SAFE DYNAMIC DECISION-REGION RESOLUTION FOR DATA-EFFICIENT WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

An agent can become better informed and still lose its last opportunity to make a justified decision. We ask what a world model must learn offline when some uncertainty can still be resolved safely at deployment, and formalize this as Safe Dynamic Decision-Region Resolution (SD-DRR): offline evidence needs to remove only the decision conflicts that safe online interaction cannot resolve within budget. A decision–nuisance factorization preserves the optimal resolution cost. In finite realizable games with deterministic passive observations, minimal budget obstructions characterize the required evidence, and their covering geometry yields optimal query allocations with matching exponential failure rates. Continuation responses specify the required prediction accuracy, and calibrated cost–risk frontiers extend the analysis to continuous stochastic histories and misspecified neural predictors. In finite games, SD-DRR needs $10.3\times$ fewer passive observations under a 90% retention criterion, and convex query allocation attains the predicted exponent. Response supervision raises TwoRoom success by 3.50 to 5.50 percentage points at three source budgets. On 400 fresh PushT windows, the role interface adds 15.75 percentage points at a fixed source budget, and convex inference raises the return of a frozen Reacher agent by 54.48 points over 400 episodes.

1 INTRODUCTION

Consider two safe probes of equal cost: one reveals two nuisance bits, the other reveals the bit needed to choose an acceptable policy, and both leave the only state where that bit is observable. The first leaves fewer possible worlds but closes the decision window; the second leaves four worlds that share an acceptable policy (Fig. 1a). Only the second probe keeps a justified decision within reach.

What a model must learn in advance therefore depends on its deployment. Some distinctions never change which decision is acceptable, and others can be resolved by the safe interactions that remain; offline evidence is needed only for conflicts that those interactions cannot resolve within budget. For world models (Ha & Schmidhuber, 2018; Hafner et al., 2020; 2025), the required predictions thus depend on the agent's remaining opportunities to observe, act and commit.

We formalize this safe decision resolution problem as **Safe Dynamic Decision-Region Resolution (SD-DRR)**, whose value is the minimum worst-case cost of reaching a policy acceptable in every world consistent with the history. The common-decision criterion of decision-region determination (Javdani et al., 2014; Chen et al., 2015) specifies when to stop, and the physical dynamics of the probes determine which decisions remain reachable. A representation must therefore preserve both terminal agreement and the continuations that can establish it.

This view links representation, evidence and prediction accuracy through one quantity. Structural compression is valid when it preserves resolution cost. The minimal conflicts exceeding the online budget specify what passive data must eliminate, and convex covering allocates queries among them. For approximate models, continuation responses express the same target through error bounds and calibrated stochastic frontiers, which set the precision needed to predict the consequences of each distinction.

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

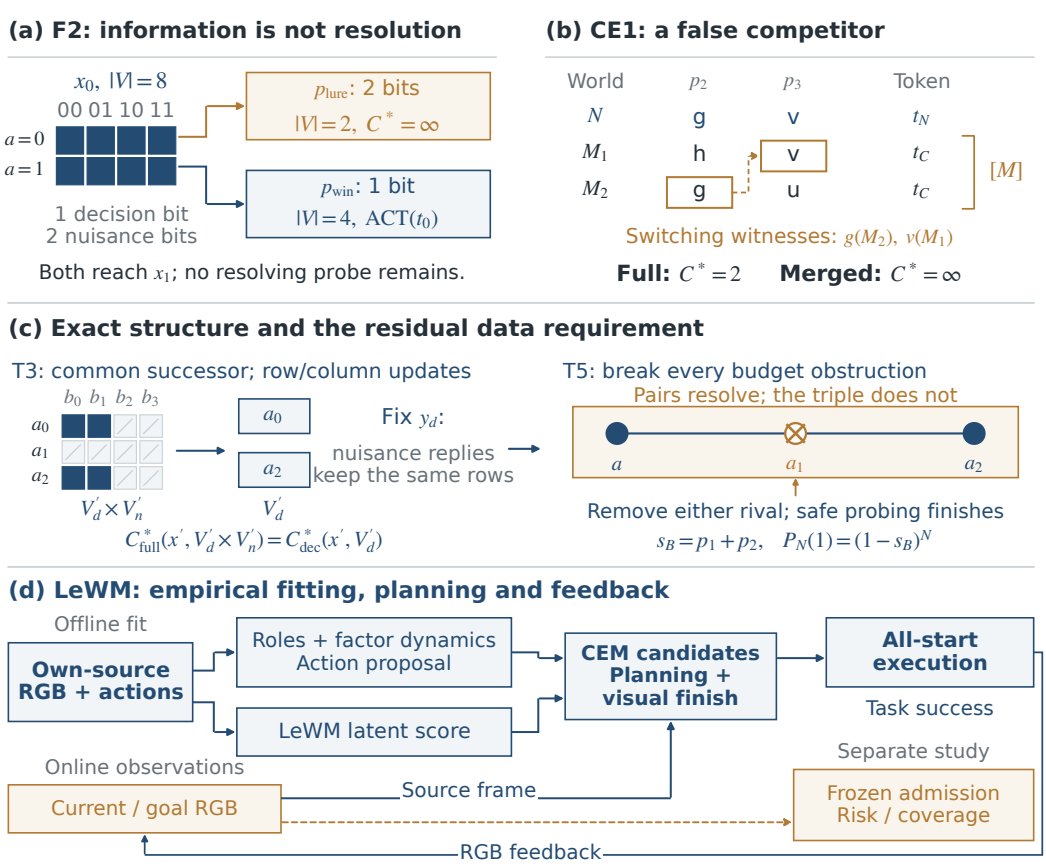

Figure 1: **Decision resolution in finite games and in a learned controller.** (a) A probe can reveal more bits while closing the decision window. (b) Merging terminal-equivalent worlds can create a spurious competitor. (c) DNF preserves safe continuation cost; offline evidence breaks the conflicts that exceed the online budget. These finite constructions motivate (d) a learned visual controller that supplies decision roles to planning and feedback (App. D.1).

Contributions.

1. **Representations that preserve resolution.** A certified structural factorization preserves optimal safe resolution cost (T3), with an exponential separation from world identification under equality-test access (T4). AutoDNF discovers and certifies this structure from explicit tables.
2. **Data requirements set by the online budget.** Minimal budget obstructions characterize exact passive-data requirements (T5). Their convex covering problem yields optimal sampling laws, dual complexity certificates and an attained finite-game exponent (T6).
3. **Prediction targets for learned control.** Continuation responses give decision-margin error budgets (T7) and data-calibrated stochastic risk and cost certificates (T8). Matched TwoRoom, PushT and Reacher interventions evaluate response supervision, decision-role structure and frozen-policy inference.

2 Decision Resolution as a Learning Target

We take as the learning target the remaining cost of reaching a justified decision. In the finite support game this is the resolution cost $C^*(x, V)$, and continuation responses provide its approximate and stochastic counterparts (T7–T8). Table 7 lists these decision objects and their assumptions.

The agent's information state. Let $\mathcal{H}$ be a finite world set. Starting from $(x_0, \mathcal{H})$, the state (x, V) pairs a physical state with the worlds compatible with the observed history. Each world w specifies acceptable policy tokens $A_w(x)$, probe safety, and a support $\sigma_w(x, q)$ of successor/observation pairs (x', y). A token represents a policy-level contract. Commitment to t is justified when $t \in \bigcap_{w \in V} A_w(x)$. Otherwise the agent applies a probe that is safe in every surviving world, then discards worlds inconsistent with the outcome.

Resolution cost and verdicts. Let $C^*(x, V)$ be the minimum worst-case cost of safe probes needed to reach a common acceptable policy, with positive costs $c(x, q)$. A resolving plan succeeds on every observation branch allowed by the surviving worlds. With budget B, the exact interface returns **ACT** when $C^* = 0$, **PROBE** when $0 < C^* \leq B$, or **ABSTAIN** when resolution exceeds the budget or is structurally impossible ($C^* = \infty$). Certificates record the common policy, contingent probe plan, or reason for refusal. The full input specification, Bellman recursion, and certificate semantics are in App. A.1.

3 From resolution cost to evidence and prediction

The exact analysis identifies which distinctions the data must remove; the response analysis specifies how accurately learned models must predict the remaining continuations. The exact game gives the chain

$$\underbrace{C^*(x, V)}_{\text{safe continuation cost}} \longrightarrow \underbrace{\mathcal{O}_B(a)}_{\text{required distinctions}} \longrightarrow \underbrace{\tau_B}_{\text{optimal evidence allocation}} .$$

T3 preserves the resolution problem, T5 identifies its statistical bottlenecks, and T6 allocates evidence to them. T7–T8 then control the error in representing and estimating continuation responses.

Why terminal agreement is insufficient. In the first counterexample, CE1 (Fig. 1b), worlds M_1 and M_2 accept the same token but predict different observations. Once they are merged, the class can explain any history by switching its supporting member from one probe to the next. The merged game therefore keeps a spurious competitor forever, whereas the full game resolves in two probes. A useful compression must preserve both how observations eliminate worlds and which safe continuations remain; DNF is a sufficient condition for both.

3.1 Preserving decision updates through structural factorization

Decision and nuisance coordinates. Arrange worlds as a grid: rows are decision coordinates and columns are nuisance coordinates. Write $\mathcal{H} = \Theta_{\mathrm{dec}} \times \Theta_{\mathrm{nui}}$. In a *decision–nuisance factorization* (DNF), acceptable policies and safety depend only on the row; each probe has a shared physical successor; and its observation support is a Cartesian product of a row-dependent decision component and a column-dependent nuisance component. Nuisance observations can differ while every allowed one preserves the same decision update (Fig. 1c).

Theorem 3 (T3, compression preserves safe resolution)**.** *Under DNF, every reachable version set has product form $V_d \times V_n$. Removing nuisance coordinates and observations gives*

$$C^*_{\mathcal{I}}(x, V_d \times V_n) = C^*_{\mathrm{dec}}(x, V_d).$$

The projected game uses $|\Theta_{\mathrm{dec}}|$ worlds and preserves optimal worst-case resolution cost.

Proof idea. For a fixed decision observation, consistency removes the same rows whichever nuisance columns survive. If a nuisance observation in the quotient conflicts with the earlier history, the product condition supplies a compatible one with the same row update, and the shared physical successor keeps the same safe continuations available. CE1 fails precisely at this repair step (App. A.4). Terminal contracts, universally safe probes and decision successors determine the entire Bellman recursion, and DNF preserves them exactly. T8 instead preserves attainable responses on general stochastic histories, with a quantitative compression bound that allows approximation error.

