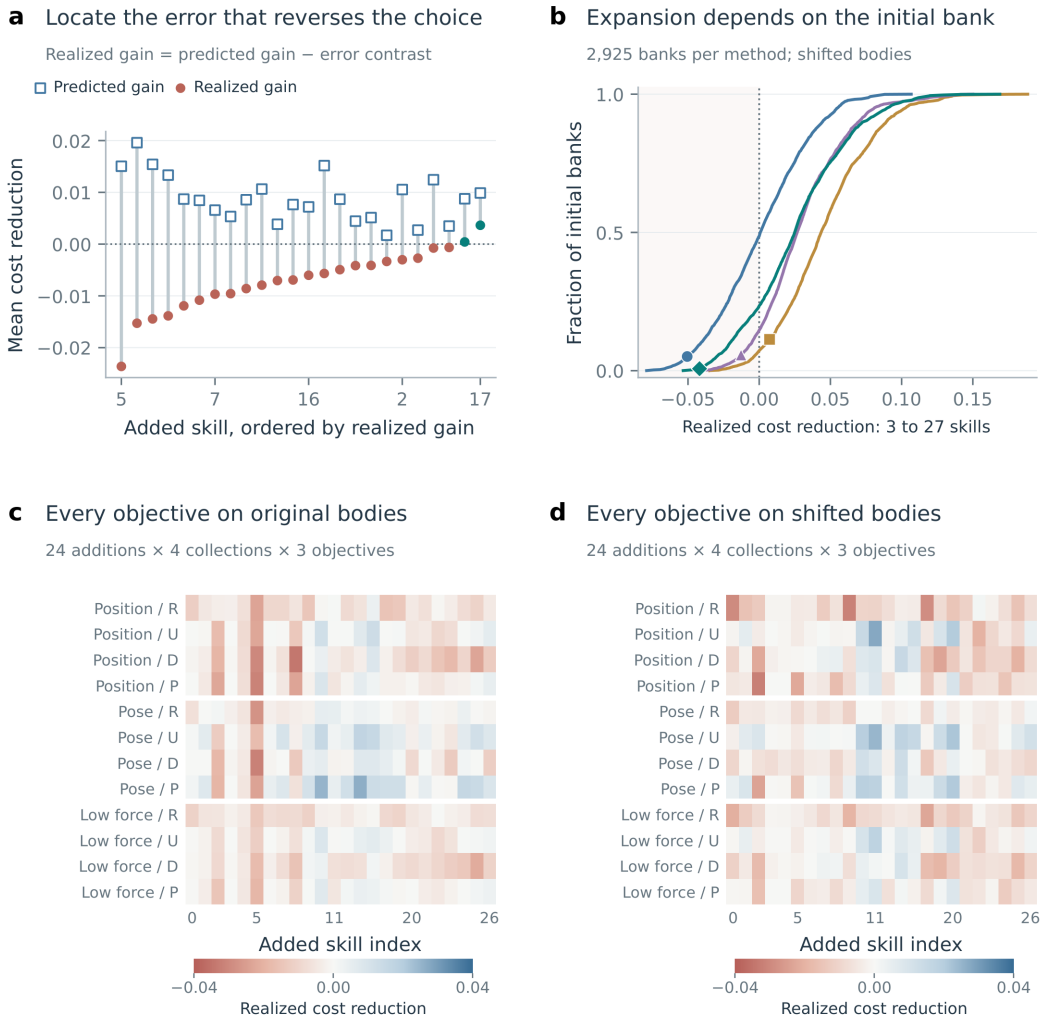


Figure 21: **An added response changes a comparison between skills.** a, All 24 additions under routine collection, decomposing the difference between predicted and realized gain. Each segment's length is the mean signed error contrast. b, Full distributions over all 2,925 initial banks on shifted bodies; marks locate the designated direct-push bank. c,d, Every addition for every collection rule and objective, on original and shifted bodies. R, U, D and P denote routine, uniform, disagreement and policy probes. Cells share one symmetric cost-reduction scale. Every entry averages the same contexts and three initialization seeds. Banks and cells are dependent comparisons of the saved physical responses.

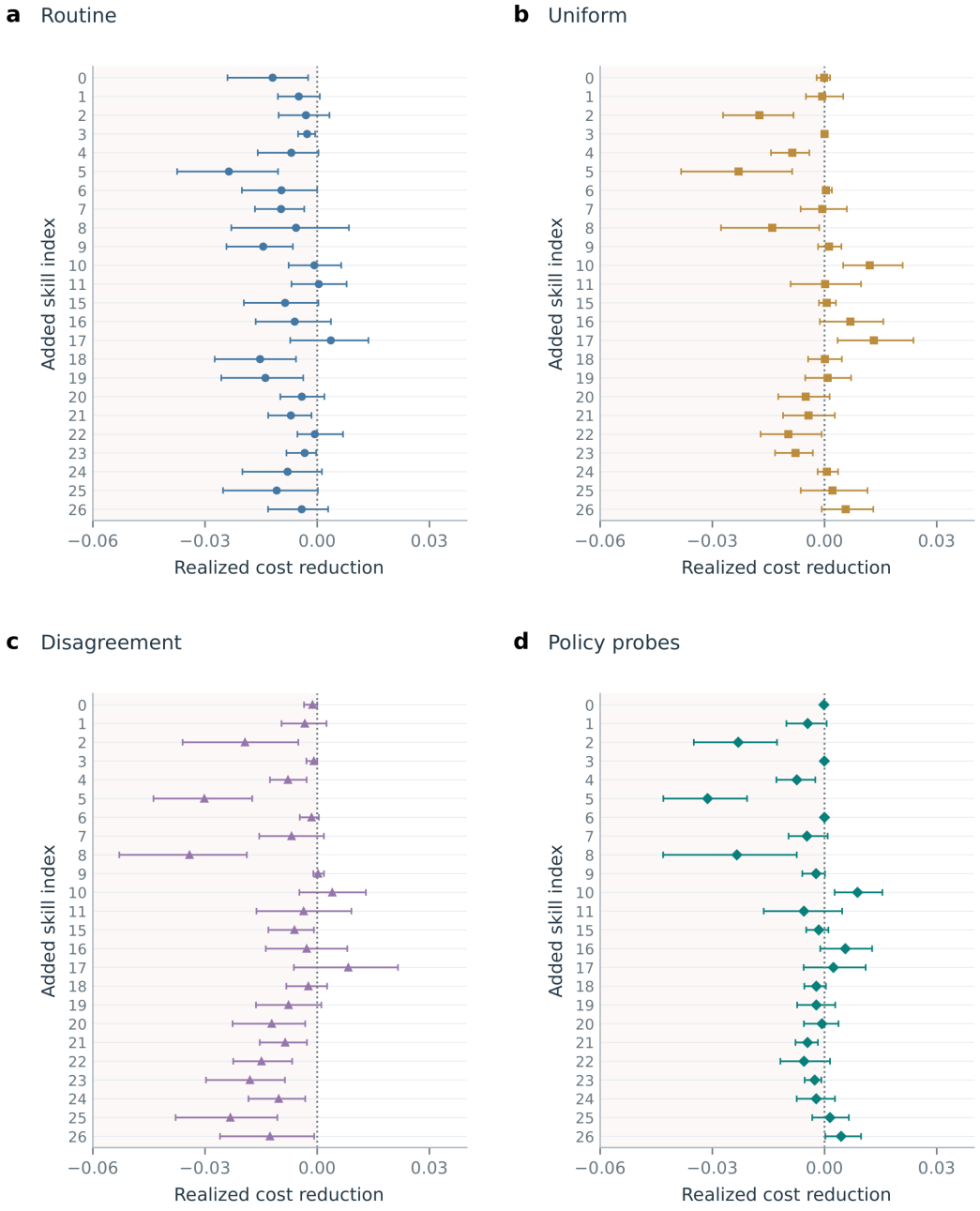


Figure 22: **Every single-skill addition with paired uncertainty.** Each panel shows all 24 additions to the same initial three-skill bank, for one collection rule at 6,000 labels. Points and bars are mean realized position-cost reductions and paired 95% intervals over 256 original-body contexts and three initialization seeds. Negative values indicate deterioration. Skill indices follow the stored Cartesian product of distance multiplier, steering, and speed; the source table retains the arguments of every skill.

C FINITE PROCESSES AND LEARNED VISUAL MODELS

C.1 EXPERIMENTAL DESIGN AND REPORTING

The experiments address two questions in using a world model: whether a selected policy retains its predicted consequences, and whether legal actions can produce a requested trajectory. Studies A and B examine the first question by calculating complete interaction laws. C and D locate the requirements of a requested trajectory in its initial conditions, information, and physical responses. E and F ask which parts of those relations a prediction objective measures. G changes the responses available to the agent. The continuous studies then bring these questions together in object pushing: H varies the time at which a target cue can affect motion, I varies the responses observed during learning, and the skill-expansion analysis varies the choices presented to the same learned model.

Table 4 links each question to its intervention and methods. The illustrated guide in Appendix G introduces the task settings. Each study below describes the physical process, the information available to its controller, the comparison being made, and the interpretation of its measurements.

Table 4: **Guide to the experiments.** Each study changes a specified part of the model–control relation. A–G use finite processes; H and I use the shared continuous contact simulator.

Study	Question	Task and intervention	Details
A	Does a command retain its meaning?	Position control with reversed actuators; calibrate and translate the command map.	C.2
B	Which discrepancy controls transfer?	Three process families, 180 model pairs, and 2,160 bounded-task planning comparisons.	C.3
C	What prevents a requested path?	Preparation, sensing, and gripping; intervene on reset, information, and tool response.	C.4
D	Can successful paths be imitated?	Wind-dependent actions and outcome-conditioned demonstrations; compare exact projection with fitted policies.	C.5
E	Which responses do logs identify?	Inspection, route selection, and safe or fast execution; cross coverage with hidden action selection.	C.6
F	Which prediction errors affect control?	Visual navigation and pushing; independently perturb rendered pixels and contact predictions.	C.7
G	Which outcome laws are available?	Mixtures of primitive responses and a feedback skill; vary physical interface and model accuracy.	C.8
H	When can a cue affect execution?	A target revealed during a push; vary action commitment and cue correspondence with exact prediction.	B.2
I	Which experiences support selection?	Learn 27 pushing-skill responses; vary collection and validation distributions.	B.3
I	How do new skills affect selection?	Hold models and responses fixed; add skills and enumerate all three-skill starting banks.	B.4

Every experiment has a finite horizon. A trajectory includes the actions and the observations available before subsequent actions. C and D additionally retain hidden variables when evaluating the full-law projection. An outcome-only calculation is identified explicitly. Total variation uses the convention $\mathrm{TV}(P, Q) = \frac{1}{2}\sum_{\omega} |P(\omega) - Q(\omega)|$; KL uses natural logarithms and is reported in nats. Rewards in A, B, and E are bounded by one. F retains the specified utility in its numerical tables. Its regret figures divide utility differences by the utility range 1.48 to place them on a unit scale.

Numerical and statistical units. A uses rational arithmetic for its principal calculations. B computes finite path laws, with some transition matrices estimated from multinomial samples. C and G solve finite convex projections. D evaluates the population law of each fitted policy exactly. E evaluates fitted tabular models against exact environment laws. F estimates deployment values from episodes. A displayed spread across training seeds is a sample standard deviation. E averages its eight datasets within each environment before calculating intervals across the 128 environments. G displays every specified target direction and tilt strength; these are exact geometric calculations.

Policy classes. A and B maximize their principal discrepancy over all feedback policies on the stated public histories. B also computes the model-selected reward policy and the physical optimum

over this full history-dependent class; a stationary registry supplies additional comparison metrics. E reports a maximum over 48 specified randomized policies. F uses a selected discriminator and three probing policies, and estimates regret within a four-controller class. G's one-step mixture problems permit all randomized choices among the listed coarse actions. These distinctions determine which quantities are exact discrepancies, which are lower bounds, and which are deployment measurements.

Files and reconstruction. The package contains all nine experiment families, their resolved configurations, numerical tables, and 75 neural checkpoints: 36 for F and 39 for I. H and I additionally include continuous trajectories, action sequences, contact forces, acquired response labels, and predictions for every registered test skill. Tables retain the experiment letter in their filename. The figure source-data directory contains the exact rows and transformations used for each display. The commands in the package README rebuild the numerical experiments, figures, and manuscript separately. The descriptions below specify the scientific protocol; the configuration supplies every numerical array and seed in machine-readable form. Finite implementations are `code/experiments/experiment_a.py` through `experiment_g.py`, with shared settings in `code/config/publication_config.json`. Their output tables are grouped by the same letter under `results/tables/`. Continuous implementations, configurations, and arrays are mapped at the end of Appendix B.

C.2 A: COMMAND MEANING AND CALIBRATION

An agent can know the attainable goals while assigning the wrong physical meaning to its commands. We examine two position-control systems whose commands differ by a permutation. Their attainable outcome values agree. A policy selected in the model is then executed with the original commands and with a calibrated translation. This comparison asks whether correcting one command map restores the consequences of every policy, including policies used for different goals.

C.2.1 ENVIRONMENT AND TASKS

The physical position lies in $\{-2, -1, 0, 1, 2\}$ and starts at zero. Commands are $-1, 0, 1$. A nonzero command moves in its intended direction with probability 0.60, in the opposite direction with probability 0.25, and stays with probability 0.15. A zero command stays in place. Motion beyond a boundary is clipped to the boundary. The model uses the stated command meanings. The environment interchanges the two nonzero commands. The corrected interface sends $-a$ whenever the policy requests a.

The task is to finish at a designated goal. Goals $-2, -1, 1, 2$ are crossed with horizons $1, 3, 5, 8$, giving 16 tasks. Backward dynamic programming computes the optimal success probability in each system. The selected model optimum is evaluated with the same feedback rule in the environment. Ties use expected final distance, action effort, and the fixed direction order specified in the source. The table of tasks also retains the best and worst deployment values among model-optimal ties. An unreachable goal can therefore have zero regret while a selected command still produces different path laws.

C.2.2 EXACT ADAPTIVE DISCRIMINATION

The discrimination problem places equal prior probability on the model and environment. Its state consists of current position, remaining time, and the posterior over the two systems. At a terminal history, the best decision chooses the hypothesis with larger posterior probability. Backward induction chooses the action that maximizes the expected terminal decision accuracy. Deterministic choices suffice because action randomization gives an average of the action values. Converting accuracy p to $2p - 1$ gives the exact adaptive TV. A separate enumeration of all deterministic feedback trees at horizons one and two agrees with this calculation. Saved witness laws allow direct evaluation of the terminal likelihood test.

C.2.3 CALIBRATION AND PROBING

A calibration trial resets position to zero and issues command $+1$. Trials are independent conditional on the unknown command map. The posterior after the observations selects a map, with equal treatment of a posterior tie. The seven trial counts are $0, 1, 2, 4, 8, 16, 32$. Exact multinomial enumeration

Table 5: Exact interaction discrepancy in A. The correction applies the same fixed translation at every decision.