Related guarantees. Worlds with identical observation, safety and acceptability profiles can be merged exactly (T1). Uniform safety and acceptability alone give a sound quotient whose successful plans transfer to the full game (T2). A second counterexample, CE2, shows that equal terminal profiles cannot identify valid DNF fibers. On the family $F_5(d)$, one decision probe resolves the game,

whereas identifying the world takes at least 2^d probes in the worst case and $\Omega(2^d)$ in expectation against the worst world for any randomized strategy (T4). This exponential separation relies on equality-test access; richer test families can reduce it.

3.2 Minimal obstructions determine the data requirement

After nuisance coordinates are removed, decision uncertainty may remain as long as safe interaction can resolve it within budget B. Fix the true decision coordinate a. A *budget obstruction* $O \ni a$ is an inclusion-minimal set, among sets containing a, with $C^*_{\mathrm{dec}}(x_0, O) > B$. Let $\mathcal{O}_B(a)$ collect these sets; the data must eliminate at least one rival from every obstruction. Pairwise disagreement is insufficient: each pair of three worlds may admit a safe distinguishing probe while the triple admits none (Fig. 1c).

Assume DNF and realizable deterministic passive observations. Queries $e \sim \nu$ are i.i.d., reveal separate decision and nuisance components, and leave x_0 unchanged. Let A_N be the decision set consistent with N samples. Write $h(a, e)$ for the decision response and $s_\nu(a, O) = \Pr_e[\exists a' \in O : h(a, e) \neq h(a', e)]$ for the probability that one query breaks O.

Theorem 5 (T5, obstruction characterization and sample rate)**.** *Budget failure occurs exactly when an obstruction survives:*

$$C^*_{\mathrm{dec}}(x_0, A_N) > B \quad \Longleftrightarrow \quad \exists O \in \mathcal{O}_B(a) : O \subseteq A_N.$$

If $M_B = |\mathcal{O}_B(a)| > 0$ and $s_B = \min_{O \in \mathcal{O}_B(a)} s_\nu(a, O)$, then for $N \geq 1$,

$$(1 - s_B)^N \;\leq\; P_N(B) \;\leq\; \min\{1, M_B(1 - s_B)^N\},$$

where $P_N(B) = \Pr[C^*_{\mathrm{dec}}(x_0, A_N) > B]$. For $0 < s_B < 1$, the exact exponential decay rate is $\rho_B = -\log(1 - s_B)$. If $M_B = 0$, no passive data are needed; if $s_B = 0$, failure is certain at every sample size.

Proof idea. A resolving plan remains valid on every subset, so every infeasible set contains a minimal obstruction, and an obstruction survives only if every query misses its distinguishing events. The hardest obstruction gives the lower bound, and a union bound, which needs no independence across obstructions, gives the upper bound. For exact budget certification, $\log(1/\eta)/\rho_B$ samples are therefore necessary and $\log(M_B/\eta)/\rho_B$ are sufficient for $P_N(B) \leq \eta$, up to integer rounding.

Budget dependence. A larger budget B enlarges the resolvable family and cannot decrease ρ_B, and for a fixed decision channel the number of nuisance values does not enter the rate. In the three-world construction, the zero-probe sample requirement diverges as one passive distinction becomes rare, while a single safe probe keeps it bounded (App. A.6).

3.3 The continuation game induces a convex acquisition problem

T5 identifies the statistical bottleneck as the hardest obstruction to break. The sequential game determines which obstructions matter, and convex optimization allocates observations across them. For a finite query menu of size K, index obstructions across true worlds by $j = (a, O) \in \mathcal{J}_B$, with $m = |\mathcal{J}_B| > 0$. Let $R_{je} = 1$ when query e breaks obstruction j, and zero otherwise. For a query distribution ν, $(R\nu)_j$ is its hit probability; Δ_d denotes the probability simplex on d elements.

Theorem 6 (T6, convex design and dual obstruction witness)**.** *Under T5's assumptions, the optimal worst-obstruction hit rate is*

$$z^*_B = \max_{\nu \in \Delta_K} \min_j (R\nu)_j = \min_{\lambda \in \Delta_m} \max_e (R^\top \lambda)_e. \tag{1}$$

For $z^*_B > 0$, its reciprocal is the fractional covering cost

$$\tau_B = \min_{w \geq 0} \{\mathbf{1}^\top w : Rw \geq \mathbf{1}\} = 1/z^*_B.$$

The best worst-world sample exponent is $-\log(1 - 1/\tau_B)$, with $-\log 0 = +\infty$. Every primal–dual pair gives computable bounds $\ell(\nu) := \min_j (R\nu)_j \leq z^*_B \leq u(\lambda) := \max_e (R^\top \lambda)_e$.

Dual weights concentrate on the least-hit obstructions and positive-mass queries maximize covered dual weight, so τ_B measures the fractional effort to break every budget conflict. The gap $u - \ell$ certifies optimization accuracy and translates through T5 into sample counts; standard LP duality supplies the certificate (Boyd & Vandenberghe, 2004).

Example. Take three worlds with distinct acceptable tokens: every pair admits a safe unit-cost resolving probe, the triple admits none, and two passive queries separately distinguish the two rivals. Omitting an uninformative query, the distinct obstruction-hit rows are

$$R_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix}, \qquad R_1 = (1 \quad 1), \qquad \tau_0 = 2, \quad \tau_1 = 1.$$

With one online probe, any informative sample leaves a resolvable pair. The world class and query menu are unchanged; only the available continuation changes what the data must distinguish.

Finite-sample design. For integer $N \geq 1$,

$$\min_{\nu \in \Delta_K} U_N(\nu), \qquad U_N(\nu) = \max_a \sum_{O \in \mathcal{O}_B(a)} [1 - (R\nu)_{a,O}]^N \tag{2}$$

minimizes a convex upper bound on worst-world failure by retaining every obstruction survival term at the specified sample size. Appendix A.6 proves T6 and gives a first-order gap certificate for this finite-sample objective.

3.4 Decision margins set the required approximation accuracy

For learned models, the corresponding question is which prediction errors can change the next decision. Continuation responses express this target on general histories. For history h, representation $z = \phi(h)$, and admissible action a, let $\mathcal{P}_t(h, a)$ be the possible next-history kernels. For a bounded vector of continuation tests ψ_{t+1}, define

$$K_t(h, a) = \overline{\mathrm{conv}}\{\mathbb{E}_P[\psi_{t+1}(\phi(h'))] : P \in \mathcal{P}_t(h, a)\}.$$

Define $\widehat{K}_t(z, a)$ analogously. Convexification preserves linear continuation evaluations. Support-function duality identifies their ℓ_∞ Hausdorff error δ_t with the largest error over ℓ_1-unit linear tests (App. A.7).

Theorem 7 (T7, continuation-response transfer)**.** *Consider bounded costs over horizon H, with identical admissible actions and well-defined robust Bellman minima in both models. Uniformly bound stage-cost, response-set, and terminal-cost errors by $\epsilon_{c,t}$, δ_t, and ϵ_H. Suppose each modeled continuation value is $b + v^\top \psi_{t+1} + r$, where $\|v\|_1 \leq L_{t+1}$ and $\|r\|_\infty \leq \rho_{t+1}$. Put*

$$E_t = \epsilon_H + \sum_{j=t}^{H-1} (\epsilon_{c,j} + L_{j+1}\delta_j + 2\rho_{j+1}).$$

Then $|V_t(h) - \widehat{V}_t(\phi(h))| \leq E_t$. A model policy with stagewise optimization error η_t has true regret at most $2E_t + \sum_{j=t}^{H-1} \eta_j$. If the true best-action gap exceeds $2E_t + \eta_t$, its action is preserved exactly.

T7 converts the decision margin into a required prediction accuracy. An occupancy-weighted version measures the error only along the trajectories of the compared policies (App. A.8), and T8 complements this analysis with cost and risk certificates estimated from observed continuations.

3.5 Calibrated resolution with stochastic observations

Let Π_B contain history-dependent continuation policies that enforce probe budget B on every trajectory. Histories may be continuous and partially observed, with an unknown stochastic law. Write $r_0(\pi) = \mathbb{E}_\pi L_0$ for normalized decision cost and $r_j(\pi) = \mathbb{E}_\pi L_j$ for observed contract losses in $[0, 1]$. An indicator of any violation makes r_j its trajectory probability. A mixture q draws one complete policy before deployment; its responses are linear in q. Let $J_B^*(b)$ be the minimum true cost at risks $r_j(q) \leq b_j$ over this class.

From simultaneous response intervals, set $u_j = \widehat{r}_j + s_j$ and $l_j = \widehat{r}_j - s_j$, and solve

$$\min_q u_0(q) \quad \text{subject to} \quad u_j(q) \le b_j. \tag{3}$$

For a supplied controller library this is a linear program.

Theorem 8 (T8, calibrated stochastic resolution)**.** *Suppose all response intervals cover with probability $1-\alpha$. On this event, a feasible solution $\widehat{q}$ satisfies every true risk constraint and, for every $\lambda \ge 0$,*

$$r_0(\widehat{q}) - J_B^*(b) \le u_0(\widehat{q}) - d(\lambda), \qquad d(\lambda) = \inf_{\pi \in \Pi_B} [l_0(\pi) + \lambda^\top l(\pi)] - \lambda^\top b,$$

where $l = (l_1, \dots, l_m)$. If $s_j \le e_j$ for all channels, the class contains a mixture with risk slack $\gamma_j > 0$, and $2e_j \le \gamma_j$ for $j \ge 1$, an η-optimal upper-program solution additionally obeys

$$r_0(\widehat{q}) - J_B^*(b) \le 2e_0 + \max_j (2e_j/\gamma_j) + \eta.$$

Learning the intervals. Direct rollout losses provide unbiased response samples. With behavioral action coverage, logged trajectories also supply unbiased corrections of arbitrary fixed neural continuation scores (Jiang & Li, 2016). For K controllers fixed before n i.i.d. calibration episodes, score range D gives joint radius $D\sqrt{\log(2K(m+1)/\alpha)/(2n)}$; empirical variances sharpen it (Maurer & Pontil, 2009). Covering numbers extend this construction to infinite controller classes. Complete proofs, explicit sample counts and convex calibration allocation appear in Apps. A.9–A.11. Because the guarantee uses observed cost and risk, it allows a misspecified neural model and dependent latent variables and requires no DNF support grid.

In this stochastic setting, compression must preserve the convex set of attainable cost–risk responses. Matching this set preserves the frontier, and approximating it adds an explicit representation term (Cor. A.6). A continuous Gaussian probe example has an analytic frontier and a calibrated cost-gap certificate (App. A.12).

Convex implementations. The visual interface fits affine role readouts to point or convex-set labels by the strongly convex objective in Eq. (A.21), then fuses model and image estimates by metric projection onto admissible factors. The projection decreases squared factor error, while an L-Lipschitz continuation readout turns factor error ζ into response error $L\zeta$ (Props. 1 and 2).