Horizon	Original Δ	Identification accuracy	Corrected Δ
1	0.350000	0.675000	0
3	0.515375	0.757688	0
5	0.625408	0.812704	0
8	0.735026	0.867513	0

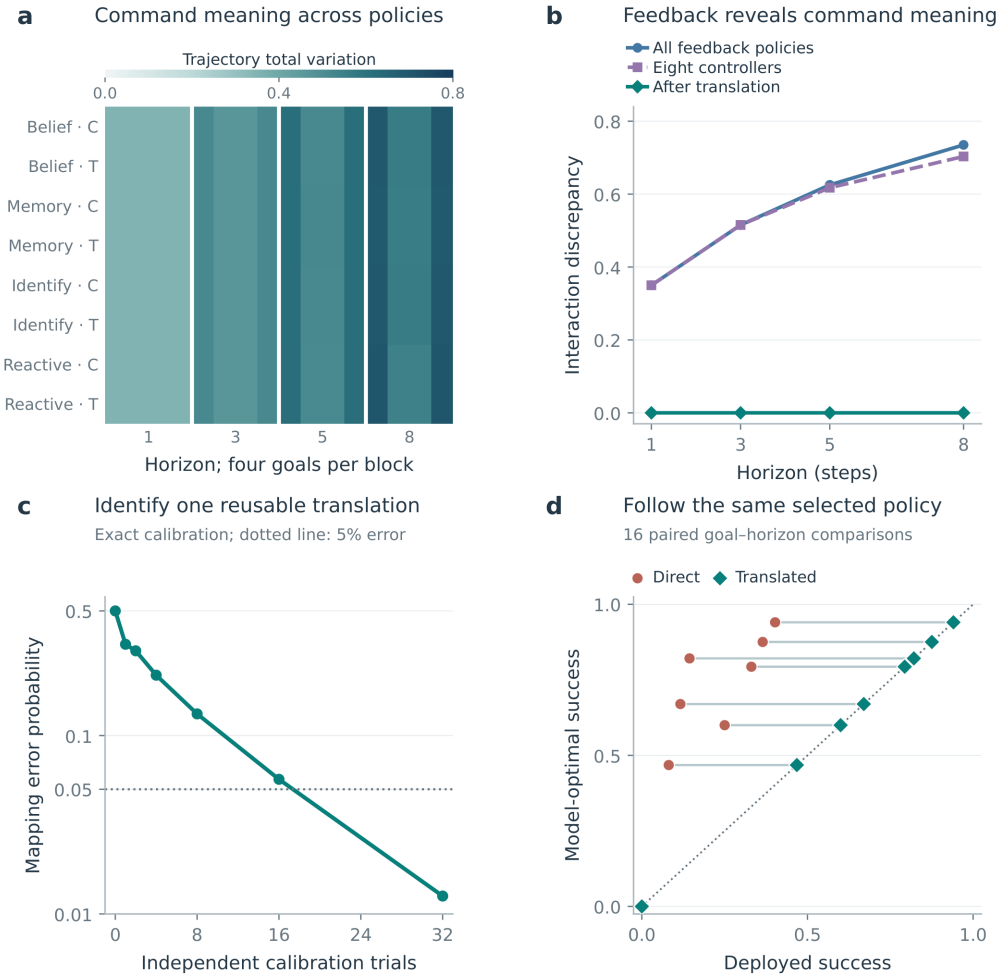


Figure 23: **Command calibration repairs a correspondence between interventions and outcomes.** a, Exact trajectory TV for all eight controllers, four horizons and four goals. C and T denote canonical and translated command maps. Every column retains its goal within the horizon block. b, Discrepancy over all public-history policies, the eight registered controllers and the repaired interface. c, Exact command-mapping error probability over every calibration sample size, on a logarithmic axis. The reference line marks 5% error. d, All 16 selected-policy transfer comparisons. Each segment joins direct and translated execution of the same selected behavior. The diagonal is equal success. These panels enumerate finite probability laws.

gives the mapping accuracy and the expected task loss after choosing the inferred translation. At 32 trials, accuracy is 0.987397 and the mean residual task loss is 0.005481; at zero trials they are 0.5 and 0.217458.

The additional probe registry contains a fixed repeated command, uniform random commands, an optimized open-loop sequence, and an adaptive feedback policy. Comparisons use the same horizon. The command-swap construction can give equal information to an open-loop and an adaptive probe. A separate recovery perturbation makes a response after a deviation informative. The complete probe and recovery tables retain these cases. Eight additional controller definitions examine the command translation across policies with different feedback rules. Translation acts on the command interface, so its equality of path laws extends beyond the task-selected policies. The calibration data identify which physical response belongs to each command. This explains why the same correction supports every tested goal and every feedback rule: it repairs the correspondence used at each decision.

C.3 B: TRANSFER OVER A FINITE COLLECTION OF MODEL ERRORS

The transfer bound concerns the consequence of replacing one controlled process with another. To examine this relation across different errors, we construct models with altered command meanings, ambiguous hidden states, and inaccurate recovery responses. For each model–environment pair, the planner receives twelve bounded task objectives and selects a policy using the model. The same rule then responds to the environment's own observations. Finite path enumeration gives its deployed value, the physical optimum, and the largest interaction discrepancy over the same policy class. The design therefore compares the quantities appearing in the bound directly.

C.3.1 THREE ENVIRONMENT FAMILIES

B contains 60 pairs in each of three families, all at horizon four, using seed 20260828 for the model construction. Each environment has five hidden states and three actions. The command-swap and recovery families expose all five states. The partially observed family maps hidden states to observations $(0, 1, 1, 2, 3)$, so states one and two have the same observation.

In the command family, the intended movement probability is $0.60 + 0.025(v \bmod 4)$ and the opposite probability is $0.25 - 0.015(v \bmod 3)$ for variant v. The remaining probability stays in place. The middle action moves left and right with probability 0.10 each and otherwise stays. The process starts at the middle position and clips at the boundaries.

The partially observed family starts at state zero. Its action-conditioned start rows, in action order, are

$$(0, .74, .21, .05, 0), \quad (0, .45, .45, .10, 0), \quad (0, .21, .74, .05, 0).$$

At hidden state one, the rows are

$$(0, .10, 0, .05, .85), \quad (0, .30, 0, .15, .55), \quad (0, .20, 0, .60, .20).$$

State two reverses the action ordering and places the continuation mass in state two. State three returns to state zero with probability 0.12, increased to 0.15 for the middle action. State four is absorbing. A variant adds $0.006(v \bmod 5)$ to the probability of state four from states one and two, taking that mass from state three, followed by row normalization.

In the recovery family, actions zero and one initially enter state one; action two enters state two. State one persists under the first two actions. Its third action succeeds at state four with probability $0.53 + 0.01(v \bmod 5)$ and otherwise enters absorbing failure state three. At state two, action zero succeeds with probability 0.9 and otherwise stays in state two. The other actions fail. The model perturbation reduces this recovery success probability by the specified severity, placing the removed mass on continued recovery.

C.3.2 MODEL CONSTRUCTION

The model collection cycles through action permutation, kernel smoothing, branch bias, transition delay, fitted kernels, and a family-specific condition. Action permutation mixes the original kernel with its reversed action ordering. Smoothing mixes it with the uniform five-state law, with weight $0.45s$. Branch bias transfers up to $0.40s$ probability to the final state, proportionally removing donor mass. Delay mixes in a self-transition with weight $0.60s$. Hidden-state merging averages transition rows that share an observation, then interpolates with the original rows. Here s is the row's recorded distortion strength. The cycle uses a linear strength schedule starting at 0.10 over a span of 0.85; all resolved strengths are retained in the pair table.

Fitted kernels use $N \in \{16, 64, 256\}$ multinomial draws per state–action row and add 0.5 to each outcome count. There are 15 such rows, so a fit uses $15N$ transition draws. These are samples from explicitly queried transition rows. Other perturbed models use a probability floor of 10^{-10} and row normalization. A complementary control changes an event of probability 0.02 to this floor in the middle action's central row. It can make KL large while changing a small amount of probability mass. The recovery family provides the reverse separation: the passive middle-action policy never queries the altered recovery row.

C.3.3 Policies, rewards, and exact calculation

For each reward, reverse dynamic programming optimizes over all deterministic policies that can use the complete public observation history. Randomization cannot improve the expected reward beyond a deterministic optimum. The recursion carries the environment and model hidden-state masses along each history. It chooses the action with the largest model continuation value, records the smallest action index at an exact numerical tie, and evaluates that chosen continuation with the environment masses. A separate recursion using the environment in both positions supplies the physical optimum. The implementation also propagates the best and worst physical values among exactly model-optimal continuations. A tolerance of 10^{-12} is used to describe near ties in supplementary comparisons; the selected optimum uses the exact numerical ordering.

A stationary comparison registry contains all current-observation maps: $3^5 = 243$ in the fully observed families and $3^4 = 81$ in the partially observed family. This registry supplies restricted discrepancy, value-error, ranking, and reachable-set measurements. The headline planning values use the larger history-dependent class.

The 12 rewards per pair include terminal-observation indicators and normalized path objectives. The latter measure success, avoidance of unchanged observations after directional commands, action energy, movement length, and avoidance of a designated unsafe observation. Directional actions have energy cost one and the middle action cost 0.15. Path contributions are divided by the horizon. Remaining task slots use a convex combination of a terminal-position score and the four path scores, with weights drawn from a five-dimensional Dirichlet distribution whose entries are 1.5. A terminal target is sampled uniformly. The task generator uses seed $20260828 + 7919n_O$, where n_O is the observation count; systems with the same observation count share these task arrays. Every reward lies in $[0, 1]$. The arrays are generated independently of the fitted transition matrices and supplied to the planner during evaluation.

The principal Δ maximizes over the larger class of all public-history feedback policies. A reverse recursion carries the unnormalized hidden-state masses under both hypotheses. For each possible action, it propagates both masses through their respective kernels, splits them by the next public observation, and sums the optimal continuation values. At the horizon, the contribution is half the absolute difference between the total masses. Maximizing before the observation split permits feedback at the next step. The witness policy and its terminal Bayes decisions are saved. The resulting discrepancy and the headline reward-planning values use the same history-dependent policy class.

C.3.4 Interpretation of the measurements

Figure 24 displays all 2,160 reward evaluations, including coincident points. The maximum regret is 0.486440. Every selected policy and every exactly tied model optimum satisfies the adaptive 2Δ bound within tolerance 10^{-10}. The auxiliary metric comparisons summarize each pair by its largest regret among the twelve tasks.

The supplementary metric table includes passive path TV, teacher-forced hidden-state error over three steps, averaged open-loop endpoint TV, observation-occupancy MMD, task value error, sampled policy-ranking error, separate optimum gaps, and endpoint-set distance. Teacher forcing supplies the actual hidden state before each prediction; it is a diagnostic with more information than the deployed agent. The full metric definitions and all pair rows are included in the numerical tables. The exact adaptive calculation supplies the reference discrepancy over the declared policy class. The recovery family explains why a prediction score measured under a passive policy can miss an error relevant to control: that policy never requests the changed response. The rare-event control shows the complementary effect, where a large log loss comes from a small amount of probability mass. The

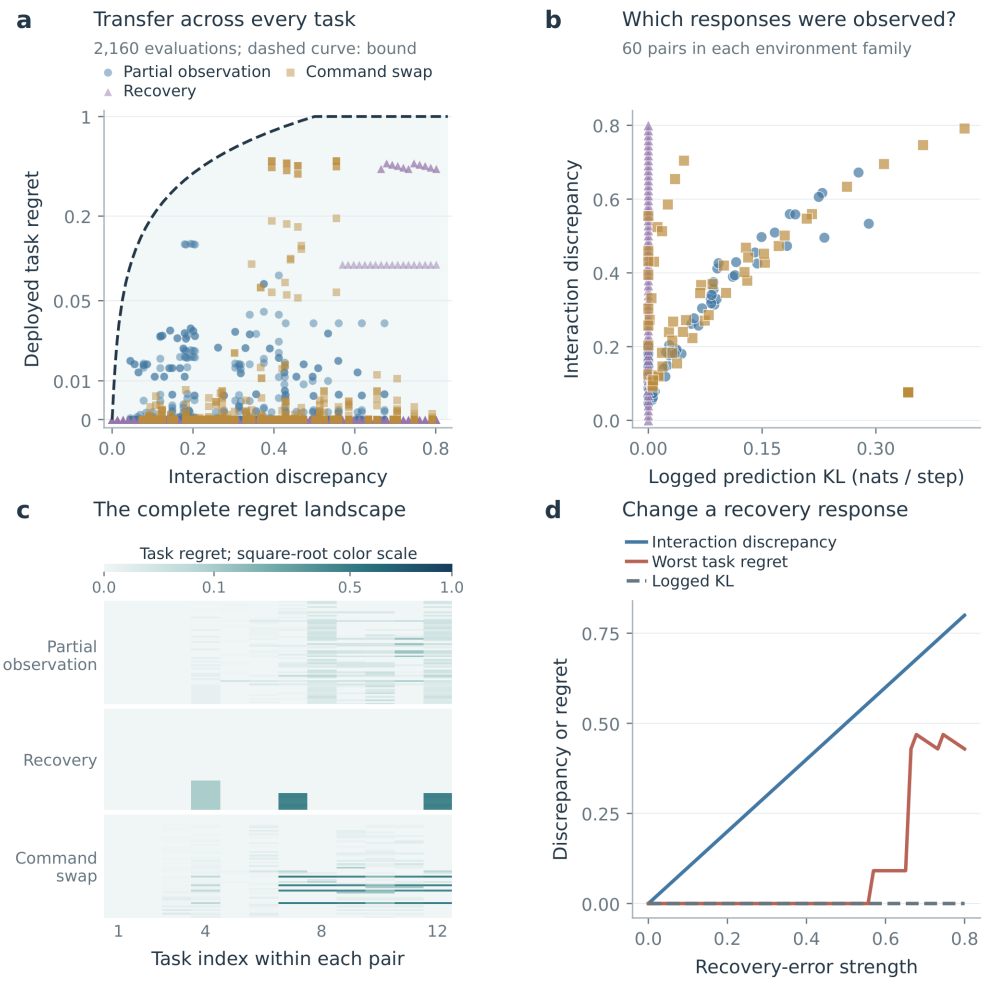


Figure 24: **The prediction score depends on the interaction being evaluated.** a, All 2,160 bounded-task transfer comparisons from 180 model–environment pairs. The envelope is $\min\{1, 2\Delta\}$. An inverse-hyperbolic-sine ordinate resolves small regrets and preserves exact zeros; ticks give untransformed values. b, Logged one-step KL and exact adaptive discrepancy for every pair. c, The complete pair-by-task regret map. Rows group the three environment families and then order pairs by adaptive discrepancy; task indices are local to each pair. The color scale uses square-root normalization on [0, 1]. d, All 60 controlled recovery-error strengths, comparing interactive discrepancy, worst task regret and logged KL. All quantities are exact finite calculations.

response discrepancy retains the policy that exposes an error together with the size of the resulting change in interaction.

C.4 C: PREPARATION, INFORMATION, AND PHYSICAL RESPONSE

A successful grasp can require a suitable starting pose, knowledge of the material, and a tool with the right mechanical response. We represent these requirements in one trajectory law. The experiment asks whether interventions at the corresponding times remove the terms identified by the causal projection. A second calculation asks how the answer changes when only the final outcome distribution is requested.

C.4.1 CHRONOLOGY AND PHYSICAL KERNELS

A path records, in order, pre-reset pose, material, reset action, operating pose, preparation action, resulting pose, noisy pose observation, sensing action, original material cue, added scanner reading, final action, and outcome. Pose and material are hidden binary variables. Material is uniform. Public decisions use the observations already received. The final outcomes are success, drop, and break.

Preparation changes the probability of alignment. An aligned pose remains aligned with probability 0.96; an unprepared misaligned pose aligns with probability 0.04. A preparation action applied to a misaligned pose uses the condition's preparation reliability. Pose observations have accuracy 0.80. A requested scan gives a material cue with a missing probability of 0.04 and a condition-dependent accuracy when present. Skipping gives the missing symbol. When sensing is requested, the additional scanner supplies a binary reading with its specified accuracy and no missing values. When sensing is skipped, both channels return the missing symbol. The cue and scanner draws are conditionally independent given material and the sensing action. Final actions are gentle grip, firm grip, and a tool.

Table 6: Final response kernels in C. Entries give probabilities of (success, drop, break). The tool column gives its selective endpoint kernel; intermediate tool strengths interpolate from the equal mixture of the two grips.

Pose	Material	Gentle	Firm	Tool endpoint
Aligned	0	$(.82, .13, .05)$	$(.52, .08, .40)$	$(.60, .12, .28)$
Aligned	1	$(.70, .27, .03)$	$(.91, .06, .03)$	$(.80, .15, .05)$
Unaligned	0	$(.40, .42, .18)$	$(.25, .20, .55)$	$(.18, .20, .62)$
Unaligned	1	$(.34, .58, .08)$	$(.54, .36, .10)$	$(.82, .12, .06)$

At tool strength λ, its kernel is $(1 - \lambda)$ times the equal mixture of the grip kernels plus λ times the table's tool endpoint. Thus a tool can change outcome probabilities in a direction that has different value for different materials. Every outcome law and observation channel belongs to the path calculation.

C.4.2 DIAGNOSTIC TARGET AND INTERVENTION PACKAGES

The diagnostic target begins aligned with probability 0.90; the physical initial probability is 0.35. Both use preparation reliability 0.82, pose observation accuracy 0.80, and original cue accuracy 0.80. The target prepares with probability 0.75 and requests sensing. Its final action probabilities are $(.45, .05, .50)$ for material zero and $(.05, .45, .50)$ for material one; the missing-reading rule is $(.25, .25, .50)$. The target uses the selective tool kernel at strength one. The physical baseline has strength zero.

The reset package adds a legal earlier reset to an alignment distribution with probability 0.90, and aligns the target's path prefix and time origin with that preparation. The scanner package exposes material through a perfect additional public channel and expresses the target's final action rule in that channel. The tool package changes the physical tool response to the target's response. The eight combinations and a ninth negative control are evaluated. In the negative control, the original cue accuracy is changed to 0.99 while its target description stays fixed. The resulting response mismatch includes the changed observation channel.

For every condition, the full-path target induces a legal randomized policy by marginalizing its actions over hidden variables at each public history. Direct KL between the target and this policy's physical rollout equals the three-term expression. The baseline is

$$0.801627 = 0.662835 + 0.126389 + 0.012403.$$

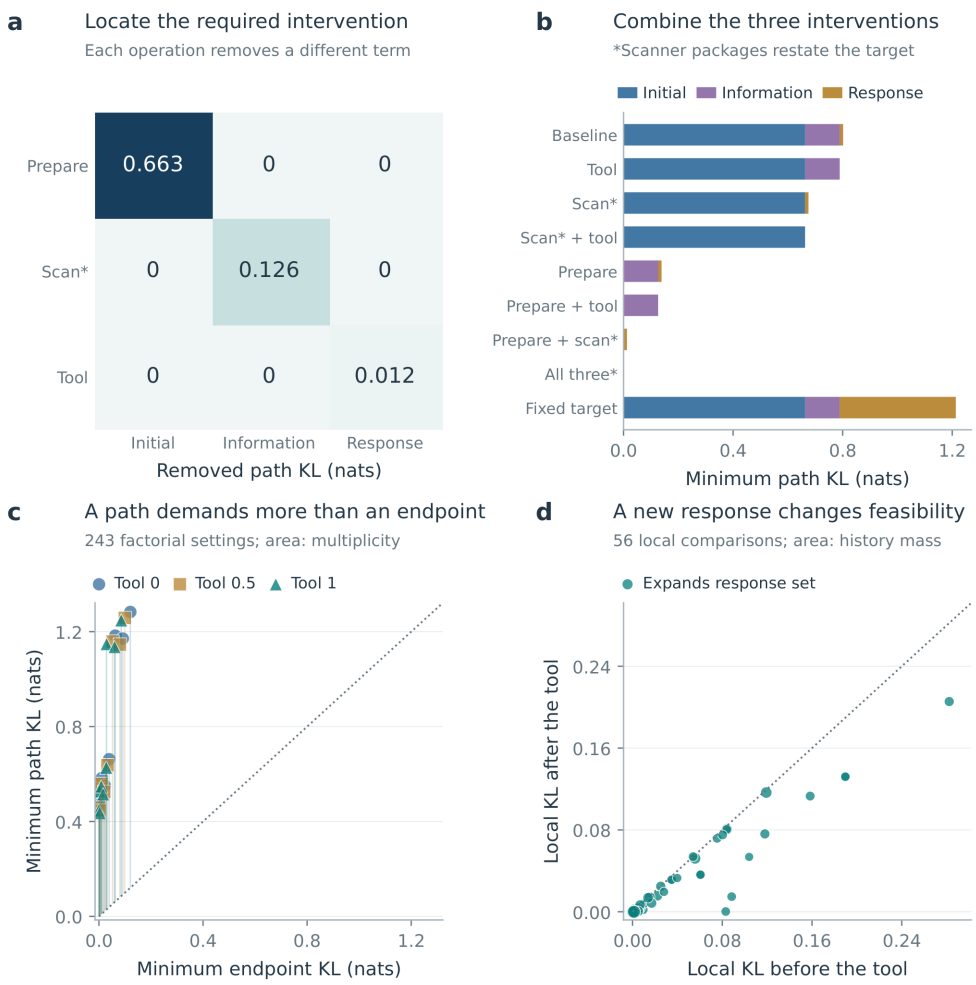


Figure 25: **Preparation, information and response constraints occupy different parts of a path.** a, Each aligned intervention's reduction of the three chronological KL terms. b, The complete decomposition for all nine diagnostic conditions. Scanner conditions marked with an asterisk restate the target in the new observation language; the fixed-target scanner keeps the original target. c, All 243 factorial settings. Symbols identify tool strength and marker area records coincident calculations. Vertical segments show the path cost beyond the endpoint cost. d, Every one of the 56 target–history–tool comparisons. A tool can reduce local KL when its response enlarges the previous set. Axes use nats; the calculations have no sampling error.

The three aligned packages respectively remove these terms. Their combination gives zero. The negative control instead gives 1.213152 nats. The largest absolute discrepancy between direct KL and the component sum is 6.33×10^{-15}.

C.4.3 Endpoint and utility objectives

A second calculation enumerates 1,800 reachable deterministic decision trees for the original preparation–sensing–gripping interface. A tree selects preparation and, after each of two pose observations, either skips sensing and selects one of three final actions or senses and selects a final action at each of three cue values. The count is $2(3 + 3^3)^2 = 1800$. Unreachable decision duplicates are removed. Mixing these trees gives their endpoint-law convex hull. The endpoint projection

minimizes KL over this hull. The full-path projection remains the randomized policy obtained from the target conditionals.

For this second calculation, the target and physical system both start aligned with probability 0.35 and perform no reset. The target prepares with probability 0.75 and requests sensing with probability 0.70. Its preparation reliability is 0.82, pose and material cue accuracies are 0.80, and the additional scanner has accuracy 0.50. Conditional on the resulting hidden pose and material, its final action probabilities for (gentle, firm, tool) are

	Material 0	Material 1
Aligned	$(.88, .04, .08)$	$(.05, .88, .07)$
Unaligned	$(.87, .03, .10)$	$(.04, .16, .80)$.

The target's tool strength is one. These probabilities specify the requested full law, including action dependence on hidden conditions. The projection supplies the closest policy restricted to public histories.

A $3^5 = 243$ grid crosses physical preparation reliability $(.10, .55, .82)$, cue accuracy $(.50, .70, .90)$, tool strength $(0, .5, 1)$, sensing cost $(0, .04, .12)$, and tool cost $(0, .04, .12)$. Physical pose accuracy remains 0.80 and additional scanner accuracy remains 0.50. Preparation costs 0.03. Outcome utility is $(1, .15, 0)$ before action costs. Each cell records full-path KL, endpoint KL, maximum net utility, and the range of preparation, sensing, and tool use among tied optima.

A separate eight-condition comparison crosses preparation reliability $.10$ or $.82$, cue accuracy $.50$ or $.80$, and tool strength zero or one. Preparation, sensing, and tool costs are $.03$, $.04$, and $.04$. With all three improved responses, endpoint KL is 0.000075 and full-path KL is 0.418364. An endpoint request leaves open how the result was reached. The full target also prescribes which actions accompany each hidden circumstance and which transitions occur along the way, so it can retain a substantial realization cost even when its endpoint distribution is nearly attainable.

C.4.4 A TOOL'S EFFECT AT EACH OBSERVED HISTORY

The local comparison considers every final public history with positive target probability across the target specifications above, including the versions with a reset prefix and an added scanner. For the second target we use the baseline physical condition, with preparation reliability 0.10 and cue accuracy 0.50. Each diagnostic target uses its corresponding reset and sensing condition before the tool upgrade. At a given history, we compute the physical posterior over pose and material, then average each final action's outcome kernel under that posterior. The two grip distributions span a line segment in the outcome simplex. Adding the selective tool gives their convex hull with a third point. For the target's conditional outcome law, we calculate the minimum KL over the old and enlarged hulls, retaining the optimizing mixture weights and distributions. A Euclidean projection of the tool distribution onto the old segment determines whether the tool adds a response outside the previous set. This produces all 56 target–history comparisons in Figure 25.

The local gain depends on the direction of the requested change. A tool can add an available distribution while leaving a particular target's nearest distribution unchanged. Together with the chronological decomposition, this calculation identifies both the stage at which an intervention acts and the target requirements it can satisfy.

C.5 D: LEARNING FROM A REQUESTED FUTURE

A demonstrator can succeed through a useful action, favorable hidden conditions, or a favorable disturbance. Selecting successful episodes changes the frequencies of all three. This study varies how much success depends on action, then compares the exact legal-policy projection with policies fitted from finite demonstrations. The comparison distinguishes the cost of learning the action rule from the residual cost of reproducing the conditioned trajectory law.

C.5.1 GENERATIVE PROCESS

Hidden wind $W \in \{-1, 1\}$ is uniform. The first public cue S agrees with W with probability 0.70; a subsequent cue S' agrees with probability 0.85. The two cues are conditionally independent given

wind. The first action observes S. The second action observes (S, A_0, S'). Both actions take values $-1, 1$.

A public demonstrator follows S with probability 0.72 and follows S' with probability 0.78. A privileged demonstrator instead follows the hidden wind with probability 0.90 at each action. Let B be an independent Bernoulli variable of mean β, and let E be an independent fair Bernoulli variable. The terminal label is

$$G = B\,\mathbf{1}\{A_0 = W,\ A_1 = W\} + (1 - B)E.$$

Thus β controls which outcome mechanism is selected. It takes the eleven values $0, .1, \ldots, 1$. The full path retains $(W, S, A_0, S', A_1, B, E, G)$. Conditioning demonstrations on the requested G defines the target family. Initial public cues can be fixed first and the resulting identities averaged over those cues.

C.5.2 Exact projection and fitted policies

The exact projection is $P(A_t \mid H_t, G)$ under each demonstrator law. For a legal public demonstrator, the mean projection cost obeys Equation (13). For the privileged demonstrator, the additional baseline information cost in Equation (15) applies. Changing demonstrator information also changes the conditioned target family.

Each fitted learner receives 4,000 demonstrations for each condition and seed. The five seeds are $11, 23, 37, 53, 71$. Common random numbers couple the declared conditions. Conditional behavioral cloning estimates action counts at each (H_t, G) with Laplace pseudocount 0.5, clips action probabilities below 10^{-5}, and uses probability 0.5 at an unseen history. The return-conditioned MLP uses a single hidden layer of 20 tanh units, Adam with initial learning rate 0.01, a maximum of 300 iterations, and regularization coefficient 0.001. Inputs contain the encoded public history and requested label.

The third learner reweights the conditional cloning probabilities by an outcome association computed from the same data. For a history h, action a, and label g, its unnormalized score is

$$\widehat{\pi}_{\mathrm{BC}}(a \mid h, g) \exp\left\{ \mathrm{clip}\left(\log \frac{\widehat{P}(g \mid h, a)}{\widehat{P}(g \mid h)}, -3, 3 \right) \right\}.$$

The label probabilities use the same Laplace pseudocount. Scores are normalized and floored as above. This is a finite association baseline. Its label probability measures an observational association in the demonstrations. After fitting, population rollouts enumerate every path in the physical process. There are 330 fitted policies across the two demonstrators, eleven values of β, three methods, and five seeds.

C.5.3 Outcome success and path agreement

For public demonstrations, exact projection KL is 0.693147 at $\beta = 0$ and 0.409319 at $\beta = 1$. At $\beta = 1$, the fitted cloning, MLP, and association policies respectively have path KL 0.411618 ± 0.000621, 0.414021 ± 0.001475, and 0.474108 ± 0.008001, with sample standard deviations over five seeds. Their success-probability deficits under a requested success are 0.359085, 0.357339, and 0.350365.

The requested-success deficit is $1 - P_{\mathrm{rollout}}(G = 1 \mid \text{requested } G = 1)$. The overall label-match rate additionally averages over requests for failure and success using their target frequencies. The association policy improves the label match while increasing full-path KL. A label records the final outcome; the full path also records the hidden wind, cues, actions, and chance variables accompanying it. At $\beta = 0$, success is an independent fair coin, so selecting successes fixes a chance outcome that no action controls. The exact projection cost is then $\log 2$. At $\beta = 1$, actions can affect success and the cost falls, while partial observation still constrains which action can accompany each wind. The residual thus changes with the causal route to success. Fitting the conditional action rule removes estimation error in that rule; reproducing the conditioned physical circumstances remains governed by the process.

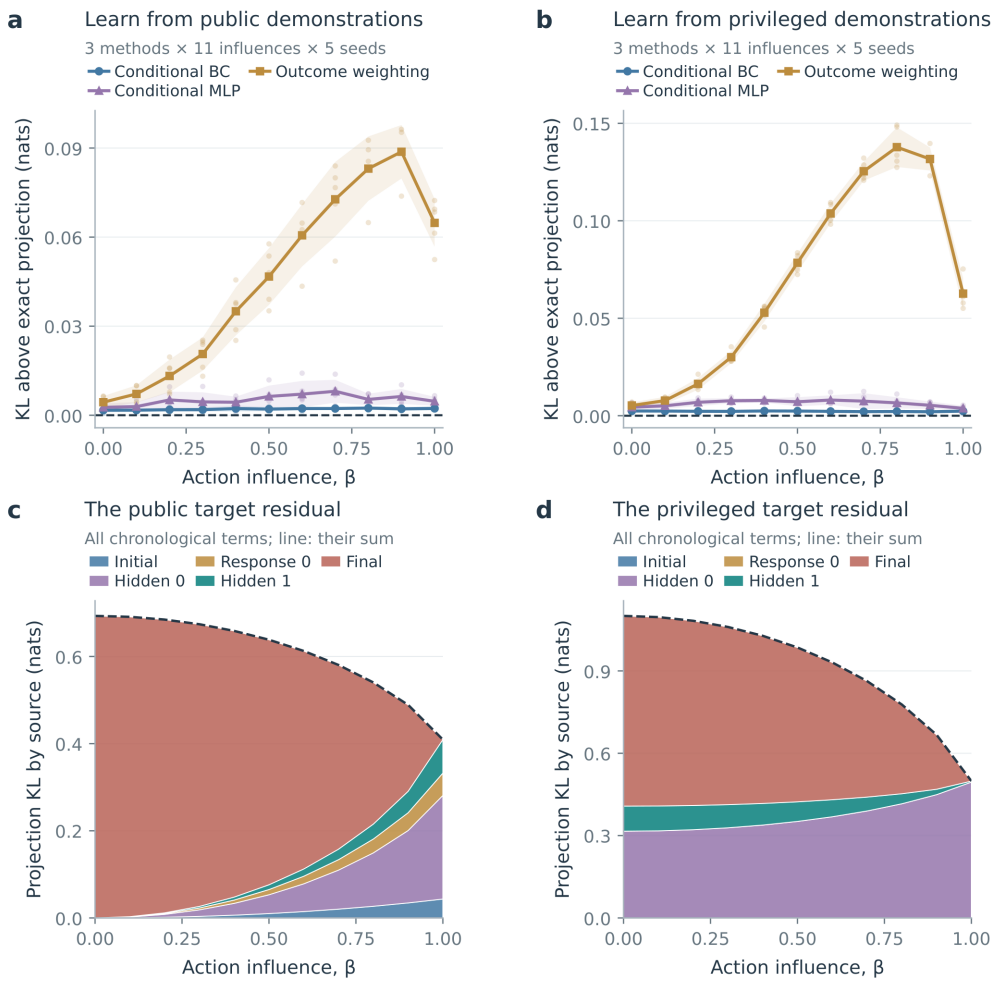


Figure 26: **Conditioning on success leaves a causal residual.** a,b, All 330 fitted policies, split by public and privileged demonstrations. Curves show method means across five fitted seeds; bands show one seed SD and small marks retain individual fits. The ordinate subtracts the independently computed exact projection, placing the optimum at the dashed zero line. c,d, Every term in the chronological decomposition of that projection. Their sum equals the exact residual. Each demonstrator's information defines its own conditioned target family. These panels distinguish finite estimation error from the information and physical responses required by the target.