4 Implementing the decision target

Certified structure discovery. AutoDNF turns the structural condition of T3 into an executable procedure: it proposes a decision–nuisance grid from a flat singleton-support table and then certifies that the grid preserves resolution cost. All 300 controlled instances yield certified compressions, 285 through the direct proposal and 15 through bounded search. On the 33.56-million-entry $F_5(12)$ table, the array implementation discovers and certifies the structure in 64.25 s, versus 642.83 s for the reference. The counting test, algorithms and opaque-table confirmation are in App. C.

Learned decision interfaces. The learned-controller studies intervene at three stages: the responses predicted during training (TwoRoom), the decision factors supplied to planning (PushT), and the factor estimates supplied to a trained actor (Reacher). PushT role coordinates are robot position, object position and object orientation, and each source budget fits its own mapper, factor predictor and proposal. Both PushT controllers use three LeWorldModel members (Maes et al., 2026), 300 CEM candidates, 30 rounds and 50 physical steps. Latent-only control removes the role modules, proposal and terminal visual servo; the complete controller also fixes one right-angle frame from training-source support (Eq. D.1). Figure 2 reports the paired gains. Admission is evaluated separately with a frozen rule (Apps. D.1 and D.5).

5 Experiments

We test how available probes, query allocation, response supervision, role structure and deployed inference each affect decision quality at their stated data budgets.

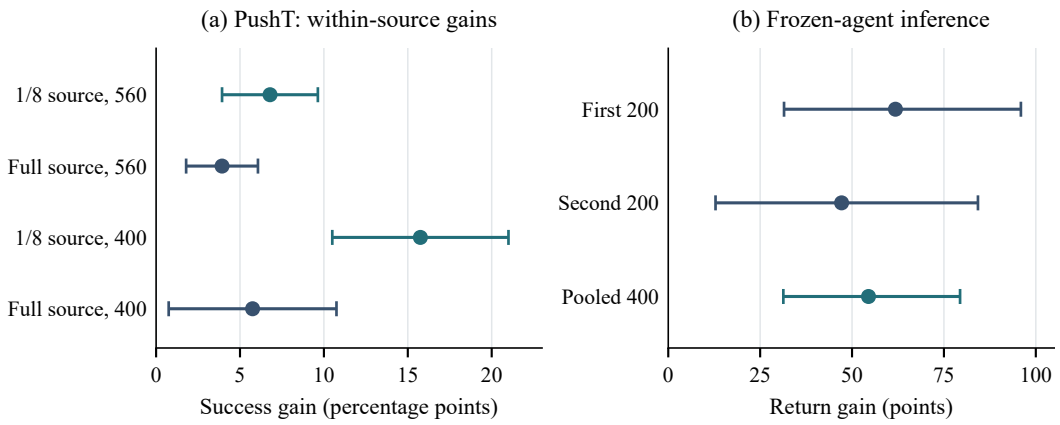

Figure 2: **Paired gains of the decision interfaces.** (a) PushT success gains over latent-only control within each source budget, on standard starts and on fresh trajectory windows. (b) Reacher return gains from the calibrated inference layer on one fixed neural checkpoint. Error bars are paired 95% intervals; the pooled estimate combines both independent cohorts.

5.1 Safe continuations reduce the required passive evidence

The finite games measure how much decision uncertainty offline data may leave for a fixed budget of safe interaction. Retention requires the paired-bootstrap fifth percentile of reduced-to-full success to reach 0.90. The 64-world game has eight decision and eight nuisance values; all methods share exact consistency learning (Mitchell, 1982) and four online interactions. Fresh 600-trial blocks per family and data size compare $N \in \{12, 14, 18, 144\}$. At $N = 14$, a $10.3\times$ reduction, the two families reach 95.33% and 94.50% success versus 100% at full data, with retention lower bounds 0.938 and 0.930. None of the 4,800 trials has a false ACT or an unsafe probe (Fig. 4).

At $N = 12$ on the original grid, SD-DRR reaches 94.2%, versus 44.3% for point estimation and 15.8% for committing once 80% of surviving worlds share a token. Greedy elimination with common-decision stopping matches it at $N \geq 18$; the families F2 (information window) and F3 (state-dependent availability) isolate the added value of resolving continuations (Figs. 8 and 5).

Converting two nuisance bits into decision bits cuts the reduction from $12\times$ to $8\times$ at budget four, and one more probe raises it to $16\times$ (lower bound 0.990). The world count is unchanged, so the data requirement follows the remaining access to decisive probes (App. D).

Exact-support transfer on Boxoban-Slip.

Boxoban-Slip adds hidden ice, on which a wrong slide can deadlock the level, to 200 external levels (Guez et al., 2018). SD-DRR checks every stop consistent with its ice knowledge; competing planners use eight posterior ice-map samples. With 32 rather than 1,024 own-level interactions, SD-DRR solves 81% of levels versus 79% for the strongest full-data competitor, with no irreversible failure in 1,200 episodes (Fig. 7).

5.2 Convex allocation attains the predicted exponent

This experiment uses T5 and T6 to choose a sampling distribution. Four worlds form two pairs (App. D.7). At budget zero, a test that only identifies the pair leaves a within-pair obstruction intact. The optimal allocation spreads its mass equally over the four identity tests, giving hit rate $1/2$ and covering number $\tau_0 = 2$. Every pure test leaves some world unresolved at every sample size, and a uniform mixture of the six non-constant tests has hit rate $1/3$. Six samples from the optimal mixture fail with probability 0.0461, within the T5 upper bound 0.0469 and below the 5% target; the uniform mixture first meets the target at eight samples, the earliest size its lower bound 0.0390 allows (Fig. 3a). At budget one, the pair test, unused at budget zero, suffices alone and one sample drives failure to zero; at budget two no passive sample is needed. Across 280 random allocations, predicted and exact decay rates differ by a median of 2.8×10^{-4} (Fig. 3b).

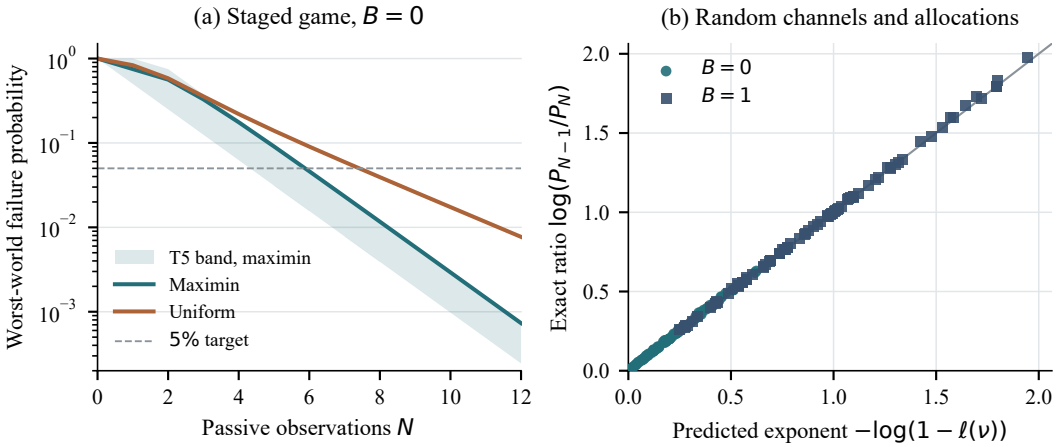

Figure 3: **Convex query allocation in the four-world game.** (a) Worst-world failure at budget zero. The band is the T5 bound; the maximin curve tracks its upper edge, and the dashed line is 5%. (b) Exact decay ratio against $-\log(1-\ell(\nu))$ on 280 pairs. The median absolute gap is 2.8×10^{-4}.

Source episodes	Rollout training	+ Response loss	Mean gain
75	52.17	**55.94**	+3.78
150	49.00	**52.50**	+3.50
1,200	53.33	**58.83**	+5.50

Table 1: **TwoRoom success (%) with and without the response loss.** The 75-episode source crosses three source draws with three training seeds; the other budgets use three training seeds. Every trained policy executes the same 200 tasks. Gains are averaged over unrounded paired results. The 1,200-episode paired gain is 5.50 ± 2.78 (sample SD, three seeds).

5.3 Continuation responses improve the learning objective

TwoRoom tests whether response supervision improves rollout training. The added objective predicts bounded goal discrepancies along source continuations, following T7's prescription to prioritize continuation values. Within each source and seed, both arms share the initialization, source order, 4,000 updates, rollout loss and planner (App. D.2).

The response loss improves mean success at all three budgets (Table 1). Including initialization, the budgets are 148,678, 299,619 and 2,397,936 frames, and every policy plans with 300-candidate, 30-round CEM within 50 physical steps (App. D.2).

5.4 Decision roles improve visual planning at fixed source budgets

The two PushT sources contain 291,969 and 2,336,736 frames. Evaluation uses 560 standard starts outside 80 fixed development starts and 400 fresh windows held out from fitting and feedback (auxiliary counts in Table 8).

Held-out windows. With the same frames and planning caps, the reduced-source interface solves 184 of the 400 held-out windows versus 121 for latent-only planning; with the full source it still leads, 214 versus 191 (Table 2).

Component contributions. Two matched interventions in the original image frame separate the contributions of latent prediction and explicit decision coordinates (App. D.6). Removing the latent score while keeping every forward pass lowers success from $519/560$ to $397/560$ (reduced source) and from $533/560$ to $405/560$ (full source). Replacing grouped roles by a role-free refit with matched source, fitting and selection budgets lowers success from $526/560$ to $489/560$.

Admission risk. A reduced-source admission rule, fixed before sampling 640 of 11,018 catalogue items, accepts 502 starts, of which 488 succeed. Exact count inversion gives a 95.13% success lower

Source	Cohort	Latent-only	Interface	Gain [95% interval]
291,969	Fresh 400	30.25	**46.00**	$+15.75$ $[10.50, 21.00]$
2,336,736	Fresh 400	47.75	**53.50**	$+5.75$ $[0.75, 10.75]$
291,969	Standard 560	85.71	**92.50**	$+6.79$ $[3.93, 9.64]$
2,336,736	Standard 560	91.79	**95.71**	$+3.93$ $[1.79, 6.07]$

Table 2: **PushT success (%) with identical source frames and planning budgets.** The first two rows are windows excluded from source fitting and from controller feedback; the last two rows are source-trajectory starts outside the 80 development tasks. Gains are paired percentage points.

bound on the remaining accepted items; the bounds of all three rules hold jointly at 90% confidence (App. D.5).

5.5 Convex inference improves a frozen Dreamer agent

Reacher tests whether better factor inference improves an already trained policy. With the world model and actor frozen, convex inference fuses a source-calibrated readout, the learned endpoint dynamics and a static target into the actor's role estimate. On the same 400 task seeds, the return rises from 909.48 to 963.96, a paired gain of 54.48 (95% interval $[31.31, 79.39]$), and both independent 200-episode cohorts improve, by 61.81 and 47.16. Calibration reuses 2,580 images already counted in the 1,010,000 training steps, so the gain is attributable to deployed inference (App. D.4); App. A.14 evaluates the point/set readout on held-out source episodes.