C.6 E: FINITE LOGS, COVERAGE, AND HIDDEN ACTION SELECTION

Training logs determine which responses a prediction objective measures. Here the same public action frequencies can arise from public randomization or from a demonstrator using an extra hidden signal. We independently vary this selection mechanism and the coverage of the actions needed at deployment. The task has three stages so that information gathered early changes the context in which later actions are chosen.

C.6.1 THREE-STAGE ENVIRONMENT

The task contains inspection, route choice, and a final safe or fast action. A hidden mode $M \in \{0, 1\}$ is uniform. Inspection produces a noisy mode cue; skipping produces a missing cue. After route choice R, a hidden action-time signal S is sampled. Its probability of agreeing with the mode is the route reliability for $R = M$, and the mismatched reliability otherwise. The agent then receives a

noisy public reading of S. The final outcome depends on (M, S, A) and has three values: success, failure, and damage.

The 128 independently sampled environments use seed 20263826 and identifiers starting at 10000. Inspection accuracy is uniform on $[.72, .90]$, matched route reliability on $[.80, .92]$, mismatched reliability on $[.12, .28]$, and public signal accuracy on $[.58, .68]$. Each outcome row is a Dirichlet draw with concentration 220 times the corresponding vector below. These parameters specify a distribution of finite environments, rather than sampling uncertainty about a single fixed transition matrix.

Table 7: Dirichlet mean outcome probabilities for E, ordered as (success, failure, damage). Each environment independently samples every row.

Mode M	Signal S	Safe action	Fast action
0	0	$(.58, .36, .06)$	$(.35, .30, .35)$
0	1	$(.82, .15, .03)$	$(.94, .04, .02)$
1	0	$(.50, .30, .20)$	$(.22, .25, .53)$
1	1	$(.75, .17, .08)$	$(.78, .07, .15)$

C.6.2 Policies and data conditions

The policy library crosses three inspection probabilities $(.2, .5, .8)$, four route maps, and four final-action maps. The route maps specify the probability of route one for cue zero, cue one, and missing:

$$(.2, .8, .5), \quad (.8, .2, .5), \quad (.2, .2, .2), \quad (.8, .8, .8).$$

Final maps specify the probability of the fast action for public signal zero and one:

$$(.2, .8), \quad (.8, .2), \quad (.2, .2), \quad (.8, .8).$$

There are $3 \cdot 4 \cdot 4 = 48$ randomized feedback policies. The fixed reference policy has inspection probability 0.5, the first route map, and the first final map. It is policy 16 under the enumeration order.

All loggers use the reference policy for inspection and route choice. In matched coverage, the marginal fast-action probability at the final public history follows the reference map. In poor coverage, it follows the reversed map. Under public randomization, that probability is implemented with a public-history conditional coin. Under hidden selection, the logger also observes S and uses

$$P(A = \text{fast} \mid H, S) = \text{sigmoid}(b_H + \gamma S),$$

where γ is uniform on $[2, 2.5]$. The intercept b_H is solved so that averaging over $P(S \mid H)$ gives exactly the desired marginal action probability at that public history. This construction holds public history–action occupancy fixed between the two selector variants while changing which hidden signal accompanies an action.

For each environment and condition, eight datasets contain 1,000 episodes each. The three fitted models use current observation, full public history, or full public history with an adjustment for the recorded action-time signal. All tabular rows use pseudocount 0.5 on structurally possible outcomes. At a missing row, this gives the uniform law on those outcomes. The adjusted model estimates $P(S \mid H)$ and $P(Y \mid H, S, A)$ separately and forms their mixture. It uses the signal as a training adjustment variable; deployed actions still use the specified public-history policy.

C.6.3 Prediction and deployment quantities

One-step prediction KL is averaged over the three stages under the behavior occupancy. The fixed-reference trajectory KL and TV compare complete laws for policy 16. Restricted Δ maximizes complete path TV over the 48 policies. The score used for deployment is 0.98 times outcome score $(1, .15, 0)$ plus 0.02 for skipping inspection. The model-selected policy maximizes this score within the registry, and its regret uses the best true registry value. Exact path enumeration evaluates every fitted model and gives 12,288 model–condition–dataset comparisons.

Table 8: Full-history model means in E. Prediction is in nats per transition; path KL is in nats per trajectory. TV and regret have their respective normalized probability and score scales.

Coverage / selector	Prediction KL	Path KL	Path TV	Δ_{48}	Regret
Matched / public	.008453	.025360	.073979	.149280	.004819
Matched / hidden	.008396	.029986	.079644	.163225	.016753
Poor / public	.008469	.047131	.101323	.148983	.007325
Poor / hidden	.008364	.066508	.119530	.162515	.018147

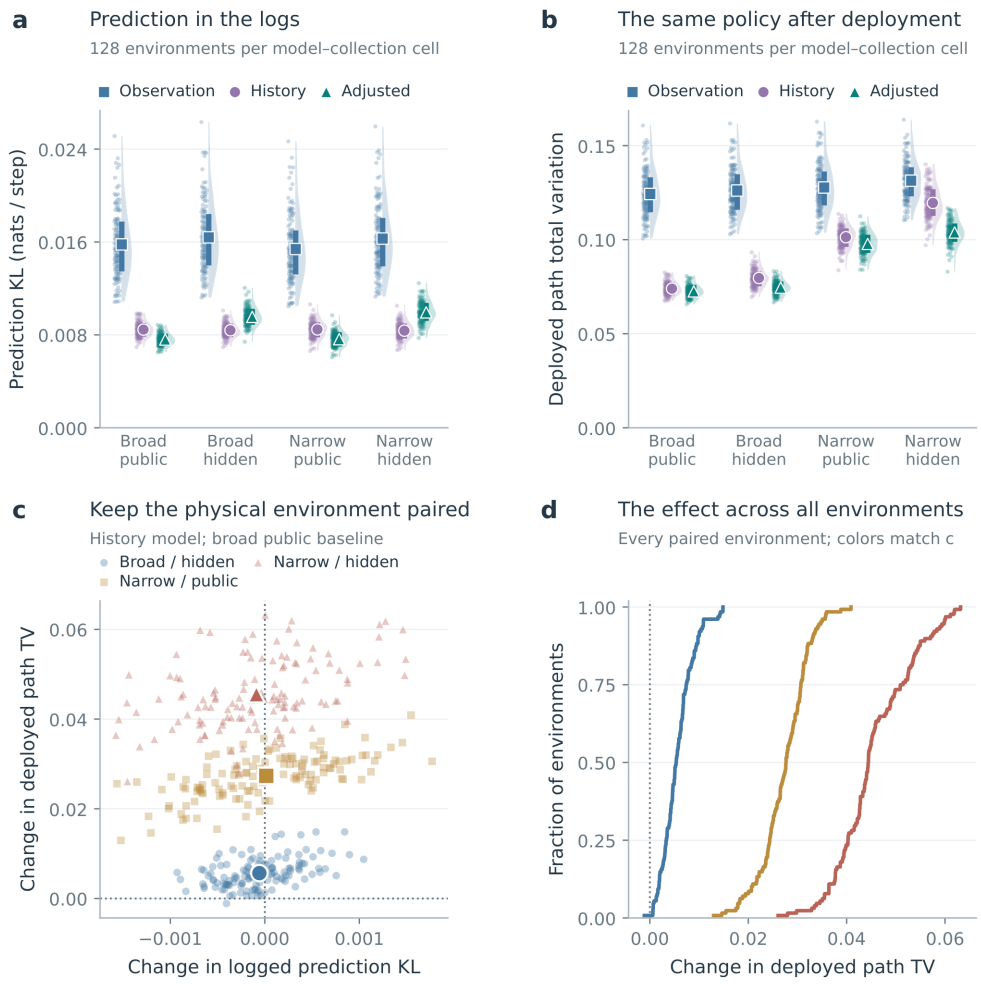


Figure 27: **Data collection determines the scope of a prediction score.** a,b, Prediction KL and deployed path TV for all collection–model combinations. Each small dot is one environment mean over eight fitted datasets. Half violins show their descriptive density, thick segments their empirical interquartile range, and larger markers their mean. All 12,288 fits contribute. c, Changes in the history model relative to broad public collection, paired within each environment. d, The complete distributions of these deployment changes. Colors match c. The comparison retains variation between environments instead of counting repeated datasets as additional independent environments.

C.6.4 Reading the comparison

The full-history model has nearly equal behavior-weighted prediction KL in the four conditions, between 0.008364 and 0.008469 nats per transition. Its deployed reference-path KL ranges from

0.025360 to 0.066508. At matched coverage, hidden selection raises mean regret from 0.004819 to 0.016753 despite preserving the public action frequencies. The hidden signal changes which physical circumstances accompany each recorded action. Consequently, a fitted conditional outcome distribution can describe the logger's selected cases while misrepresenting the response to that action when a public policy chooses it. The adjusted model uses the recorded signal to average response conditionals under the public-history distribution of that signal.

Poor coverage changes a different part of the comparison. It supplies fewer observations for the responses used by the reference policy. The additional-data controls below hold the number of new episodes fixed and change where those observations are collected. These interventions connect an empirical prediction score to the controlled responses its data can identify.

C.6.5 Statistical analysis and additional controls

Each condition mean first averages the eight datasets within an environment, then averages across environments. The 95% intervals use Student's t distribution on these 128 environment means. Primary contrasts concern reference-path KL and deployment regret. Coverage, hidden selection, and their interaction are calculated within each environment. Intervals use 10,000 environment-cluster bootstrap samples with seed 20271826. Two-sided paired t tests receive Holm adjustment across the six primary contrasts. The raw condition and contrast tables specify all denominators.

The sample-only probe uses separate training, validation, and test trajectories. For each domain and candidate policy it uses 2,048 training, 2,048 validation, and 8,192 test trajectories. Validation selects the policy and discriminator; the balanced untouched test accuracy supplies a distinguishing estimate. The confidence calculation uses a one-sided Hoeffding bound with the stated within-environment correction. Such a lower bound detects some discrepancies; it cannot certify a small supremum over all policies. The stored correlations therefore describe the selected finite probe. Added-data controls begin with the same poorly covered public dataset and add 1,000 episodes. One collection repeats final actions on branches that carry little deployment weight; another selects the complementary, deployment-relevant branches. A third adds records in an auxiliary history that the deployment policies never reach. All three comparisons preserve the original observations. In the full-history model, the resulting changes in reference-path KL are -0.003315, -0.034531, and zero, respectively. The paired interval for the targeted-minus-low-weight comparison is $[-0.031919, -0.030514]$ nats after averaging datasets within environments. This isolates the placement of additional responses at a common episode count.

A zero-support control removes a branch entirely and constructs the two-environment witness from Appendix A.8. Its recorded behavior laws are identical. Under the deployment policy, the previously unvisited branch has probability 0.8 and the two environments give different outcomes there, yielding deployment TV 0.8 and a minimax identification lower bound of 0.4. The released tables give both transition kernels, behavior laws, and deployment laws.

C.7 F: compact learned models and controlled prediction channels

A learned visual model can predict several aspects of the same future, while a planner uses particular outputs to choose an action. We make this computational relation explicit with models that predict both an image and operational variables such as position, contact, and success. A background intervention changes the image comparison in a region masked before encoding. A second intervention changes the contact predictions read by the planner. Both are applied to each fitted checkpoint and evaluated on matched physical episodes. Their paired effects show how the route from a prediction to a decision determines the control consequence of an error.

C.7.1 Physical processes and observations

The two processes occupy a 5×5 grid and have four actions: up, down, left, and right. Episodes last eight steps. Navigation starts at $(0, 0)$ with goal $(4, 4)$. Its fixed hidden wind is -1 or 1 with equal probability. At each step, wind affects horizontal movement with probability 0.22, after clipping horizontal displacement to $[-1, 1]$. An attempted move outside the grid produces a collision. Success is absorbing.

Pushing starts with the agent at $(0, 2)$ and the object at $(1, 2)$; the object goal is $(4, 2)$. Moving into the object attempts a push. A valid push fails with probability 0.08 in the physical environment. Invalid moves and unsuccessful pushes produce a collision. Each active step consumes one unit of energy. The public image is a 12×12 grayscale rendering with walls, goal, agent, and object. Navigation includes a wind cue correct with probability 0.72. The recorded operational variables are normalized position, contact, success, collision, and normalized cumulative energy. The input encoder masks the nuisance border while preserving the wind cue.

Data collection is balanced across the two tasks. The concentrated policy chooses the goal-directed action with probability 0.88 and otherwise samples a boundary-safe action. The broader policy chooses the goal-directed action with probability 0.55 and otherwise samples any of the four actions. Dataset sizes are 96 and 384 trajectories. Seeds are $11, 29, 47$. These choices are crossed with three architectures, giving 36 trained checkpoints.

C.7.2 Architectures and training

The image encoder contains two stride-two convolutions, with 8 and 16 channels and 3×3 kernels, followed by a linear map from $16 \cdot 3 \cdot 3$ features to the model width. Rectified linear units follow the convolutions and linear map. The small and large recurrent models have widths 32 and 64. A Gaussian prior is predicted from the recurrent state; a Gaussian posterior also observes the encoded current image. Both latent width and recurrent width equal the model width. Log variances are clipped to $[-6, 2]$. A GRU cell takes the sampled latent, a four-dimensional action code, and a two-dimensional task code. Sigmoid heads predict six operational variables and 144 pixels.

The Transformer has width 32, one layer, four attention heads, a feedforward width of 64, and zero dropout. Each input token combines the encoded image, action, and task. A causal mask blocks future tokens. Linear sigmoid heads produce the same operational variables and pixels. It has no stochastic latent KL term.

Training uses 60 epochs, batch size 64, Adam learning rate 0.002, and gradient norm clipping at 2. Continuous operational variables use squared loss and binary variables use binary cross-entropy, both with weight one. Pixel squared error has weight 0.35. The recurrent posterior-to-prior KL has weight 0.002. Recurrent latent samples are used during training. Prediction evaluation uses 96 held-out trajectories and preserves the distinction between posterior predictions with supplied observations and autonomous prior rollouts.

C.7.3 Planning and physical evaluation

The learned controller enumerates the four one-step candidates and maximizes predicted success weighted by 2, less collision weighted by 0.2, remaining distance weighted by 0.1, and energy weighted by 0.015. It replans after the next physical observation. Its recurrent imagination uses the prior after initialization. The comparison class contains this controller and three fixed public controllers: balanced coordinate progress, horizontal-first progress, and vertical-first progress. They share the stated pushing rule. The fixed controllers use publicly recoverable positions and the task; none observes navigation wind directly.

Each controller runs 64 physical episodes per task with matched episode seeds. Utility is terminal success minus 0.05 times collision count and 0.01 times energy. For an eight-step episode it lies in $[-0.48, 1]$. The reported overall regret first averages each controller's estimated value over the two equally weighted tasks, then subtracts the learned controller's value from the largest class value. Task-specific gaps separately allow different best controllers and are reported as separate quantities. Paired bootstrap intervals resample matched within-task episode indices 500 times. Raw episodes, controller labels, and task labels are retained.

C.7.4 Two changes to the trained model

Rendering nuisance fills the masked top and right image border with independent values from $[0.75, 1]$. The encoder input after masking, physical trajectory, and operational response remain paired with baseline. This change tests a channel that contributes to pixel error but is excluded from the operational encoder by construction.

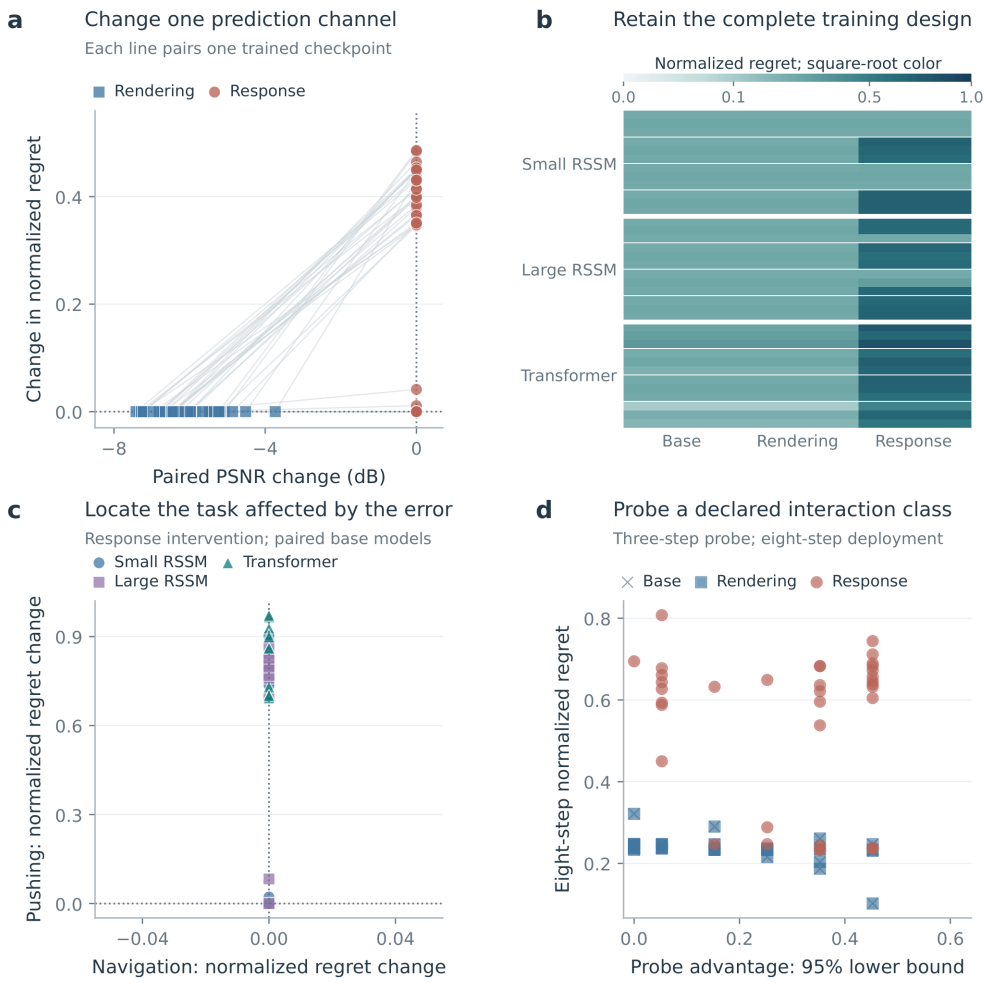


Figure 28: **Rendering accuracy, physical response and interactive probes can change separately.** a, All 72 paired rendering and response perturbations, each relative to its own trained checkpoint. Segments join the two interventions within a checkpoint. b, The complete 36×3 matrix of checkpoint–condition regrets, grouped by architecture and model size. Square-root color normalization on [0, 1] resolves small errors. c, Paired response-channel changes in navigation and pushing. The perturbation acts on pushing responses. d, Every checkpoint–condition combination, comparing a three-step discriminator lower bound with eight-step deployment regret. Regret is normalized by the utility range 1.48; physical evaluation uses 64 episodes per task. These are controlled visual environments.

The response perturbation changes the learned pushing response on contact or collision branches. Put $s = .65\ \mathrm{clip}(p_{\mathrm{contact}} + p_{\mathrm{collision}}, 0, 1)$. The perturbation subtracts $0.35s$ from predicted horizontal position, multiplies predicted success by $1 - s$, and adds $.65s(1 - p_{\mathrm{collision}})$ to predicted collision, with valid-range clipping. Planning uses these predictions. Pixel predictions and the physical environment are unchanged. The navigation response law is also unchanged. These interventions act on the separate heads of the specified model and expose the different information in the associated measurements.

Mean PSNR is 15.003909 dB at baseline and 8.857979 under the background intervention. Mean estimated overall regret is 0.350054 at baseline, remains 0.350054 under the background intervention, and is 0.808420 under the response perturbation. Dividing by the utility range 1.48 gives an increase of 0.310 in normalized regret for the response intervention. The background changes the image score while preserving the information and predictions used for action selection. The contact perturbation

changes that selection pathway. These paired changes expose the relation between each evaluated prediction channel and this planner's decisions. Numerical tables retain raw utility differences, which can exceed one because utility ranges from -0.48 to one.

C.7.5 INTERACTION PROBES AND VIDEO DISTANCE

Three probing policies repeat right, cycle through actions, or respond to contact and collision. They run for three steps and use the recorded operational feedback. For each policy, 48 task-balanced environment/model trajectory groups are generated with independent domain random streams. The group split is stratified by task: 60% training, 20% validation, and the remainder test. A recurrent trajectory discriminator with hidden width 16 trains for 24 epochs at learning rate 0.008; a second discriminator uses terminal operational variables with logistic regression. Neither receives task identity as an input. Policy and discriminator selection use validation accuracy.

For the selected probe, let $\widehat{p}$ be untouched test accuracy and n the number of test trajectories from both domains. The reported one-sided confidence correction at $\alpha = .05$ is

$$\max\left\{0,\ 2\left[\widehat{p} - \sqrt{\frac{\log(1/\alpha)}{2n}}\right] - 1\right\}.$$

It uses independent test draws conditional on the fitted classifier and the task-stratified sampling design. Environment and model draws have separate random streams within each group. Selection does not revisit the test data. The resulting lower bound concerns the specified three-step operational-record experiment. It does not upper-bound the eight-step discrepancy needed for a planning guarantee.

The base, nuisance, and response conditions have descriptive probe–regret Spearman correlations -0.528486, -0.528486, and 0.015657. Under the response intervention, the probe quantity increases in 25 checkpoints, is unchanged in seven, and decreases in four. This heterogeneity motivates keeping the probe, its horizon, and deployment evaluation together.

The supplementary video comparison uses 32 paired environment-rendered videos with seed 9107, an eight-step source sequence resampled to 16 frames, and the bundled frozen I3D Kinetics-400 feature extractor. Frames are resized and replicated to the required three-channel input. FVD uses the difference of feature means and the covariance expression of Unterthiner et al. (2018). Its values are zero for baseline, 403.997 for rendering nuisance, and zero for response perturbation. These videos isolate the rendering intervention. They provide a single paired video comparison, separate from per-checkpoint learned pixel predictions. The bundled model notice identifies the feature extractor.

C.8 G: TRANSITION LAWS AND THE PHYSICAL INTERFACE

An accurate prediction of the available actions can coexist with a requested outcome distribution that those actions cannot produce. We keep the target fixed and cross model accuracy with three nested physical interfaces. Exact optimization separates the error caused by selecting a mixture in the model from the remaining distance to the physical response set. Additional target directions and public branches test how this geometry depends on what the request specifies.

C.8.1 AVAILABLE KERNELS AND A REALIZABLE SKILL

The three primitive outcome kernels are

$$K_1 = (.8, .1, .1), \qquad K_2 = (.1, .8, .1), \qquad K_3 = (.1, .1, .8).$$

The base interface exposes the first two. The macro interface also exposes the third. The skill interface additionally exposes a fixed feedback procedure. Its provisional outcome is distributed as $(.45, .45, .10)$ and is observed exactly. At provisional outcome zero, a repair command produces $(1/9, 2/9, 2/3)$. At outcomes one and two, the procedure holds the observed outcome. Its aggregate kernel is therefore

$$K_4 = .45(1/9, 2/9, 2/3) + .45(0, 1, 0) + .10(0, 0, 1) = (.05, .55, .40).$$

The physical repair command also has rows $(.2, .8, 0)$ and $(.1, 0, .9)$ at the two other provisional states; the hold command has the identity matrix. The stated skill uses only the first repair row. This construction makes its internal observation and feedback explicit.

The planning experiment has one coarse decision. The skill contains two physical stages, provisional evolution followed by observation and action. Its coarse availability is compared at this specified interface. Treating the internal stages as primitive time requires the duration and state augmentation described in Appendix A.6.1.

C.8.2 KL AND TV PROJECTION ALGORITHMS

For kernels collected as rows of a matrix K and mixture weights λ, the realized law is $p = K^\top \lambda$. The KL problem minimizes $\sum_j q_j \log(q_j/p_j)$ on the simplex of weights. The implementation uses deterministic initial candidates and eight additional seeded starts, with optimization tolerance 10^{-11}. Its solution is checked through normalization, attained objective, and feasible directional derivatives. At a finite optimum, the directional condition can be written $\sum_j q_j K_{aj}/p_j \le 1$ for every primitive row a. A mixture draw of 60,000 outcomes checks the sampled law against the exact projected law; a TV tolerance of 0.015 is used for that sampling check.

TV projection is solved independently by a linear program. Auxiliary variables u_j satisfy $-u_j \le q_j - p_j \le u_j$, the weights are nonnegative and sum to one, and the objective is $\frac{1}{2}\sum_j u_j$. This produces both the physical TV floor and the model-selected TV mixture. A sampled discrepancy from a KL-optimal mixture remains a sampling check on that mixture.

C.8.3 A FIXED TARGET ACROSS FIDELITY AND CAPABILITY

The target is $q = (.08, .22, .70)$. A model of each interface replaces every true kernel by $(1-s)K_a + s(1/3, 1/3, 1/3)$, at smoothing values $s = .25, .08, 0$. The one-step discrepancy is the largest row TV, giving $0.116667, 0.037333, 0$. Randomizing actions cannot exceed the largest row discrepancy, and a deterministic action attains it. For each of the nine cells, a TV linear program selects the model-optimal mixture and evaluates that exact mixture using the physical kernels.

Table 9: Exact fixed-target TV in G. Each column evaluates the mixture selected at the specified model smoothing. The last column is the independently optimized true floor.

Interface	$s = .25$	$s = .08$	$s = 0$	Physical floor
Base	.600000	.600000	.600000	.600000
Base + macro	.037778	.051884	.020000	.020000
Base + macro + skill	.037778	.009855	.006667	.006667

The nine deployed errors satisfy $\chi_E \le \mathrm{TV}(q, P_E^{\pi_M}) \le \chi_E + 2\Delta$, with zero model-planning slack. Their largest numerical residual is 4.44×10^{-16}. At intermediate fidelity, the selected mixture's physical error can be nonmonotone because the optimum changes within the model. The exact-model column reaches the physical floor in every row. This illustrates the difference between a uniform perturbation bound and a monotonic ranking of individual policies.

C.8.4 TARGET DIRECTIONS AND OBSERVED BRANCHES

The tilt sweep starts from the uniform passive law and uses $q_j \propto \exp(\theta v_j)$. The six direction vectors are the three permutations of $(1, -.5, -.5)$ and the three edge directions $(.5, .5, -1)$, $(-1, .5, .5)$, and $(.5, -1, .5)$. Strengths are $0, .75, 1.5, 2, 2.25, 2.75, 3.5$. All six directions, seven strengths, and three interfaces give 126 projection rows. A second six-target set contains the fixed target and $(.20, .25, .55)$, $(.06, .34, .60)$, $(.22, .06, .72)$, $(.04, .12, .84)$, and $(.07, .50, .43)$. Adding the skill strictly lowers KL for four of these six targets and leaves two unchanged.

The branch experiment exposes a public signal with probabilities .85 and .15. Its full-strength target requests K_1 on the common branch and K_3 on the rare branch. A branch selector implements both exactly. A signal-independent mixture matches their marginal average but differs conditionally. At full strength its endpoint TV is below 10^{-15}, its full-path KL is 0.168971, its rare-branch KL is 0.895407, and its path TV is 0.1785. Seven branch-contrast strengths $0, .1, .25, .4, .55, .75, 1$ interpolate from identical branch requests to this full contrast.

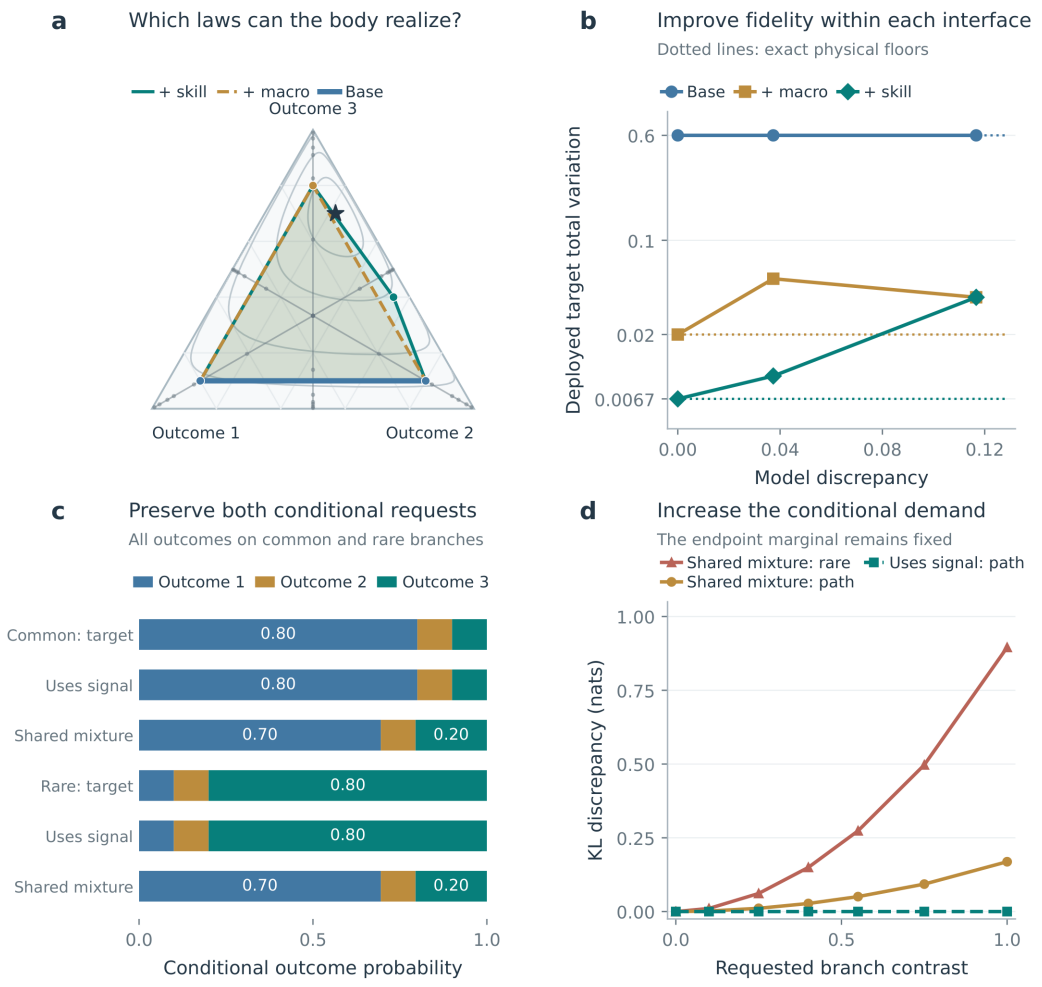


Figure 29: **The available responses define which probability laws can be produced.** a, Nested response sets and all tilted requests in the probability simplex. The star marks the fixed target. Fine contours show its analytic KL objective at 0.05, 0.2, 0.6, and 1.2 nats. b, All nine fixed-target comparisons between model fidelity and physical capability; dotted lines show the exact attainable floors. c, Complete outcome distributions on both public branches. A shared action mixture matches the endpoint marginal while differing from each conditional request. d, All seven branch-contrast strengths. The signal-dependent selector realizes both branches, while the shared mixture accumulates path KL and rare-branch KL. All panels enumerate finite probability laws.