6 Related work

World models are usually trained to predict action effects (Hafner et al., 2020; 2025; Hansen et al., 2024; Ye et al., 2021; Chua et al., 2018; Maes et al., 2026). Bisimulation and value equivalence ask which of these predictions a decision actually needs (Givan et al., 2003; Ferns et al., 2004; Gelada et al., 2019; Zhang et al., 2021; Grimm et al., 2020; Poli et al., 2023; Thorpe et al., 2026). SD-DRR adds the remaining interaction opportunities to this question: it preserves the cost of resolving uncertainty through safe probes until the surviving worlds share an acceptable policy.

Decision-region determination encodes common-decision conflicts (Javdani et al., 2014; Chen et al., 2015); belief-support games incorporate dynamic uncertainty (Chatterjee et al., 2006; Bertrand et al., 2017). Bayes-adaptive planning, information-directed sampling and experimental design value information through decisions (Duff, 2002; Guez et al., 2012; Russo & Van Roy, 2014; Huang et al., 2024; Go et al., 2026; Wan et al., 2026). Disagreement-based active learning and adaptive submodular querying score conflicts inside a fixed hypothesis class (Dasgupta, 2005; Hanneke, 2014; Golovin & Krause, 2011). T5 and T6 instead target only the conflicts that exceed the online budget and allocate passive queries to break them through a convex covering problem (Boyd & Vandenberghe, 2004), so the remaining deployment opportunities decide which evidence is worth collecting.

Shielding, reachability analysis and constrained policy optimization address the safety of actions (Alshiekh et al., 2018; Bansal et al., 2017; Achiam et al., 2017). The admission study relates to calibrated help requests, agent risk assessment and conformal risk bounds (Ren et al., 2023; Ruan et al., 2024; Vovk et al., 2005; Angelopoulos & Bates, 2023).

7 Conclusion

SD-DRR measures data efficiency against the decisions that remain reachable at deployment. Resolution cost determines which representations are valid, budget obstructions specify the evidence offline data must supply, and continuation responses convert prediction accuracy into decision guarantees. Interventions derived from these results reduced passive data in finite games and improved control at fixed source budgets in TwoRoom, PushT and Reacher.

Limitations. The visual results cover image-based tasks with fixed planning horizons and task-specific source and evaluation protocols.

AI USE STATEMENT

Large language models assisted with editing this manuscript and polishing its language.

ETHICS STATEMENT

This work aims to reduce the risk of irreversible failures by deciding, before commitment, whether to act, probe or abstain; in the exact finite setting these verdicts are certified. The benchmarks involve no human subjects or personal data. For agentic systems that execute irreversible actions, abstention and escalation to a human are treated as explicit outputs.

REPRODUCIBILITY STATEMENT

Appendix A states the assumptions and proofs. Appendices B and C specify the finite constructions, reference solvers, and AutoDNF algorithms. Appendix D gives source definitions, learned-interface objectives, controller settings, fixed task cohorts, statistical calculations, and the exact staged-game exponent. Appendix E describes the supplementary archive, which contains the code, the per-trial records and a script that recomputes the reported numbers.

REFERENCES

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *International Conference on Machine Learning*, pp. 22–31, 2017.

Mohammed Alshiekh, Roderick Bloem, Rüdiger Ehlers, Bettina Könighofer, Scott Niekum, and Ufuk Topcu. Safe reinforcement learning via shielding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

Anastasios N Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

Somil Bansal, Mo Chen, Sylvia Herbert, and Claire J Tomlin. Hamilton-Jacobi reachability: A brief overview and recent advances. In *IEEE 56th Annual Conference on Decision and Control*, pp. 2242–2253, 2017.

Nathalie Bertrand, Blaise Genest, and Hugo Gimbert. Qualitative determinacy and decidability of stochastic games with signals. *Journal of the ACM*, 64(5):1–48, 2017.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. ISBN 9780521833783. URL https://web.stanford.edu/~boyd/cvxbook/.

Krishnendu Chatterjee, Laurent Doyen, Thomas A Henzinger, and Jean-François Raskin. Algorithms for omega-regular games with imperfect information. In *International Workshop on Computer Science Logic*, pp. 287–302, 2006.

Yuxin Chen, Shervin Javdani, Amin Karbasi, J Andrew Bagnell, Siddhartha Srinivasa, and Andreas Krause. Submodular surrogates for value of information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2015.

Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Sanjoy Dasgupta. Coarse sample complexity bounds for active learning. In *Advances in Neural Information Processing Systems*, volume 18, pp. 235–242, 2005.

Michael O'Gordon Duff. *Optimal Learning: Computational Procedures for Bayes-Adaptive Markov Decision Processes*. PhD thesis, University of Massachusetts Amherst, 2002.

Norm Ferns, Prakash Panangaden, and Doina Precup. Metrics for finite Markov decision processes. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pp. 162–169, 2004.

Carles Gelada, Saurabh Kumar, Jacob Buckman, Ofir Nachum, and Marc G Bellemare. DeepMDP: Learning continuous latent space models for representation learning. In *International Conference on Machine Learning*, pp. 2170–2179, 2019.

Robert Givan, Thomas Dean, and Matthew Greig. Equivalence notions and model minimization in Markov decision processes. *Artificial Intelligence*, 147(1-2):163–223, 2003.

Jinwoo Go, Xiaoning Qian, and Byung-Jun Yoon. Goal-driven Bayesian optimal experimental design for robust decision-making under model uncertainty. *arXiv preprint arXiv:2605.26093*, 2026.

Daniel Golovin and Andreas Krause. Adaptive submodularity: Theory and applications in active learning and stochastic optimization. *Journal of Artificial Intelligence Research*, 42:427–486, 2011. URL https://www.jair.org/index.php/jair/article/view/10731.

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

Arthur Guez, David Silver, and Peter Dayan. Efficient Bayes-adaptive reinforcement learning using sample-based search. In *Advances in Neural Information Processing Systems*, volume 25, 2012.

Arthur Guez, Mehdi Mirza, Karol Gregor, Rishabh Kabra, Sébastien Racanière, Théophane Weber, David Raposo, Adam Santoro, Laurent Orseau, Tom Eccles, Greg Wayne, David Silver, Timothy Lillicrap, and Victor Valdes. An investigation of model-free planning: Boxoban levels. https://github.com/google-deepmind/boxoban-levels, 2018.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

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

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

Steve Hanneke. Theory of disagreement-based active learning. *Foundations and Trends in Machine Learning*, 7(2–3):131–309, 2014. doi: 10.1561/2200000037.

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

Daolang Huang, Yujia Guo, Luigi Acerbi, and Samuel Kaski. Amortized Bayesian experimental design for decision-making. In *Advances in Neural Information Processing Systems*, 2024.

Shervin Javdani, Yuxin Chen, Amin Karbasi, Andreas Krause, Drew Bagnell, and Siddhartha Srinivasa. Near optimal Bayesian active learning for decision making. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33 of *Proceedings of Machine Learning Research*, pp. 430–438. PMLR, 2014. URL https://proceedings.mlr.press/v33/javdani14.html.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 652–661. PMLR, 2016. URL https://proceedings.mlr.press/v48/jiang16.html.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*, pp. 267–274. Morgan Kaufmann, 2002. URL https://homes.cs.washington.edu/~sham/papers/rl/aoarl.pdf.

Lucas Maes, Quentin Le Lidec, Damien Scieur, Yann LeCun, and Randall Balestriero. LeWorldModel: Stable end-to-end joint-embedding predictive architecture from pixels. *arXiv preprint arXiv:2603.19312*, 2026. doi: 10.48550/arXiv.2603.19312. URL https://arxiv.org/abs/2603.19312.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009. URL https://arxiv.org/abs/0907.3740.

Tom M Mitchell. Generalization as search. *Artificial Intelligence*, 18(2):203–226, 1982.

Naoki Morihira, Amal Nahar, Kartik Bharadwaj, Yasuhiro Kato, Akinobu Hayashi, and Tatsuya Harada. R2-Dreamer: Redundancy-reduced world models without decoders or augmentation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Je2QqXrcQq.

Michael Poli, Stefano Massaroli, Stefano Ermon, Bryan Wilder, and Eric Horvitz. Ideal abstractions for decision-focused learning. In *International Conference on Artificial Intelligence and Statistics*, volume 206, pp. 10223–10234, 2023. URL https://proceedings.mlr.press/v206/poli23a.html.

Allen Z. Ren, Anushri Dixit, Alexandra Bodrova, Sumeet Singh, Stephen Tu, Noah Brown, Peng Xu, Leila Takayama, Fei Xia, Jake Varley, Zhenjia Xu, Dorsa Sadigh, Andy Zeng, and Anirudha Majumdar. Robots that ask for help: Uncertainty alignment for large language model planners. In *Proceedings of the 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 661–682, 2023. URL https://proceedings.mlr.press/v229/ren23a.html.

Yangjun Ruan, Honghua Dong, Andrew Wang, Silviu Pitis, Yongchao Zhou, Jimmy Ba, Yann Dubois, Chris J Maddison, and Tatsunori Hashimoto. Identifying the risks of LM agents with an LM-emulated sandbox. In *International Conference on Learning Representations*, 2024.

Daniel Russo and Benjamin Van Roy. Learning to optimize via information-directed sampling. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

Adam J. Thorpe, Stepan Tretiakov, Cheng-Hsi Hsiao, Su Ann Low, Xingjian Li, Hassan Iqbal, Neel P. Bhatt, Ufuk Topcu, and Krishna Kumar. Physically viable world models: A case for query-conditioned embodied AI. *arXiv preprint arXiv:2605.30542*, 2026.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Beichen Wan, Mo Liu, Paul Grigas, and Zuo-Jun Max Shen. Decision-focused sequential experimental design: A directional uncertainty-guided approach. *arXiv preprint arXiv:2602.05340*, 2026.

Andrew Chi-Chih Yao. Probabilistic computations: Toward a unified measure of complexity. In *18th Annual Symposium on Foundations of Computer Science*, pp. 222–227, 1977.

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

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.

Appendix roadmap. Appendix A proves T1–T8, the obstruction sample bounds, convex design duality, convex response transfer, set-supervised readouts, factor inference, and the counterexamples. Appendix B gives the finite-model and Boxoban figures and verification records. Appendix C specifies AutoDNF, its soundness, search procedure and opaque-table confirmation. Appendix D gives the visual implementation, evaluation protocols, matched controls, the frozen-policy risk bound, and the exact staged-game exponent. Appendix E indexes the manuscript sources and summary-figure inputs. The experimental protocols identify each fixed cohort, intervention, source budget, and paired comparison.

A Proofs and the passive-data connection

This appendix proves T1–T8, gives the CE1/CE2 constructions and the randomized identification bound, and characterizes passive data through minimal budget obstructions. It includes sample upper and lower bounds, a budget frontier, and convex query-design certificates. The original T1–T4 verification suite is mapped to its finite-instance checks in Table 5.

A.1 Setup (self-contained)

A finite SD-DRR instance is $\mathcal{I} = (\mathcal{H}, X, Q, Y, \mathrm{sig}, \mathrm{Safe}, A, c)$, making the interface in Sec. 2 explicit: $\mathcal{H}$ a finite set of worlds, X physical states, Q probing actions, Y observations; support maps $\mathrm{sig}_w(x,q) \subseteq X \times Y$; safety predicate $\mathrm{Safe} \subseteq \mathcal{H} \times X \times Q$; acceptability sets $A_w(x) \subseteq \mathcal{T}$; strictly positive costs $c : X \times Q \to \mathbb{R}_{>0}$. Well-formedness: $(w,x,q) \in \mathrm{Safe} \Rightarrow \mathrm{sig}_w(x,q) \neq \emptyset$.

Game (sure semantics). Information states $s = (x, V)$ with $\emptyset \neq V \subseteq \mathcal{H}$; authorization set $G = \{(x,V) : \bigcap_{w\in V} A_w(x) \neq \emptyset\}$; admissible probes $Q_{\mathrm{pre}}(x,V) = \{q : \forall w \in V,\ (w,x,q) \in \mathrm{Safe}\}$. After q, for each $(x',y) \in \bigcup_{w\in V} \mathrm{sig}_w(x,q)$ the successor support is $V' = \{w \in V : (x',y) \in \mathrm{sig}_w(x,q)\}$ (nonempty by construction), and the adversary picks the branch. The value C^* is the least fixed point of

$$C^*(s) = \begin{cases} 0, & s \in G, \\ \min_{q\in Q_{\mathrm{pre}}(s)} \left[c(x,q) + \max_{(x',V')\in \mathrm{succ}(s,q)} C^*(x',V')\right], & \text{otherwise,} \end{cases}$$

with the convention $\min_\emptyset = +\infty$. Since costs are strictly positive and the information-state space is finite, the least fixed point exists, is unique, and is reached by Bellman relaxation initialized at $+\infty$ (this is exactly what the reference solver implements).

Closure quotients. For a partition Π of $\mathcal{H}$ in which every class has a constant acceptability profile and a constant safety profile (an *acc/safe-uniform* partition), the quotient instance $\mathcal{I}/\Pi$ has worlds Π, $A_K := A_w$ and $\mathrm{Safe}_K := \mathrm{Safe}_w$ for any $w \in K$, and *closure signatures* $\mathrm{sig}_K(x,q) := \bigcup_{w\in K} \mathrm{sig}_w(x,q)$ (on safe combinations). The projection is $\alpha_\Pi(V) = \{K \in \Pi : K \cap V \neq \emptyset\}$. Two special partitions: the *fine congruence* $\approx_f$ (equal acceptability, safety, *and* support profiles) and the *profile equivalence* $\approx_c$ (equal acceptability and safety profiles only).

A.2 T1: the fine quotient is an isomorphism

Theorem T1 (fine quotient, restated). *Let Π_f be the partition induced by $\approx_f$ and $k_f = |\Pi_f|$. Then: (i) every V reachable from $(x_0, \mathcal{H})$ is $\approx_f$-saturated (a union of classes); (ii) the map $\varphi(x,V) = (x, \kappa(V))$ with $\kappa(V) = \{K \in \Pi_f : K \subseteq V\}$ is an isomorphism between the reachable subgraph of the full game and the reachable subgraph of the quotient game (preserving G-membership, admissible probes, branches, and costs); (iii) $C^*(x,V) = C^*_{/f}(\varphi(x,V))$ on all reachable states, and verdicts and optimal policies correspond under φ; (iv) the quotient game has at most $|X| \cdot 2^{k_f}$ information states, so solving is fixed-parameter tractable in k_f, and Π_f itself is computable in input-polynomial time by one-shot signature bucketing.*

Lemma A.1 (saturation)**.** *Every V reachable from $V_0 = \mathcal{H}$ is a union of $\approx_f$-classes.*

Proof. Induction on the play. V_0 is trivially saturated. Let V be saturated, $q \in Q_{\mathrm{pre}}(x,V)$, and let (x',y) be a branch with survivor set $V' = \{w \in V : (x',y) \in \mathrm{sig}_w(x,q)\}$. Take $w \approx_f w'$ in the same class; by saturation either both or neither belong to V, so assume both. Their safety

profiles agree, so q is safe for both; their support profiles agree at the safe pair (x, q), so $(x', y) \in \mathrm{sig}_w(x, q) \iff (x', y) \in \mathrm{sig}_{w'}(x, q)$. Hence $w \in V' \iff w' \in V'$: V' is saturated. □

Lemma A.2 (isomorphism)**.** *On saturated states φ is a bijection onto quotient states and: (a) $(x, V) \in G \iff \varphi(x, V) \in G_{/f}$; (b) $Q_{\mathrm{pre}}(x, V) = Q_{\mathrm{pre}}(\varphi(x, V))$; (c) the branch sets after any admissible q are in bijection, successors satisfy $\varphi(x', V') = (x', \kappa(V)')$, and costs agree.*

Proof. Bijectivity on saturated sets is immediate ($V = \bigcup_{K \in \kappa(V)} K$). (a) A_w is constant on classes, so $\bigcap_{w \in V} A_w(x) = \bigcap_{K \in \kappa(V)} A_{\mathrm{rep}(K)}(x)$. (b) Likewise with Safe. (c) Since support profiles are constant on classes, $\bigcup_{w \in V} \mathrm{sig}_w(x, q) = \bigcup_{K \in \kappa(V)} \mathrm{sig}_{\mathrm{rep}(K)}(x, q)$, so the branch sets coincide; and by the argument of Lemma A.1 class members enter and leave survivor sets together, so the quotient survivor set after branch (x', y) is exactly $\{K \in \kappa(V) : (x', y) \in \mathrm{sig}_{\mathrm{rep}(K)}(x, q)\} = \kappa(V')$. □

Proof of Theorem T1. (i) is Lemma A.1. (ii) is Lemma A.2, with reachability transported in both directions by induction along (c). (iii) By (ii) the two game graphs restricted to reachable states are isomorphic, so $C^* \circ \varphi^{-1}$ and $C^*_{/f}$ satisfy the same Bellman least-fixed-point equation on the same graph; uniqueness of the least fixed point gives equality, and verdicts/optimal policies are graph-defined objects, hence correspond under φ. (iv) Quotient information states number at most $|X| \cdot 2^{k_f}$; value iteration is polynomial per round with the number of rounds bounded by the number of reachable states, giving total time $2^{O(k_f)} \cdot \mathrm{poly}$. For the bucketing: map each w to the canonical string of its full profile triple $(A_w(\cdot), \mathrm{Safe}_w(\cdot, \cdot), \mathrm{sig}_w(\cdot, \cdot))$ and bucket by equality. Because the *entire* support profile is part of the signature, one pass suffices; no iterative refinement (as in bisimulation computations) is needed. □

A.3 T2: CLOSURE QUOTIENTS ARE ONE-SIDED SOUND; CE1

Theorem T2 (one-sided soundness, restated). *For every acc/safe-uniform partition Π and every state (x, V): $C^*_{\mathcal{I}}(x, V) \le C^*_{\mathcal{I}/\Pi}(x, \alpha_\Pi(V))$. In particular a winning quotient certificate transfers to an executable winning strategy of the full game at no extra cost.*

Proof (strategy transfer with a simulation invariant). Let σ be a quotient strategy from $(x_0, \hat{A}_0)$, $\hat{A}_0 = \alpha_\Pi(V_0)$, guaranteeing arrival in $G_{/\Pi}$ at cost $\le B := C^*_{/\Pi}(x_0, \hat{A}_0)$ (if $B = \infty$ there is nothing to prove). Define a full-game strategy τ that maintains a shadow quotient state $(x, \hat{A})$ subject to the invariant

$$\alpha_\Pi(V) \subseteq \hat{A}. \tag{Inv}$$

Initially both sides are equal. Inductive step:

1. If $(x, \hat{A}) \in G_{/\Pi}$, pick a common token $t \in \bigcap_{K \in \hat{A}} A_K(x)$. For every $w \in V$ we have $[w]_\Pi \in \alpha_\Pi(V) \subseteq \hat{A}$ and $A_w = A_{[w]_\Pi} \ni t$, hence $(x, V) \in G$: the full game commits immediately at cost 0.

2. Otherwise σ prescribes a probe q safe for every class in $\hat{A}$. By (Inv) the class of every $w \in V$ lies in $\hat{A}$, and safety is constant on classes, so $q \in Q_{\mathrm{pre}}(x, V)$: the probe is legal in the full game and has the same cost $c(x, q)$.

3. The full-game adversary picks a branch (x', y) with $V' \neq \emptyset$. Choose any $w \in V'$: then $(x', y) \in \mathrm{sig}_w(x, q) \subseteq \mathrm{sig}_{[w]_\Pi}(x, q)$ and $[w]_\Pi \in \hat{A}$, so (x', y) is also a legal branch of the shadow quotient state. Update $\hat{A}' = \{K \in \hat{A} : (x', y) \in \mathrm{sig}_K(x, q)\}$. The invariant is preserved: if $K \in \alpha_\Pi(V')$ then some $w \in V' \cap K$ exists, so $K \in \alpha_\Pi(V) \subseteq \hat{A}$ and $(x', y) \in \mathrm{sig}_w(x, q) \subseteq \mathrm{sig}_K(x, q)$, hence $K \in \hat{A}'$.

Every play of τ thus projects to a legal play of σ with identical actions and costs. Since σ reaches $G_{/\Pi}$ within cost B on every play, step 1 fires at the latest at that point, and the full game is then in G. Hence $C^*_{\mathcal{I}}(x_0, V_0) \le B$. The argument applies verbatim from any (x, V) with shadow $\alpha_\Pi(V)$. □

Proposition A.1 (CE1: profile merging changes resolution cost)**.** *There is an instance with $C^*_{\mathcal{I}} = 2$ while $C^*_{\mathcal{I}/\approx_c} = \infty$.*

Proof. **Construction.** $\mathcal{H} = \{N, M_1, M_2\}$, a single physical state x_0, two probes p_2, p_3, all safe, deterministic observations, and tokens $A_N = \{t_N\}$, $A_{M_1} = A_{M_2} = \{t_C\}$ with $t_N \neq t_C$:

world	obs. under p_2	obs. under p_3	tokens
N	g	v	$\{t_N\}$
M_1	h	v	$\{t_C\}$
M_2	g	u	$\{t_C\}$

Full game. $(x_0, \mathcal{H}) \notin G$. Play p_2: branch h gives $\{M_1\} \in G$; branch g gives $\{N, M_2\} \notin G$, from which p_3 splits into $v \to \{N\} \in G$ and $u \to \{M_2\} \in G$. Worst case two probes. No single probe wins: p_2 has the branch $g \to \{N, M_2\}$ and p_3 has the branch $v \to \{N, M_1\}$, both mixed-token pairs. Hence $C^*_{\mathcal{I}} = 2$.