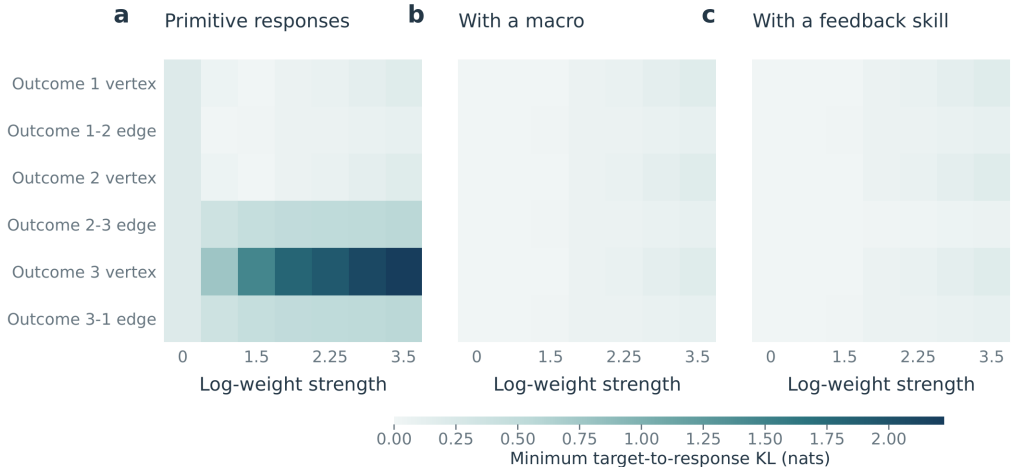


Figure 30: **Feasibility depends on the direction of the requested change.** Each cell is the exact minimum KL from one tilted target to an available response set. Rows retain all six target directions; three column blocks distinguish the primitive, macro and feedback-skill interfaces, each at all seven tilt strengths. The passive law is uniform. The complete $6 \times 7 \times 3$ grid contains 126 calculations. A common scale preserves differences between interfaces; no directions are averaged.

A final independent witness uses target $(.25, .15, .60)$ and available kernels $(.60, .15, .25)$, $(0, 1, 0)$, and $(.25, .35, .40)$. Every action mixture assigns at most 0.40 to outcome three. The target assigns 0.60, giving a TV lower bound of 0.20. The third kernel achieves exactly this distance. This witness has its own interface and is evaluated separately from the fixed-target matrix. The outcome indicator itself certifies the shortfall: every legal action mixture falls short of the requested third-outcome probability by at least 0.20. The branch calculation supplies the corresponding lesson for conditional requests. A signal-independent mixture can reproduce the endpoint marginal while failing to produce the requested association between the cue and outcome. Preserving that association requires selecting among responses after the cue has been observed.

D Connections to theory and contemporary systems

D.1 Controlled interactions and the comparison of models

Stochastic control begins with a physical process and the information available when each decision is made. A policy then determines a probability law over the complete interaction. Strategic measures make these laws the objects of analysis, with information restrictions expressed through their factorization (Bertsekas & Shreve, 1978; Yüksel & Saldi, 2017; Saldi & Yüksel, 2022). The response map retains the association between a policy and its induced law. This association allows a question about a learned model to be stated before choosing its architecture: what happens when the same feedback rule is connected to either system?

This question belongs to a broader tradition of comparing models through the decisions they support. Blackwell's comparison of experiments orders information by its usefulness across decision problems, while approximate sufficiency relates the closeness of statistical experiments to the transfer of decision procedures (Blackwell, 1953; Le Cam, 1964). The comparison here keeps the execution interface fixed. An action has the physical meaning assigned to it by the receiving system, and each feedback rule reads that system's own observations. Under this convention, preserving the consequences of a selected policy is a different requirement from preserving the optimum attained after policies are selected separately in each system. The latter determines the support function of an attainable set; the former preserves the policy correspondence needed at execution.

Value equivalence makes this correspondence explicit for selected policy and value-function classes (Grimm et al., 2020; 2021). Distributional model equivalence extends the preserved object to return distributions, including objectives that depend on risk (Kastner et al., 2023). Our choice of observable determines the corresponding requirement. A return suffices when all intended decisions depend on that return. A transcript containing measurements and subsequent actions can express requirements about conditional behavior, information use, and later composition. Total variation controls every bounded objective measurable on the declared observable. A narrower objective class permits a correspondingly narrower discrepancy.

Simulation lemmas connect local transition errors with errors in policy value (Kearns & Singh, 2002; Lobel & Parr, 2024). The substitution bound starts from a discrepancy on the completed interaction, so its definition already includes the histories visited and the feedback used by a policy. The two contributions to the planning bound arise from evaluating the selected policy and its physical comparator. Appendix A.8 connects this trajectory discrepancy to local prediction errors weighted by policy visitation. The resulting account separates the statistical problem of learning responses from the decision problem of comparing the responses that have been learned.

D.2 Which distinctions a predictive state should preserve

The belief state of a POMDP summarizes the history needed for future controlled prediction (Åström, 1965; Smallwood & Sondik, 1973; Kaelbling et al., 1998). Predictive state representations express the same ambition in terms of future observable tests (Littman et al., 2002; Singh et al., 2004). Bisimulation, approximate state metrics, and approximate information states provide recursive conditions under which compression preserves distinctions relevant to control (Givan et al., 2003; Ferns et al., 2004; Subramanian et al., 2022). These constructions give a functional interpretation to state: two histories can share a description when the permitted future interactions cannot distinguish them in the required way.

The argument using a common continuation in Appendix A.9 makes a consequence of this interpretation explicit. If the same continuation has different response laws at two histories, a predictor that merges the histories incurs at least half their separation on one of them. Enlarging the available continuations can therefore refine the state distinctions needed by the predictor. A new compression skill may reveal a difference between objects whose stiffness had no effect on familiar actions. The earlier tactile observation becomes relevant because the new action makes use of the hidden property. This couples the learning of state representations to the acquisition of skills.

Representation learning for control already uses value or behavioral criteria to discard irrelevant variation. Value-aware losses weight model error through downstream values (Farahmand et al., 2017; Voelcker et al., 2025); visual representations based on bisimulation can omit distractors without reconstructing the image (Zhang et al., 2021); value-equivalent sampling studies which aspects of a model need to be learned for decisions (Arumugam & Van Roy, 2022). Our visual experiment makes the relevant distinction observable: image changes can dominate reconstruction loss while leaving every candidate's task consequences unchanged, whereas errors in a contact response can reverse the selected action. The operative quantity is the error in the comparisons the planner makes.

D.3 Requested trajectories and their causal implementation

Maximum causal entropy treats the order of observations and decisions as part of a trajectory model (Ziebart et al., 2010). Relative entropy projection supplies a general language for finding the closest distribution within a constrained family (Csiszár, 1975). In our projection, the environment fixes preparation and physical response, while the policy controls action conditionals using its available history. Applying the entropy chain rule under this factorization locates a target's requirements at three points in time: before the interaction starts, before a command is chosen, and after that command enters the physical process.

Control based on path integrals and linearly solvable decision problems relates useful trajectories to changes in probability law (Kappen, 2005; Todorov, 2006). Whether such a change can be implemented depends on the stipulated control channel. An agent that selects among fixed action responses has access to their mixtures. The local criterion in Theorem 3 characterizes these mixtures at each history and connects them to realization of an entire requested state trajectory. This identifies

the physical content of a distributional goal: each conditional transition must have an available implementation at the time it is requested.

The distinction matters for learning from successful behavior. Return-conditioned learning depends on coverage and environmental stochasticity (Brandfonbrener et al., 2022). Dichotomy of Control studies how future conditioning can separate controlled effects from exogenous randomness (Yang et al., 2023b). The hindsight projection in Appendix A.5 gives an account of complete trajectories of the same issue. Selecting successful records can favor both favorable initial conditions and disturbances occurring after commitment. The closest legal policy uses the target's action probabilities conditional on the observations available at each decision. The remaining divergence measures requirements placed on preparation, unavailable information, and physical response. Optimizing over hidden completions then gives the corresponding criterion for a target specified only through observed outcomes.

Causal transport and adapted distances also retain temporal information when comparing stochastic processes (Backhoff et al., 2017; Backhoff-Veraguas et al., 2020). Their couplings express which information can be used at each stage. Our realization problem fixes an environment and asks which action conditionals it permits. These perspectives share a reason for retaining chronology: two similar endpoint distributions can belong to different sequential decision problems when the information required to produce them arrives at different times.

D.4 Distributional precision and the time of intervention

Covariance steering studies how feedback can drive a stochastic system between prescribed distributions. For linear systems with complete state observation and suitable control channels, Gaussian endpoint distributions can be reached by feedback laws computed from coupled matrix equations (Chen et al., 2016). The Gaussian result here fixes the last opportunity for correction and solves the projection over all randomized controls available at that opportunity. The residual disturbance after correction determines a covariance floor. Its principal directions determine joint feasibility, so the marginal variances alone can miss an unattainable correlation pattern.

The sensor calculation has the same interpretation. An observation received before correction changes the conditional uncertainty that remains to be controlled. An observation received afterward can refine knowledge of the realized outcome while leaving that outcome's distribution fixed. Blackwell's ordering describes the potential usefulness of an observation in a decision problem (Blackwell, 1953); the timing calculation identifies the decision at which this usefulness can be exercised. Adding another corrective action changes that decision problem and can lower the earlier floor.

Dual control further allows actions to change the observations available for future decisions (Klenske & Hennig, 2016). A contact probe can reveal friction, alter the object's pose, and consume time needed for correction. All three effects enter the law of the complete probe-and-correction policy. The relevant prediction is therefore the joint consequence of learning and acting. The Gaussian case isolates one part of this dependence by fixing the observation channel and varying when its output becomes available.

D.5 Skill libraries, evaluation error, and composition

Options describe temporally extended policies through their internal decisions and termination rule (Sutton et al., 1999). Generalized policy improvement establishes how a collection of evaluated policies can support improved behavior and how approximation error enters its guarantee (Barreto et al., 2017). Selecting among imperfect estimates also has a long history in decision analysis and reinforcement learning: maximizing estimated values can favor alternatives with optimistic errors (Smith & Winkler, 2006; van Hasselt, 2010).

The result on skill expansion places this selection mechanism inside a physical response model. What matters for a switch is the error in the new skill's predicted advantage over an incumbent. A common cost bias cancels from that comparison. A difference in errors can overwhelm a predicted improvement and select the physically inferior skill. Thus an expanded repertoire creates two learning requirements at once: the state should retain properties that affect the new responses, and evaluation should compare those responses under the contexts in which deployment will choose between them.

The skill expansion study varies the available skills while keeping the predictor fixed, making these requirements visible separately from the additional physical capability.

Composition adds another requirement. Two skills can terminate at the same visible pose and leave different hidden contact modes. Equality of their visible endpoints then leaves the response to a subsequent lift undetermined. Appendix A.6.1 gives a constructive example and conditions for composing local errors. Recording duration, failure, resources, and terminal information makes the exported skill interface informative about the next interaction. Local feedback can also remove distinctions from that interface: a tactile controller that compensates for stiffness internally can expose a stable response to a planner that does not itself estimate stiffness.

The behavioral approach to control studies systems through their admissible signals and interconnections (Willems, 2007). This is a natural setting for the same compositional question. Empowerment measures how many effects an action channel can distinguish (Klyubin et al., 2005). A response map additionally retains the particular outcome distributions and the policies that produce them, allowing feasibility and transfer to be tested for a specified task.

D.6 Connections to current world-model architectures

Learning behavior inside a predictive environment. World Models, Dreamer, IRIS, and DIAMOND provide learned environments for selecting or improving behavior (Ha & Schmidhuber, 2018; Hafner et al., 2025a; Micheli et al., 2023; Alonso et al., 2024); Dreamer 4 scales the training of agents inside such models (Hafner et al., 2025b). An improving policy changes the distribution on which the model is used. A rare recovery branch under routine data collection may become common under the new policy. Model-based policy optimization addresses related distribution shifts through short model rollouts and data from the current policy (Janner et al., 2019). The coverage identity here makes the evaluated policy's visitation explicit. Experiment I tests whether collecting the branches that distinguish candidate policies improves deployment under a fixed learning budget.

Planning in learned features. DINO-WM plans using predicted visual features, and V-JEPA 2 combines video representation learning with action-conditioned robot prediction (Zhou et al., 2025; Assran et al., 2025). MuZero and TD-MPC2 learn latent quantities used by their planners (Schrittwieser et al., 2020; Hansen et al., 2024). The response criterion asks what a feature needs to preserve for the declared continuation class. A useful test takes histories with similar representations and applies one common diagnostic policy. A systematic difference in the resulting responses identifies information needed for that continuation. When feature vectors are close rather than identical, a quantitative bound also requires a bound on how sensitively the predictor changes with its features.

Generating interactive visual worlds. Genie, UniSim, and Cosmos expose different forms of action or scenario conditioning (Bruce et al., 2024; Yang et al., 2023a; NVIDIA, 2025; 2026). A camera command, an ego action, and an edit to the initial scene alter different parts of the process. Declaring these meanings makes a test of physical correspondence possible. For a generator conditioned on a complete command sequence, prefix consistency asks whether a changed future command can alter the distribution of earlier observations. Appendix A.2.1 gives the causal realization associated with this condition. ForgeWM develops causal training and the alignment of controls with compressed video time steps (Li et al., 2026b).

Evaluation is increasingly organized around interaction. World-in-World compares predictive models through planning and execution in a common feedback interface (Zhang et al., 2026). WorldExam distinguishes visual quality, control adherence, spatial consistency, and the surrounding world's reactions (Yang et al., 2026). Such evaluations instantiate useful parts of a response comparison. The policy class and recorded variables determine the claim each benchmark supports. The theoretical criterion supplies this specification and relates its error to downstream decisions.

Generating actions and choosing when to commit. Diffuser, Decision Diffuser, and Diffusion Policy generate trajectories or action sequences for control (Janner et al., 2022; Ajay et al., 2023; Chi et al., 2023). Their execution interface determines when a new observation can change the next action. Real-time action chunking, correction of action chunks, and DREAM-Chunk address inference and feedback with different mechanisms (Black et al., 2025; Sendai et al., 2025; Chen et al., 2026). Our timing experiment removes model error to isolate the resulting physical restriction.

Additional computation helps an ongoing interaction when it changes a command before the relevant correction opportunity has passed. Internal skill feedback can provide that correction even when a higher-level planner updates slowly.

Joint prediction and action. DreamDojo learns robot world models from human videos using latent proxy actions; DreamZero jointly predicts future video and robot actions (Gao et al., 2026; Ye et al., 2026). LingBot-VA combines prediction with execution, incorporating actual observations during rollout and asynchronous inference (Li et al., 2026a). Riemann-1.0 places observations, robot state, and actions in a causal autoregressive sequence (Sun et al., 2026). Joint architectures bring prediction and behavior into one model. Evaluating an externally specified policy still asks for its conditional consequences, while generating a policy asks for actions based on information available at execution. A shared architecture can support both uses when the command translation and information timing are preserved.

Adaptation and embodiment enter through the same interface. RMA estimates changing physical context from recent experience (Kumar et al., 2021); broad vision-language-action systems learn relationships between instructions, observations, and commands for particular bodies (Kim et al., 2024; Physical Intelligence et al., 2025; NVIDIA et al., 2025). A latent or linguistic action obtains its physical meaning through the controller that implements it. Matched interventions under a fixed translator test that meaning. Observational training also needs the relevant information recorded: action frequencies alone do not resolve a logger's use of hidden cues, as causal confusion in imitation learning demonstrates (de Haan et al., 2019). Together, these connections identify what should be measured when a world model is asked to support a new body, skill, or feedback schedule.