Profile quotient. $\approx_c$ merges M_1, M_2 (equal acceptability and safety profiles), giving $\Pi = \{K_N = \{N\}, K_C = \{M_1, M_2\}\}$ with closure signatures $\mathrm{sig}_{K_C}(p_2) = \{g, h\}$ and $\mathrm{sig}_{K_C}(p_3) = \{u, v\}$. From $\hat{A} = \{K_N, K_C\}$: under p_2 the branch g keeps both classes alive ($g \in \{g\}$ and $g \in \{g, h\}$), and under p_3 the branch v keeps both alive ($v \in \{v\}$, $v \in \{u, v\}$). The adversary always picks that branch, so $\{(x_0, \{K_N, K_C\})\}$ is a losing set closed under all admissible probes: $C^*_{/\approx_c} = \infty$. □

The mechanism: taking unions of support profiles *inside* a class creates a covering branch that no single member world could produce, and the adversary hides in it forever. Profile equality alone does not preserve value; what is missing is exactly the rectangularity that T3 declares.

A.4 T3: DECLARED DNF FIBERS ARE SOUND AND COMPLETE

DNF condition. $\mathcal{H} = \Theta_{\mathrm{dec}} \times \Theta_{\mathrm{nui}}$, $w = (a, b)$, such that (i) $A_{(a,b)} = A_a$ and $\mathrm{Safe}_{(a,b)} = \mathrm{Safe}_a$; (ii) $Y = Y_1 \times Y_2$ and there are a shared successor $x'(x, q)$ and support tables $D_a(x, q) \subseteq Y_1$, $N_b(x, q) \subseteq Y_2$ with

$$\mathrm{sig}_{(a,b)}(x, q) = \{x'(x, q)\} \times \big(D_a(x, q) \times N_b(x, q)\big);$$

(iii) $D_a, N_b \neq \emptyset$ at safe combinations. The *declared fiber partition* is $\Pi_{\mathrm{dec}} = \{C_a = \{a\} \times \Theta_{\mathrm{nui}}\}$, which is acc/safe-uniform by (i).

Theorem T3 (declared fibers, restated). *Under DNF, for every state reachable from $(x_0, \mathcal{H})$ (all of which are products $A \times B$):*

$$C^*_{\mathcal{I}}(x, A \times B) \;=\; C^*_{\mathcal{I}/\Pi_{\mathrm{dec}}}(x, A),$$

with verdicts and optimal policies in step-wise correspondence. The quotient game is independent of $|\Theta_{\mathrm{nui}}|$, and when the DNF structure is given as a declaration (tables $A_a, \mathrm{Safe}_a, D_a, N_b, x'$), the quotient instance is built in time polynomial in the declaration.

Lemma A.3 (product invariance)**.** *Every V reachable from $\Theta_{\mathrm{dec}} \times \Theta_{\mathrm{nui}}$ has the form $A \times B$ with $A, B \neq \emptyset$.*

Proof. Induction. After probe q and branch $(x', (y_1, y_2))$:

$$V' = \{(a, b) \in A \times B : \; y_1 \in D_a(x, q)\} \cap \{(a, b) : \; y_2 \in N_b(x, q)\} = A' \times B'$$

with $A' = \{a \in A : y_1 \in D_a(x, q)\}$ and $B' = \{b \in B : y_2 \in N_b(x, q)\}$; a branch is nonempty exactly when both factors are. □

Lemma A.4 (quotient survival reads only y_1)**.** $\mathrm{sig}_{C_a}(x, q) = \{x'(x, q)\} \times D_a(x, q) \times \bar{N}(x, q)$ *with $\bar{N} := \bigcup_{b \in \Theta_{\mathrm{nui}}} N_b$. Hence after a quotient branch $(x', (y_1, y_2))$ the surviving class set is $\hat{A}' = \{a \in \hat{A} : y_1 \in D_a(x, q)\}$, independent of y_2.*

Proof. Immediate from the closure definition $\mathrm{sig}_{C_a} = \bigcup_b \{x'\} \times D_a \times N_b$ and the fact that $y_2 \in \bar{N}$ holds for every legal branch, uniformly in a. □

Proof of Theorem T3. **($\leq$).** Theorem T2 applied to Π_{dec} gives $C^*_{\mathcal{I}}(x, A\times B) \leq C^*_{/\Pi_{\mathrm{dec}}}(x, \alpha(A\times B))$ and $\alpha(A \times B) = A$.

($\geq$), i.e., quotient $\leq$ full. Let τ be an optimal full-game strategy from $(x_0, A_0 \times B_0)$ with value $B^* = C^*_{\mathcal{I}}(x_0, A_0 \times B_0)$ (assume finite, else nothing to prove). We build a quotient strategy σ that maintains a *shadow full-game state* $(x, A \times B)$ and shadow history, with the invariant $\hat{A} = A$ (quotient class set equals the shadow decision component). Initially both equal A_0.

At a non-terminal state σ plays $q := \tau(\text{shadow history})$; this is legal on the quotient side because safety depends only on the decision coordinate and $\hat{A} = A$. The quotient adversary answers with a branch $(x', (y_1, y_2))$, where necessarily $y_1 \in \bigcup_{a\in\hat{A}} D_a$ and $y_2 \in \bar{N}$. Update the shadow as follows:

- If $B' := \{b \in B : y_2 \in N_b(x, q)\} \neq \emptyset$, advance the shadow along the *same* branch $(x', (y_1, y_2))$. This is a legal full-game branch: picking any $a \in A$ with $y_1 \in D_a$ and any $b \in B'$, the world (a, b) realizes the pair by rectangularity.
- If $B' = \emptyset$ (the quotient adversary played a nuisance observation that no shadow-alive nuisance value can realize), perform the y_2**-repair**: choose any $b \in B$ (nonempty) and any $\tilde{y}_2 \in N_b(x, q)$ (nonempty by well-formedness (iii), since q is safe and safety ignores b), and advance the shadow along $(x', (y_1, \tilde{y}_2))$. This is a legal full-game branch with survivor factors $A' = \{a \in A : y_1 \in D_a\} \neq \emptyset$ (from the origin of y_1) and $\tilde{B}' \ni b$.

In both cases the shadow's decision component updates to $A' = \{a \in A : y_1 \in D_a(x, q)\}$, which by Lemma A.4 equals the quotient survivor set $\hat{A}'$; the invariant is preserved, and costs agree step-wise. Termination: the ACT conditions are equivalent,

$$(x, A \times B) \in G \iff \bigcap_{a\in A} A_a(x) \neq \emptyset \iff (x, \hat{A}) \in G_{/\Pi},$$

because tokens depend only on a and $\hat{A} = A$. The shadow play is a legal full-game play consistent with τ, so it reaches G within cost B^*; at that same step and cost the quotient play is in $G_{/\Pi}$. Hence $C^*_{/\Pi_{\mathrm{dec}}}(x_0, A_0) \leq B^*$. □

Remark A.1 (the y_2-repair is the mathematical core of DNF). The repair step is legal precisely because the rectangular support $D_a \times N_b$ makes y_1 and y_2 independently realizable: an "impossible" nuisance observation carries no decision information, so the simulation may substitute any realizable one without changing the decision update. CE1 is exactly a failure of rectangularity (the observations of M_1, M_2 interleave across the two probes and do not factorize), and CE2 below shows that even under DNF, taking the quotient over *emergent* profile classes instead of *declared* fibers breaks Lemma A.4 (the D-table is no longer constant within a class) and completeness collapses.

Proposition A.2 (CE2: profile equivalence can merge distinct fibers)**.** *There is a DNF instance and an acc/safe-uniform partition strictly coarser than the fiber partition (obtained by merging two fibers with equal acceptability and safety profiles) for which $C^*_{\mathrm{full}} = 2$, the quotient over the merged (profile) classes has value ∞, and the quotient over declared fibers has value* 2*.*

Proof. **Construction.** Three decision fibers a_1, a_2, a_3, $|\Theta_{\mathrm{nui}}| = 1$ (so worlds may be identified with fibers), one physical state, two probes q_0, q_1, all safe, deterministic observations:

fiber	obs. under q_0	obs. under q_1	tokens
a_1	u	p	$\{t_A\}$
a_2	u	r	$\{t_B\}$
a_3	v	p	$\{t_B\}$

Full game (= declared-fiber quotient, since fibers are singletons here): from $\{a_1, a_2, a_3\} \notin G$, play q_0: branch $v \to \{a_3\} \in G$; branch $u \to \{a_1, a_2\}$, then q_1 splits $p \to \{a_1\}$, $r \to \{a_2\}$, both in G. No

single probe wins (q_0 has branch $u \to \{a_1, a_2\}$, q_1 has branch $p \to \{a_1, a_3\}$). Hence both values equal 2.

Merged profile classes: a_2 and a_3 have identical acceptability ($\{t_B\}$) and safety profiles, so $\Pi_c = \{K_A = \{a_1\}, K_B = \{a_2, a_3\}\}$ is acc/safe-uniform, with closure signatures $\mathrm{sig}_{K_B}(q_0) = \{u, v\}$, $\mathrm{sig}_{K_B}(q_1) = \{r, p\}$. From $\{K_A, K_B\}$: under q_0 branch u keeps both classes; under q_1 branch p keeps both classes. The adversary picks these, so the state is a losing fixed point and the merged quotient value is ∞. □

CE2 isolates the blocking pattern caused by merging equal acceptability and safety profiles; the independent calculation in Table 5 reproduces its three values exactly. The DNF conditions in T3 certify value preservation for the declared fiber quotient.

Proposition A.3 (compression). $|\Pi_{\mathrm{dec}}| = |\Theta_{\mathrm{dec}}|$ *independently of* $|\Theta_{\mathrm{nui}}|$*; quotient information states number at most* $|X| \cdot 2^{|\Theta_{\mathrm{dec}}|}$*. Given the declaration, building the quotient reads each table once:* $\mathrm{poly}(|\Theta_{\mathrm{dec}}|, |X|, |Q|, |Y_1|)$ *time, an exponential compression relative to* $|\mathcal{H}| = |\Theta_{\mathrm{dec}}| \cdot |\Theta_{\mathrm{nui}}|$ *when* $|\Theta_{\mathrm{nui}}|$ *is exponential in the declaration (e.g.,* 2^d*).* □

A.5 T4: INTERACTION SEPARATION AND LOWER BOUNDS

The family $F_5(d)$**.** $\Theta_{\mathrm{dec}} = \{0, 1\}$, $\Theta_{\mathrm{nui}} = \{0, \ldots, 2^d - 1\}$, a single physical state, all probes safe, deterministic observations. Probes: p_{dec} outputs the decision bit a; the equality tests e_j ($j \in \Theta_{\mathrm{nui}}$) output $[n = j]$. Tokens $A_{(a,n)} = \{t_a\}$ with $t_0 \neq t_1$. DNF holds: the D-tables depend only on a (with e_j constant), the N-tables only on n (with p_{dec} constant).

Proposition A.4 (upper bound)**.** $C^*(x_0, \mathcal{H}) = 1$ *on* $F_5(d)$.

Proof. Initially $\{t_0\} \cap \{t_1\} = \emptyset$, so $(x_0, \mathcal{H}) \notin G$ and $C^* \geq 1$. Play p_{dec}: each branch a has survivor set $\{a\} \times \Theta_{\mathrm{nui}}$, on which t_a is a common token, so the state is in G. Hence $C^* = 1$. □

Theorem T4.a (deterministic identification lower bound). *Call a strategy* identification-complete *if it commits only at singleton supports* $|V| = 1$*. On* $F_5(d)$*, every deterministic identification-complete strategy makes at least* 2^d *probes in the worst case, and* 2^d *is achieved.*

Proof (adversary argument). The adversary maintains the consistent set $V = A \times B$, initially $\{0, 1\} \times \Theta_{\mathrm{nui}}$, and answers:

- to e_j: if $|B \setminus \{j\}| \geq 1$, answer "no" (consistent, since some world with $n \neq j$ survives) and set $B \leftarrow B \setminus \{j\}$, so each such query removes at most one element of B; otherwise the answer "yes" is forced (then $B = \{j\}$ already);
- to p_{dec}: if $|A| = 2$, answer $a = 0$ and set $A \leftarrow \{0\}$; repeated queries are answered consistently and change nothing.

All answers remain consistent with every world in the current $A \times B$, and observations are deterministic, so these are the only information channels. Identification-completeness requires reaching $|A| = |B| = 1$. The component $|B|$ shrinks from 2^d to 1 only via distinct e_j-queries, each removing at most one value, so at least $2^d - 1$ of them are needed (after $2^d - 1$ "no"s the last value is inferred without a query). The component $|A|$ shrinks from 2 to 1 only via p_{dec} (equality tests carry no information about a), so at least one p_{dec} is needed. Total: $\geq 2^d$. Achievability: play p_{dec} once, then $e_0, e_1, \ldots$ in order; in the worst case $2^d - 1$ equality tests are needed, for a total of exactly 2^d. □

Corollary A.1 (interaction separation)**.** *On* $F_5(d)$ *the SD-DRR optimum commits after* 1 *probe while every deterministic identification-complete strategy needs* 2^d *in the worst case: the ratio* 2^d *is unbounded in* d*.* □

Theorem T4.b (randomized lower bound via Yao). *Every (possibly adaptive) randomized identification-complete strategy on* $F_5(d)$ *has worst-case-world expected probe count at least*

$$\frac{(N-1)(N+2)}{2N} + 1 = \Theta(2^d), \qquad N := 2^d.$$

Proof. By Yao's principle (Yao, 1977), the worst-case expected cost of a randomized strategy is at least the optimal expected cost of a *deterministic* strategy against any fixed input distribution μ. Take μ: a uniform on $\{0,1\}$ and n uniform on Θ_{nui}, independent. Observations are deterministic, so a deterministic strategy's play against a fixed world is a fixed query sequence. Equality tests carry no information about a, so every identification-complete strategy must query p_{dec} at least once on every play: expected contribution ≥ 1. For the nuisance component: conditioned on any sequence of "no" answers, the posterior of n stays uniform on the remaining values (symmetry), so adaptivity does not help and any strategy is equivalent to a fixed query order $j_1, j_2, \ldots$ over Θ_{nui}. If $n = j_i$ (probability $1/N$) the strategy spends i equality tests; if n is the last unqueried value (probability $1/N$) it spends $N-1$ (the final value is inferred for free). Hence

$$\mathbb{E}[\#e\text{-queries}] \;\geq\; \frac{1}{N}\Big(\sum_{i=1}^{N-1} i + (N-1)\Big) = \frac{(N-1)(N+2)}{2N} \;\sim\; \frac{N}{2}.$$

Adding the p_{dec} query gives the bound; combined with the deterministic upper bound of 2^d, the randomized identification-complete complexity is $\Theta(2^d)$, against 1 for the SD-DRR optimum. □

Remark A.2 (Query classes and resource measures). (a) The separation is against the *identification-complete* strategy class. Region-level commitment heuristics (e.g., worst-case elimination greedy) also solve F_5 at cost 1 (they select p_{dec} and then hold a region certificate); those archetypes are separated instead by the F2 (information window) and F3 (state-dependent availability) families (Fig. 5). (b) Representation cost is measured for the belief-support solver: the unquotiented implementation reaches its 5×10^5-state guard at $d \geq 5$, while the fiber quotient uses 3 states. Theorems T4.a/T4.b bound interaction complexity. (c) The distribution and symmetry argument of T4.b uses the single-value-reveal structure of equality tests. For arbitrary subset-membership queries, the corresponding identification lower bound is $\Omega(\log N)$. The query family determines the separation.

A.6 Refinement and the passive-data requirement

Partition refinement uses representation and computation to reduce conservative planning loss; passive observations shrink the compatible decision set. The following results track these resources separately.

Pessimistic quotients. Consider a finite instance $\mathcal{I} = (\mathcal{H}, X, Q, Y, \mathrm{sig}, \mathrm{Safe}, A, c)$ with strictly positive costs $c(x,q)$ and nonempty supports for safe triples. For an arbitrary partition Π of $\mathcal{H}$, write $\mathcal{I}/\Pi$ for the pessimistic quotient of Proposition A.18: for each class $K \in \Pi$,

$$A_K(x) = \bigcap_{w \in K} A_w(x), \qquad \mathrm{Safe}_K(x,q) \iff \forall w \in K : \mathrm{Safe}_w(x,q),$$

and, whenever the class action is safe,

$$\mathrm{sig}_K(x,q) = \bigcup_{w \in K} \mathrm{sig}_w(x,q).$$

Costs are unchanged. Let $\alpha_\Pi(V) = \{K \in \Pi : K \cap V \neq \varnothing\}$ for any nonempty $V \subseteq \mathcal{H}$. The next two results hold for arbitrary partitions with these pessimistic semantics.

Lemma A.5 (Version-set monotonicity). *For any $\varnothing \neq V \subseteq U \subseteq \mathcal{H}$ and $x \in X$,*

$$C^*_{\mathcal{I}}(x,V) \leq C^*_{\mathcal{I}}(x,U).$$

Proof. If the right-hand side is infinite, the claim is immediate. Otherwise, take any strategy that resolves (x,U) within a worst-case cost bound r. Execute the same strategy from (x,V) while maintaining its shadow history and compatible set U_t. Initially $V \subseteq U$. An action safe for all of U_t is safe for all of V_t. Every outcome possible from V_t is possible from U_t; applying the same consistency update preserves $V_{t+1} \subseteq U_{t+1}$. A token accepted by every member of U_t is accepted by every member of V_t. Thus each execution from V follows a legal shadow execution from U with identical action costs and reaches authorization within r. Taking the infimum over resolving strategies proves the inequality. □

Corollary A.2 (Refinement monotonicity)**.** *If Π_f refines Π_c, then for every nonempty $V \subseteq \mathcal{H}$,*

$$C^*_{\mathcal{I}}(x, V) \le C^*_{\mathcal{I}/\Pi_f}(x, \alpha_{\Pi_f}(V)) \le C^*_{\mathcal{I}/\Pi_c}(x, \alpha_{\Pi_c}(V)).$$

Consequently, every state certified resolvable within budget B under Π_c is also resolvable within B under Π_f. The mass of such states under any fixed evaluation distribution is nondecreasing under refinement.

Proof. The first inequality is the one-sided soundness of pessimistic grouping. For the second, let $\widehat{\Pi}_c$ partition the worlds of $\mathcal{I}/\Pi_f$ by assigning each fine class to its containing coarse class. Quotienting a second time gives

$$(\mathcal{I}/\Pi_f)/\widehat{\Pi}_c \;\cong\; \mathcal{I}/\Pi_c.$$

Indeed, intersections of the fine-class acceptability intersections are exactly the coarse-class intersections, and the same associativity holds for safety conjunctions. At a coarse-safe action every involved fine class is safe, so the union of fine-class signatures is exactly the coarse signature. States and costs are unchanged. The projected initial fine-class set becomes $\alpha_{\Pi_c}(V)$. Apply one-sided soundness to this second quotient. Thresholding the inequalities at B gives set inclusion; integrating the corresponding indicators against a fixed distribution gives the final statement. □

Pessimistic-quotient soundness orders nested partitions by resolution cost. Retaining uncertain members in separate classes gives a nonincreasing upper bound on the full game's value. Refinement work is a separate computational resource.

Offline observation model. For the remaining results assume the online instance satisfies T3's DNF conditions: $\mathcal{H} = \Theta_{\mathrm{dec}} \times \Theta_{\mathrm{nui}}$, $A_{(a,b)} = A_a$, $\mathrm{Safe}_{(a,b)} = \mathrm{Safe}_a$, and

$$\mathrm{sig}_{(a,b)}(x, q) = \{x'(x, q)\} \times \big(D_a(x, q) \times N_b(x, q)\big),$$

with nonempty factors at safe triples. The state successor $x'(x, q)$ and action cost do not depend on (a, b). The online factors may be set-valued. The decision game $\mathcal{I}_{\mathrm{dec}}$ retains $X, Q, A_a, \mathrm{Safe}_a, c$ and uses the signatures $\{x'(x, q)\} \times D_a(x, q)$; its value is denoted C^*_{dec}.

The *offline* observation interface is deterministic and realizable. Fix a true world $(a_\star, b_\star) \in \mathcal{H}$. Each logged query e_i is observed, is sampled independently from a fixed distribution ν on a finite query set $\mathcal{E}$, and returns

$$o_i = \big(h_{\mathrm{dec}}(a_\star, e_i), h_{\mathrm{nui}}(b_\star, e_i)\big).$$

The query law is independent of the true world, and the deterministic maps are those used by the consistency learner. Samples are historical contextual or reset observations, with online deployment starting at fixed x_0. Historical collection and its safety protocol are specified separately from the online resolving policy. Each observation costs one query.