E Computational interpretation in biological cognition

E.1 Motor calibration preserves the consequences of commands

A motor command has meaning through the response of a body in a physical context. Load, contact, posture, and actuator gain all contribute to that response. Forward models predict the sensory consequences of commands, while inverse models use a desired consequence to select a command (Wolpert et al., 1995; Wolpert & Kawato, 1998). The response map extends this relationship to a sequence of decisions: a prediction used to choose a feedback rule should describe the consequences of executing that rule.

A reversible change in command mapping can preserve an organism's attainable movements while changing the movement produced by a familiar command. Adaptation therefore has an identifiable computational role even when the body's overall range of motion is unchanged. Human reaching provides a relevant dissociation. Learning driven by sensory prediction error and learning driven by reward prediction error can produce similar changes in movement with different effects on perceived sensory consequences (Izawa & Shadmehr, 2011). Cerebellar adaptation also depends on sensory prediction errors (Tseng et al., 2007). These findings motivate measuring command calibration and task choice together.

A direct experiment would exchange the labels of two symmetric force channels while preserving their force ranges, disturbance distributions, and targets. On probe trials, a specified command is elicited and its consequence is predicted. On choice trials, either command can be selected to reach a target. Immediately after the exchange, the old command predictions would be miscalibrated even though the same targets remain achievable. Learning can restore successful choice and command prediction at different rates. The joint measurements identify which part of the relation between policies and responses has been acquired. A separate manipulation that increases available force would test a change in capability itself.

E.2 Feedback timing determines what information can change

Biological feedback arrives after sensory transduction and neural processing. The relevant estimate for action concerns the body when a command takes effect. Human reaching responses to perturbations of visual position and motion are consistent with combining delayed observations and a prediction of ongoing movement (Saunders & Knill, 2004). Cerebellar stimulation can produce reaching errors

consistent with an outdated estimate of hand state (Miall et al., 2007). In the oculomotor system, disrupting a thalamic corollary-discharge pathway impairs the updating of frontal visual responses across eye movements (Sommer & Wurtz, 2006). These studies identify contributions of sensory feedback and issued commands to state estimation.

The formal history records measurements when they become available, together with commands already issued. A delayed measurement can carry its sampling time, allowing a predictor to propagate its implications through the intervening movement. An efference copy supplies information about the issued command. Disturbances that occur later remain uncertain until another observation reveals them. This temporal accounting specifies which corrections an internal model can support.

Feedback is also organized around the movement's purpose. Optimal feedback control relates correction to the remaining task cost (Todorov & Jordan, 2002), and responses in primary motor cortex vary with the goal under matched mechanical perturbations (Pruszynski et al., 2014). Franklin and Wolpert found opposite changes in reflex gain for visual discrepancies with the same mean and variance when one persisted to the endpoint and the other disappeared before it (Franklin & Wolpert, 2008). Thus even matched error magnitudes can have different implications for corrective action.

The timing result suggests crossing sensory reliability with the last time a correction can influence the endpoint. A cue can remain perceptually informative after that opportunity has passed. Its predicted effect on the current movement depends on whether the controller can still act on it. Introducing a later corrective pulse provides a further test: information that could previously support only a report or subsequent learning can now improve the ongoing movement. This manipulation identifies the role of the available intervention alongside the accuracy of the sensory estimate.

E.3 A QUANTITATIVE TEST OF INFORMATION AND PRECISION

The Gaussian calculation gives a reference whose parameters can be estimated independently. Let X be a random initial displacement, V a sensor measurement, and W an independent disturbance applied after a single additive correction U. Suppose (X, V) is jointly Gaussian, W is Gaussian, and correction is unrestricted. For the endpoint $Y = X + U + W$, the control $U = m - \mathbb{E}[X \mid V]$ attains mean m and the minimum covariance

$$\Sigma_{\min} = \mathrm{Cov}(X \mid V) + \mathrm{Cov}(W). \tag{27}$$

The conditional covariance is constant under the Gaussian assumption. Appendix A.13.2 proves optimality over randomized controls and derives the projection for a requested Gaussian outcome law. If V arrives after the only correction, the first term becomes $\mathrm{Cov}(X)$.

A manipulandum task could impose a hidden displacement, display a noisy estimate of it, permit a corrective pulse, and then apply a new independent displacement. Cue reliability and arrival time would be crossed with the end of the correction window. Calibration trials would estimate the initial, sensor, and subsequent disturbance covariances. The prediction would then follow from these estimates, without a separate fitted precision parameter for each timing condition.

Recording the complete endpoint covariance tests more than precision along individual coordinates. A target can require tighter concentration along an oblique direction even when its variance on each displayed coordinate appears attainable. A second corrective pulse would test whether a disturbance that was formerly unavoidable becomes correctable once it can be observed. Motor noise in biological systems can depend on the command (Harris & Wolpert, 1998); such a task is represented by the corresponding measured response kernels. The additive experiment isolates the timing mechanism for which Equation (27) is exact.

Perceptual report, movement dispersion, and learning on the next trial measure different consequences of a cue. A late cue can improve report and subsequent calibration while leaving the completed movement unchanged. Animals can also value advance information about outcomes they cannot influence (Bromberg-Martin & Hikosaka, 2009). This preference can enter a broader behavioral objective. The covariance prediction concerns the specified movement endpoint and its available correction opportunities.

E.4 A NEW SKILL CAN CHANGE THE STATE AN AGENT NEEDS

Two histories can predict the same familiar movements and different consequences of a newly acquired skill. Consider visually identical objects whose hidden stiffness was revealed during an earlier contact. Familiar movements produce equivalent responses for both objects. A new compression skill makes stiffness consequential. Choosing that skill now requires access to the earlier observation, through retained state, memory retrieval, or renewed sensing.

Human orbitofrontal activity represents hidden task states that require integrating observations across time (Schuck et al., 2016). Neural population geometry in hippocampus and prefrontal cortex can also support abstract variables and generalization across task conditions (Bernardi et al., 2020). Predictive maps and successor representations provide computational accounts of future consequences under behavior (Stachenfeld et al., 2017; Momennejad et al., 2017). These findings motivate a specific experiment on how changing the available actions changes the information used for prediction.

Subjects would first learn a repertoire under which the two stiffness conditions have matched consequences. Training would then introduce either the compression skill or a new command that duplicates an existing response. Sensory exposure, practice, and reward frequency would be matched. On held-out trials, a common continuation would test whether the earlier cue predicts the new skill's outcome. Behavioral prediction and cross-validated neural readouts could quantify access to this distinction. The bound in Appendix A.9 supplies the computational requirement: merging histories separated by that continuation forces prediction error on at least one history.

The location of feedback makes a further prediction. A compression skill with local tactile feedback may compensate for stiffness and export the same response in both conditions. Its invoking planner can then use a coarser description, provided that the exported state also predicts subsequent interactions. Comparing fixed compression with tactile correction tests where the stiffness information is needed. Appendix F.3 gives the corresponding condition for skill composition.

E.5 SIMULATION AND EXPERIENCE CAN BE ORGANIZED AROUND FUTURE CHOICES

Hippocampal sequences can represent paths toward remembered goals before navigation (Pfeiffer & Foster, 2013). A computational account of replay explains selective memory access through its expected effect on later choice (Mattar & Daw, 2018). The response framework specifies which comparisons matter when such simulation contributes to evaluating actions. A new maneuver can make a rarely visited recovery branch determine the preferred policy. Accuracy under familiar behavior can then coexist with uncertainty about the newly decisive branch.

An experiment could introduce an additional corrective action after subjects have learned a small repertoire. Equal amounts of experience would be allocated either to familiar trajectories or to contexts where the new action and the incumbent have different consequences. Probe trials would estimate both responses from matched starting conditions, and choice trials would measure selection. The theory predicts that improvement should track the reduction in error in the comparison between the new action and the incumbent. Relating prospective neural activity to these comparisons would test whether simulation changes in the contexts where it could improve the subsequent decision.

Success-conditioned memories raise a related question. A successful route may include a favorable starting state and a gate that happened to open after the subject committed to movement. The path projection identifies which parts of that record can be reproduced by a policy. A behavioral task can preserve the route while changing the probability of the later gate opening. Predictions and choices should then be measured for the joint law of movement, gate state, and outcome. The remembered route can propose a candidate, and evaluating that candidate requires the probability of its physical completion.

Active sensing follows the same dependence on future choices. Human exploratory behavior changes with the number of subsequent decisions that can use the acquired information (Wilson et al., 2014). Dual control formalizes actions that change both the physical state and the information available for later decisions (Klenske & Hennig, 2016). A diagnostic push may reveal stiffness while changing pose and consuming correction time. Comparing complete probe-and-correction policies retains these effects together. Equation (28) specifies how information gains instrumental value through the available continuation.

E.6 Preparation, bodily mechanics, and physical inference

Preparation can change future responses before producing overt movement. Motor cortical activity during preparation can occupy dimensions with little immediate effect on motor output (Kaufman et al., 2014). To relate preparation to the response geometry in Appendix A.7.2, one can vary preparation while matching the visible initial state, then apply a fixed set of commands and measure their response distributions. This distinguishes preparatory states through what they enable next.

Bodily mechanics also contribute to these distributions. Compliance, geometry, and contact determine which movements and force profiles a body can produce (Pfeifer et al., 2007). Affordances and the competition among available actions provide complementary descriptions of this dependence (Gibson, 1979; Cisek, 2007). The local mixture criterion gives a quantitative form to it: a requested response is available when the body's admissible action responses can produce the required conditional distribution. Learning a controller and changing a body's mechanics can both change this set.

A probability represented inside the organism has its own physical implementation. Predictive coding and active inference relate perception and action to probabilistic descriptions (Rao & Ballard, 1999; Friston et al., 2010); stochastic neural sampling offers a model in which neural activity represents samples (Buesing et al., 2011). Appendix A.10 considers a binary internal state whose initial distribution represents a prior. Evidence-dependent transition rates can move its endpoint distribution to a requested posterior when the requisite transitions and time are available. The external variable being inferred remains fixed. This construction makes the implementation question explicit: a probabilistic update specifies a target for an internal physical process, whose attainable distributions depend on its dynamics and available time.

F Design consequences and their mathematical basis

The response map provides a common basis for several design choices: which observations to predict, which distinctions a representation should preserve, what a skill should return, and where interaction data are needed. The following sections derive these consequences for a fixed physical interface and explain how they change when the interface changes.

F.1 Predicting information that can affect a later action

The public interaction law contains the observations on which a policy will act. This matters when an action is valuable partly because of what it reveals. A light contact can both displace an object and reveal its friction. The distribution of displacement alone leaves the value of the observation unspecified. Likewise, predicting the right marginal distribution of a cue and the right marginal distribution of a final position leaves their relationship unspecified. The joint law determines whether a subsequent decision can use the cue. Experiment H measures this distinction by changing the cue–target association while preserving cue frequency.

A sufficient information state can retain this information through a compressed history and support the required conditional predictions and value computations (Subramanian et al., 2022). The requirement is that the predicted improvement after a diagnostic action be obtained from evidence that the physical interface can return. This is the dual role of action studied in dual control (Klenske & Hennig, 2016).

The following decision-theoretic identity makes the role of action availability precise. Let Θ denote an uncertain circumstance, let Y be a signal obtained before choosing among finitely many actions, and let $u(a, \Theta)$ be a bounded reward. Acquiring the signal in this comparison has no direct physical effect or cost. The best expected rewards without and with the signal are

$$V_0 = \max_a \mathbb{E}u(a, \Theta), \qquad V_Y = \mathbb{E}\Big[\max_a \mathbb{E}[u(a, \Theta) \mid Y]\Big].$$

For any maximizer a_0 of V_0,

$$V_Y - V_0 = \mathbb{E}\Big[\max_a \mathbb{E}[u(a, \Theta) \mid Y] - \mathbb{E}[u(a_0, \Theta) \mid Y]\Big] \geq 0. \tag{28}$$

The integrand is nonnegative. Thus the signal has zero value exactly when the same fixed action is optimal for almost every posterior induced by the signal. This follows directly from the tower

property and is a task-specific application of the comparison of experiments (Blackwell, 1953). If the signal arrives after the last action, the legal value remains V_0. In a longer interaction, a may index an admissible continuation policy and u its expected remaining reward.

Equation (28) explains a complementarity between sensing and skill. A highly informative sensor can have zero value for a task whose available actions all favor the same choice. Introducing an action whose advantage depends on the signal can make that sensor useful. Conversely, a new corrective action may be useful only after a diagnostic observation. These effects arise from the same conditional response structure. The useful distinction is the one that changes an available action comparison. Sensing costs, delays, and changes in physical state can be included by evaluating the complete chronological interaction.

The sampled world and the controller's information. Suppose a simulator samples Θ from the current belief about an unknown physical parameter. A controller receiving the realized sample can optimize $\mathbb{E}[\max_a u(a, \Theta)]$. A controller receiving only Y achieves V_Y. Their difference can be positive with an exact simulator. The projection in Theorem 2 locates the discrepancy in action dependence on hidden information. An implementation can sample a physical parameter once for an imagined episode, generate the corresponding measurements, and update the controller from those measurements. At each decision, fresh private randomization is independent of the sampled world given the public history. An implicit latent model can implement the same law without explicit physical parameters. Preserving that law also preserves the temporal dependence induced by a persistent unknown parameter; resampling an independent parameter after each frame generally changes it.

F.2 Representation, geometry, and changes in capability

Fix a decision time, two physically reachable histories, a common observation and action interface, a class of continuation policies Π, and a family $\mathcal{F}$ of bounded tests of the future record. This record begins with the continuation, so the earlier observations that distinguish the histories are excluded. A continuation starts with the same controller initialization in both histories and can respond to subsequent observations. The tested difference between histories can be written

$$d_{\Pi,\mathcal{F}}(h, h') = \sup_{\sigma \in \Pi,\, F \in \mathcal{F}} \left| \mathbb{E}^{\sigma}_{E}[F \mid h] - \mathbb{E}^{\sigma}_{E}[F \mid h'] \right|. \tag{29}$$

With all tests $0 \le F \le 1$, this is the largest TV difference under a common continuation. A smaller test class measures the specified consequences. The distance is symmetric and obeys the triangle inequality; distinct histories can have distance zero when all tested continuations agree. It is therefore a pseudometric. Predictive state, behavioral equivalence, and approximate information states supply established constructions with related requirements (Littman et al., 2002; Givan et al., 2003; Subramanian et al., 2022). Bisimulation-based representation learning and approximate value equivalence make corresponding choices of distortion for compression (Zhang et al., 2021; Arumugam & Van Roy, 2022).

If an encoder ϕ and a model using only $\phi(h)$ achieve error at most ε on every declared continuation and test, then

$$\phi(h) = \phi(h') \quad \Longrightarrow \quad d_{\Pi,\mathcal{F}}(h, h') \le 2\varepsilon.$$

For each continuation and test, insert the model's common prediction between the two physical expectations and apply the triangle inequality. Taking the supremum proves the statement. This extends the single-witness argument in Appendix A.9. For nearby latent states, the corresponding conclusion requires a sensitivity assumption. If model predictions are L-Lipschitz in latent distance, then

$$d_{\Pi,\mathcal{F}}(h, h') \le 2\varepsilon + L\|\phi(h) - \phi(h')\|.$$

Such a relationship calibrates latent distance against differences in physical responses.

Enlarging either the continuation class or the test class can increase the distance in Equation (29). Thus exact equivalence classes can split when a skill or a new tested objective is added. A fixed representation remains adequate precisely when its earlier compression retained the newly required distinctions. When the earlier compression discarded a required distinction, new interactions can identify the histories that need separation, and a revised recurrent representation can preserve the relevant evidence. The source of that evidence can be visual, tactile, proprioceptive, or temporal.

Two consequences guide the choice of geometric objectives. Mass and surface properties can determine an action's effect even when the agent cannot change those properties. A representation may need to retain them for prediction. Similarity and reachability also express different relations. The distance in Equation (29) is symmetric. Physical reachability can be directional, as when a contact can be released but cannot be recovered within the remaining time. A planner can use a symmetric embedding distance to score goals and action-indexed response laws to represent the direction of physical reachability.

The same distinction applies to a complete map of attainable outcomes. In finite spaces with randomization, the best terminal value for every reward determines the convex attainable set through its support function. It leaves the assignment of policies to distributions undetermined. Appendix A.3 gives the command-permutation construction. Matching every optimal task value consequently establishes a different property from substituting the same controller between systems. Value equivalence conditions that preserve the specified policy labels and Bellman computations retain that additional correspondence (Grimm et al., 2020).

F.3 Skill composition and the location of feedback

Consider an interface at which the planner invokes a skill u. The skill uses a fixed causal controller, may observe the world during execution, and returns at a stopping time τ. Let Z be a state sufficient for the responses of all permitted skills. A returned record may include the next sufficient state Z', elapsed duration D, and task-relevant quantities B, such as accumulated energy or an indicator that a force limit was exceeded. Its response kernel is

$$J^u(dz', dd, db \mid z).$$

Absolute time and remaining resources can be included in Z'. When the interface permits interruptions or decisions during the skill, the returned stream and the corresponding decision times become part of the public interaction. An endpoint kernel is appropriate only for the declared policy class that waits for completion. These choices make explicit the information requirements behind temporal abstraction (Sutton et al., 1999).

Proposition 12 (Composition through a sufficient returned state)**.** *Suppose the conditional returned-record kernel of each skill depends on its execution history only through its current sufficient state and chosen command. If a model and environment have the same initial-state law and the same kernels J^u, every common causal policy invoking at most N skills induces the same public macro record law. If the initial TV error is at most ε_0 and the conditional kernel error at invocation j is uniformly at most ε_j, then the record discrepancy is at most*

$$\min\left\{1, \varepsilon_0 + \sum_{j=1}^{N} \varepsilon_j\right\}.$$

Proof. At equal macro histories, the common policy chooses commands from the same distribution. With equal response kernels, induction gives equality of the next returned-record laws and hence of the full record. For the approximate statement, maximally couple the initial records. Conditional on agreement so far, couple the commands identically and the next returned records maximally. Disagreement arises initially with probability at most ε_0, or at invocation j with conditional probability at most ε_j. The union bound and the coupling characterization of TV give the stated result. Early termination can be represented by an absorbing returned state. The finite setting guarantees the required couplings; the same argument applies on standard Borel spaces with the corresponding kernels. □

The returned state may need to retain an estimate of mass, an energy budget, or a contact mode because these variables affect the next response. Reporting them together with the task's path quantities permits direct testing of the proposition's premise. For a fixed additive reward, some of these quantities can be replaced by the conditional expectations sufficient for its Bellman computation. Reuse for arbitrary risks, deadlines, or path events requires the corresponding distributions or tests.

The proposition also explains how feedback can support abstraction. Suppose two microstates respond differently to primitive commands, but a stabilizing skill uses its local measurements to bring both

to the same returned-record law. Their higher-level distinction can then be omitted for a planner restricted to such skills and the tested records. The information needed to distinguish them remains inside the skill. This is a redistribution of predictive work between levels. Adding a skill while retaining every old primitive continuation can only refine exact predictive equivalence. Restricting a higher-level interface to robust feedback skills changes the continuation class and can permit a coarser description. Both conclusions follow from specifying which decisions a representation serves.

This offers a design criterion for learned skills beyond endpoint success: the skill should make its future response reliably predictable from its exported state. A long command chunk can delay the use of information that arrives during execution. Asynchronous generation and correction methods address different parts of this interface (Black et al., 2025; Sendai et al., 2025; Chen et al., 2026). Their effect can be evaluated by holding the physical predictor fixed and varying when the next observation is allowed to affect an action, as in experiment H.

F.4 Transfer across bodies through a common interaction

A common task description can conceal different physical command meanings. Suppose bodies A and B expose a common sufficient public state z. Their commands a and b induce kernels K_A^a and K_B^b over the same returned record, including duration and task-relevant observations. A causal translator $G(db \mid z, a)$ implements body A's command through body B when

$$K_A^a(dz' \mid z) = \int K_B^b(dz' \mid z) G(db \mid z, a). \tag{30}$$

With matching initial laws, applying this relation at every invocation transfers every admissible policy on the common interface. The proof substitutes the right-hand side into the chronological factorization and proceeds by induction. Approximate kernels give the composition bound in Proposition 12. When a state lacks the required information, the same condition can be stated on public histories. The translator uses the information actually available to the receiving body.

This condition gives embodiment transfer a constructive meaning. A receiving body can simulate another body's tested responses when those responses belong to its conditional mixture sets. Inclusion of terminal attainable sets alone leaves the command correspondence, intermediate information, and time of execution undetermined. The direction of transfer can also be asymmetric: a richer interface can emulate a restricted one, while a representation learned under the restricted interface may omit distinctions its additional commands require.

A model shared across bodies can therefore retain common predictive structure and learn calibrated response and observation channels for each body. Matching action embeddings can help parameterize the translator; Equation (30) specifies the physical property to test. A low-level feedback controller can supply the translation when a static command map is insufficient, provided its returned information and duration enter the common record. Calibration from interaction history, as in physical adaptation (Kumar et al., 2021), supplies evidence about the channel. The theory leaves the architecture of the shared model open while identifying what its reuse entails: correspondence between executable experiments and their consequences.

F.5 Reliability, uncertainty, and changing the physical problem

A stochastic outcome can contain uncertainty about the present world, uncertainty removable by future observations, and variation introduced after the last useful action. The chronological factorization identifies when each is relevant. These are roles within an interaction: a disturbance that remains uncertain for one interface may be observed and corrected under another. A single scalar uncertainty score leaves this distinction unspecified.

For example, suppose X is an initial Gaussian displacement, V a jointly Gaussian measurement available before correction, and W an independent disturbance that follows it. With unrestricted additive correction, choosing $U = m - \mathbb{E}[X \mid V]$ gives residual covariance

$$\mathrm{Cov}(X \mid V) + \mathrm{Cov}(W).$$

Appendix A.13.2 gives the assumptions and global projection result. Receiving V after the sole correction leaves $\mathrm{Cov}(X)$ in the outcome. Improving the sensor can reduce the first covariance down

to the sensor's informational limit; reducing the later disturbance or retaining another correction opportunity addresses the second. With partial actuation, the covariance constraints also concern correlations across actuated and unactuated coordinates. The target's joint distribution specifies the required dependence as well as its coordinate-wise variability.

The Gaussian projection further shows why uncertainty calibration and goal guidance can conflict in a planner. Below the available precision, narrowing the requested target raises its best possible KL discrepancy. Selecting favorable imagined disturbance samples can conceal this increase. A planner that evaluates candidate policies against the unchanged physical response distribution exposes it. For general dynamics, membership in the conditional response set gives the corresponding criterion. The Gaussian case makes this criterion explicit through the mean and covariance.

This distinction helps allocate learning and control effort. Additional data improve knowledge of a response channel. A better tool, a different contact preparation, or another feedback opportunity can change the channel itself and expand its attainable set. The target projection evaluates the resulting interactions by the same criterion. A world model can then guide both the choice of control under a given interface and the choice of an interface that makes the requested behavior possible.

The three terms of Theorem 2 locate requirements within a fixed interaction. A design intervention can change several terms together: preparation changes the states encountered, and sensing changes subsequent decisions and responses. An observable target also allows its hidden completion to vary (Appendix A.1.1). Evaluating each proposed design through its complete law retains these dependencies.

F.6 Evaluation that identifies the supported decisions

Evaluation can be organized around a declared decision: a controller receives a specified observation history, issues a command through a specified body or translator, and is judged on specified outcomes. The same rule is run in the model and environment. Its action sequence may differ across runs because the observations differ. This comparison tests model substitution at the point where the model is used. Common initial conditions reduce variation in paired comparisons. The comparison concerns the two response distributions; Appendix F.7 treats the additional structure involved in individual counterfactuals.

Theorem 1 gives two complementary tests. A bounded observable task detects a particular discrepancy between interaction laws. A policy optimized in the model tests the downstream effect of the discrepancies the optimizer finds. The latter depends on the search procedure and its available decisions. A more capable planner can discover both better physical opportunities and more optimistic model errors. Evaluating under an expanding search or skill class therefore tests a changing specification. This connects the response criterion with policy optimization in learned environments (Hafner et al., 2025b), feedback-based evaluation (Zhang et al., 2026), and proposed evaluations centered on decision making (Yu et al., 2026).

Data for the decisions being evaluated. The occupancy identity in Appendix A.8 weights conditional prediction errors by the histories visited under the tested policy. It motivates probes near the choices of interest and recovery branches that routine behavior seldom visits. Coverage depends on which interventions are observed. Identifying a relevant branch requires data from policies that visit it. Moreover, if a demonstrator uses unrecorded information when selecting actions, conditioning on the recorded action can select hidden circumstances. Testing the causal response then requires the relevant information or an appropriate intervention design (de Haan et al., 2019).

For a fixed task, the identity in Equation 7 identifies the error that can change a decision. Validation can compare the incumbent and candidate under matched contexts, and assess the sign and size of their predicted advantage. Selection across several candidates calls for a bound that holds simultaneously over their comparisons. The acceptance criterion in Appendix A.12 has this form. Independent validation or uniform statistical bounds can account for adaptive search. Ensemble spread can guide acquisition, with its usefulness assessed through the resulting comparison errors.

What a performance reference establishes. An exact attainable optimum permits separating physical task difficulty from regret under the declared policy class. In a large physical system, a reference controller, a registered bank, and a simulator oracle establish different quantities. A

reference controller supplies a feasible performance level. Exact optimization over an executable policy class gives that class's optimum. Selecting a separate action using each realized hidden state or future disturbance gives an optimistic reference with additional information. Experiment I compares the registered physical responses within each saved context. The relevant reference therefore depends on the information available at selection. Specifying it makes success scores interpretable across changes of body, observation timing, and search budget.

Proposition 8 supplies another route when the goal itself is in question. For a compact convex set of attainable laws, every positive target distance has a bounded task witnessing a shortfall. In practice, a force constraint, a deadline, or a target–outcome association can supply such a candidate test. An upper bound on attainable expected reward establishes whether such a demand exceeds physical capability. The witness expresses that limit through an observable task, independently of which controller is used to search for a solution.

F.7 The empirical meaning of a learned world

The response map has an interpretation independent of internal coordinates. Two models agreeing on every declared controlled interaction are equivalent for those decisions and tests. A change of invertible latent coordinates leaves that empirical content unchanged. Extending the available interventions or tested quantities can separate models that were previously equivalent. This gives the development of a representation an operational meaning: the agent can acquire new distinctions through newly available experiments.

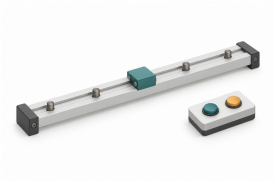
The distinction between interventional predictions and hidden mechanisms can be stated with a single decision. Consider a single decision with two actions and no observation of a fair hidden bit U. One structural model gives potential outcomes $Y(0) = U$ and $Y(1) = U$; another gives $Y(0) = U$ and $Y(1) = 1 - U$. Every legal action has the same Bernoulli outcome law in both models, including every randomized policy in this one-decision interface. The two models disagree on whether changing the action would have changed the outcome for the same individual. The controlled response map determines the outcome law under each intervention. The joint relationship between outcomes of mutually exclusive interventions is an additional property of the structural account. Counterfactual methods that infer alternative outcomes for a specific episode supply additional structural assumptions (Buesing et al., 2019).

For control, sufficiency is assessed against the interventions and tasks the model is asked to support. For a mechanistic interpretation, one also asks what additional observations or interventions distinguish the proposed hidden accounts. The two inquiries can develop together. A new sensor or tool can improve control and expose a distinction on which previously adequate models disagreed. This is the sense in which embodiment contributes to both capability and knowledge.

G ILLUSTRATED EXPERIMENTAL GUIDE

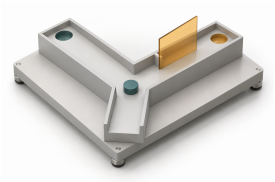
This guide introduces the physical setting and reasoning behind each experiment. The images are conceptual task illustrations generated with ImageGen. The linked methods specify the environments, observations, actions, probability laws, and quantitative comparisons.

COMMANDS AND REQUESTED TRAJECTORIES



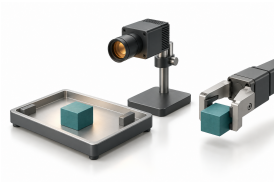
A. The meaning of a command

A noisy actuator moves among five positions. Reversing its command map preserves attainable goals while changing what each command does. Executing selected policies before and after command translation shows why a model needs the correspondence between an input and its physical consequence.



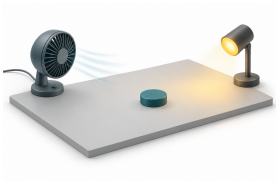
B. A policy selected inside a model

Three finite process families vary command meaning, partial observation, and recovery responses. A planner selects a feedback policy for each of twelve objectives and executes it in the corresponding environment. Exact calculation across 180 model pairs relates the resulting loss to the same-policy interaction discrepancy.



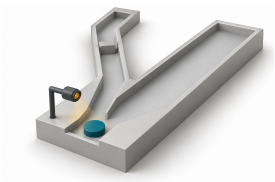
C. The requirements of a requested path

A manipulation sequence combines preparation, material sensing, and gripping. Its requested trajectory specifies favorable starts, informed actions, and physical outcomes. Reset, scanner, and tool interventions reveal when each requirement enters the path and which change to the physical interface can satisfy it.



D. What successful demonstrations contain

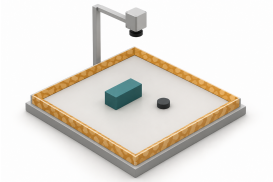
Two decisions use noisy cues about hidden wind. Selecting successful trajectories also selects the hidden circumstances and chance events that accompanied their actions. Comparing fitted policies with the exact legal projection identifies how much of that conditioned trajectory law can be reproduced through actions available to the learner.



E. What training logs reveal

An agent inspects a hidden mode, chooses a route, and selects a safe or fast action. The data vary action coverage and the hidden information used by the demonstrator. Matched public action frequencies reveal how the circumstances accompanying a recorded action affect the physical response learned from it.

PREDICTION, TIMELY INFORMATION, AND SKILL SELECTION



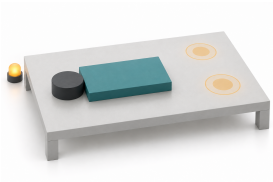
F. The predictions used for a decision

Visual models predict images and operational variables for navigation and pushing. Changing a masked image border affects rendering error; perturbing predicted contact responses affects the quantities used by the planner. Paired deployment traces each intervention through the model's specified prediction and decision pathways.



G. Available distributions of outcomes

Primitive actions and an added feedback skill each produce a distribution over three outcomes. Their mixtures determine which target laws can be attained. Exact projections show how a new response changes this set. A second task asks the chosen outcome to depend on an observed branch, retaining the relation between information and action.



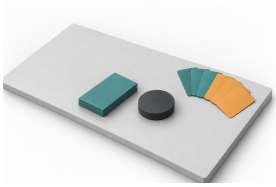
H. A cue arriving during a push

A circular pusher moves a rectangle toward a target revealed during motion. An exact predictor generates a sequence of commands at each permitted update. Changing cue time and update spacing measures when the new target can still change the push. Shuffled cues expose the correspondence between each requested target and the reached destination.



I. Experience for choosing a skill

After a short calibration history, a learned model predicts the results of 27 two-second feedback skills. Matched label budgets compare routine pushes, uniform collection, ensemble disagreement, and deployment probes. Independent validation and held-out physical contexts show how response coverage affects both model selection and skill selection.



Skill expansion with fixed predictions

The same trained models choose from growing subsets of the same physical responses. Individual additions identify the prediction error that changes an incumbent choice. Random expansion orders and all 2,925 three-skill starting banks explain how the initial options and the new comparisons jointly determine the benefit of a larger library.