After N samples, exact consistency gives

$$\begin{aligned} A_N &= \{a : h_{\mathrm{dec}}(a, e_i) = h_{\mathrm{dec}}(a_\star, e_i),\ 1 \le i \le N\}, \\ B_N &= \{b : h_{\mathrm{nui}}(b, e_i) = h_{\mathrm{nui}}(b_\star, e_i),\ 1 \le i \le N\}, \end{aligned} \qquad V_N = A_N \times B_N.$$

In particular $a_\star \in A_N$ and $b_\star \in B_N$ for every data realization. The letter B_N denotes a nuisance set; the scalar B below denotes the deployment probe-cost budget.

Lemma A.6 (Product starts)**.** *Under the stated DNF conditions, for every nonempty product $A \times B'$ and every $x \in X$,*

$$C^*_{\mathcal{I}}(x, A \times B') = C^*_{\mathrm{dec}}(x, A).$$

This includes product sets produced by the separate offline interface, whether or not they are reachable by online probes from $(x_0, \mathcal{H})$.

Proof. Consider the optimal worst-case cost of reaching authorization in at most t probes, assigning value $+\infty$ if that is impossible. For $t = 0$, authorization depends only on $\bigcap_{a \in A} A_a(x)$, so the full and decision games agree. Suppose their truncated values agree at horizon t for all product sets. The admissible action sets agree because safety depends only on a. After an admissible action and outcome (y_1, y_2) the full surviving set is $A' \times B''$, where

$$A' = \{a \in A : y_1 \in D_a(x, q)\}, \qquad B'' = \{b \in B' : y_2 \in N_b(x, q)\}.$$

Its decision successor is precisely the successor for y_1. Conversely, every possible decision observation y_1 extends to a full outcome: choose any $b \in B'$ and $y_2 \in N_b(x, q)$, which exists by nonemptiness. Thus, by the induction hypothesis, the maxima of successor values in the two games agree, as do action costs and the minimum over actions. This proves equality at horizon $t+1$. Finally, any strategy with a finite worst-case cost bound makes only finitely many probes, uniformly over branches, since the finite instance has a strictly positive minimum action cost. Taking the infimum over horizons therefore gives equality of the untruncated values, including $+\infty$. □

Minimal budget obstructions. Fix a true decision coordinate a and finite budget $B \geq 0$. Define

$$\mathcal{F}_B(a) = \{A \subseteq \Theta_{\mathrm{dec}} : a \in A,\ C^*_{\mathrm{dec}}(x_0, A) \leq B\}.$$

This family is downward closed among sets containing a by Lemma A.5. Its minimal excluded sets are

$$\mathcal{O}_B(a) = \left\{ O \ni a : \begin{array}{l} C^*_{\mathrm{dec}}(x_0, O) > B, \\ C^*_{\mathrm{dec}}(x_0, U) \leq B \quad \text{for every } a \in U \subsetneq O \end{array} \right\}.$$

These sets are determined by the online game, start and budget. They index the sample-complexity analysis; the deployment gate takes the data-consistent game as its input. This construction allows $\{a\}$ itself to be an obstruction if even a known world cannot meet the budget. Write

$$D_a(e) = \{a' : h_{\mathrm{dec}}(a', e) \neq h_{\mathrm{dec}}(a, e)\}, \qquad H_a(O) = \{e : D_a(e) \cap O \neq \varnothing\}.$$

Thus $H_a(O)$ is the set of passive queries that break O, and $s_\nu(a, O) = \nu(H_a(O))$. The product-start lemma transfers all following budget statements to the full world set V_N.

Proposition A.5 (T5: exact obstruction event and finite-sample bounds)**.** *Under the offline observation model above, for every realized dataset,*

$$\{C^*_{\mathrm{dec}}(x_0, A_N) > B\} = \bigcup_{O \in \mathcal{O}_B(a)} \{O \subseteq A_N\}. \tag{OC}$$

For a nonempty obstruction family and every $N \geq 1$,

$$\max_{O \in \mathcal{O}_B(a)} (1 - s_\nu(a, O))^N \leq P_N(B) \leq \min\left\{1, \sum_{O \in \mathcal{O}_B(a)} (1 - s_\nu(a, O))^N\right\}. \tag{OB}$$

Put $M_B = |\mathcal{O}_B(a)|$ and $s_B = \min_{O \in \mathcal{O}_B(a)} s_\nu(a, O)$. For $0 < s_B < 1$, $\rho_B = -\log(1 - s_B)$ satisfies

$$\lim_{N \to \infty} -\frac{1}{N} \log P_N(B) = \rho_B. \tag{OR}$$

For $\eta \in (0, 1)$ define $N_\eta(B) = \min\{N \geq 0 : P_N(B) \leq \eta\}$. Then

$$\left\lceil \frac{\log(1/\eta)}{\rho_B} \right\rceil \leq N_\eta(B) \leq \left\lceil \frac{\log(M_B/\eta)}{\rho_B} \right\rceil. \tag{ON}$$

If $\mathcal{O}_B(a) = \varnothing$, $P_N(B) = 0$ for all $N \geq 0$. If $s_B = 0$, $P_N(B) = 1$ for all $N \geq 0$. If $s_B = 1$, $P_0(B) = 1$ and $P_N(B) = 0$ for $N \geq 1$.

Proof. Realizability gives $a \in A_N$ for every dataset. If A_N is budget infeasible, finiteness supplies an inclusion-minimal infeasible subset containing a, which lies in $\mathcal{O}_B(a)$. Conversely, if an obstruction O survives inside A_N, monotonicity gives $C^*_{\mathrm{dec}}(x_0, A_N) \geq C^*_{\mathrm{dec}}(x_0, O) > B$. This proves (OC).

An obstruction survives precisely when no query lies in $H_a(O)$. Independence across the N samples therefore gives $\Pr[O \subseteq A_N] = (1 - s_\nu(a, O))^N$. Each survival event implies failure, yielding the maximum lower bound; the union bound gives the upper. This argument permits arbitrary dependence between the survival events of different obstructions. As the obstruction family is finite, the smallest hit probability is attained, so

$$e^{-N\rho_B} \leq P_N(B) \leq M_B e^{-N\rho_B}.$$

Taking logarithms and dividing by N proves the exact rate. The lower bound is necessary for $P_N(B) \leq \eta$, and the upper bound is sufficient, proving (ON). Here $P_0(B) = 1$ whenever an obstruction exists because $A_0 = \Theta_{\mathrm{dec}}$.

An empty obstruction family means the full decision set is feasible. When $s_B = 0$, some obstruction survives with probability one for every sample size, which forces failure. When $s_B = 1$, each query breaks every obstruction with probability one: there are finitely many obstructions, so their probability-one events intersect with probability one. This proves all boundary cases. □

A three-world construction: pairs do not suffice. Let $\Theta_{\mathrm{dec}} = \{a_0, a_1, a_2\}$ and let the nuisance coordinate be a singleton. There is one physical state x_0 and three distinct acceptable tokens, one per world. Probe q_i costs one, is unsafe only in world a_i, leaves x_0 unchanged, and reveals the world among those in which it is safe. Every pair has a safe probe, namely the probe indexed by its excluded world, and resolves at cost one. The triple has neither a common token nor a universally safe probe, hence its cost is $+\infty$. The construction satisfies DNF, with singleton observation supports at safe triples.

For true a_0 and budget $B = 1$, the only obstruction is the triple. Take passive queries e_0, e_1, e_2 with probabilities $1 - p_1 - p_2, p_1, p_2$, where $p_1, p_2 > 0$ and $p_1 + p_2 < 1$. All responses at a_0 are zero. Only a_1 responds one at e_1; only a_2 responds one at e_2; other responses are zero. Therefore

$$P_N(1) = (1 - p_1 - p_2)^N. \tag{EX1}$$

At budget zero the two pairs containing a_0 are the minimal obstructions. Both rivals must be removed, giving

$$P_N(0) = (1 - p_1)^N + (1 - p_2)^N - (1 - p_1 - p_2)^N. \tag{EX0}$$

The asymptotic rate changes from $-\log(1 - \min\{p_1, p_2\})$ to $-\log(1 - p_1 - p_2)$ when one safe probe becomes available. The improvement is caused by a change in the obstruction family, with no change to the world count or passive channel. The displayed probabilities follow analytically from this construction.

Corollary A.3 (An arbitrarily large passive-data separation)**.** *In the three-world construction, fix $\eta \in (0, 1)$ and $p_2 \in (0, 1)$. As $p_1 \downarrow 0$ with $p_1 + p_2 < 1$, $N_\eta(0)/N_\eta(1) \to +\infty$. The hypothesis class, acceptable tokens, and online probes remain unchanged.*

Proof. The pair $\{a_0, a_1\}$ survives with probability $(1 - p_1)^N$. Together with (EX1), this gives

$$N_\eta(0) \geq \left\lceil \frac{\log(1/\eta)}{-\log(1 - p_1)} \right\rceil, \qquad N_\eta(1) \leq \left\lceil \frac{\log(1/\eta)}{-\log(1 - p_2)} \right\rceil.$$

The first bound diverges as $p_1 \downarrow 0$, while the second is a fixed finite positive integer. One safe online probe substitutes for the increasingly rare passive distinction, producing an unbounded sample-ratio separation as $p_1 \downarrow 0$ at fixed p_2. □

Corollary A.4 (Budget frontier and nuisance invariance)**.** *At fixed true coordinate, online game and passive channel, $P_N(B)$ is nonincreasing in B at every N. Assign decay rate $\rho_B = +\infty$ when the failure probability is eventually zero and $\rho_B = 0$ when it is always one. Then ρ_B is nondecreasing in B. It is piecewise constant between the finitely many finite values $C^*_{\mathrm{dec}}(x_0, A)$ for $A \ni a$. Under DNF these quantities are unchanged by adding nuisance coordinates while holding the decision game, decision observation map, and query law fixed.*

Proof. For $B_1 \leq B_2$, the failure event at B_2 is contained in that at B_1 for every realized consistency set. Taking probabilities proves monotonicity. Applying the exact rate limit in Proposition A.5, including its boundary cases, proves the rate ordering. A budget change can alter the feasible family only when it crosses the resolution cost of one of the finitely many decision sets, proving the step structure. Lemma A.6 removes the nuisance set from the value, and the offline product model removes it from the law of A_N. This proves the final assertion. The qualification on the channel is essential: increasing the frequency of nuisance-only queries changes the hit probabilities. □

Convex design, dual witnesses, and finite-sample optimization

Fix the finite online game, start, budget B, and deterministic passive channel of T5. Let $\mathcal{E}$ be a nonempty query menu, $K = |\mathcal{E}|$, and $\mathcal{J}_B = \{(a, O) : a \in \Theta_{\mathrm{dec}}, O \in \mathcal{O}_B(a)\}$, of size $m > 0$. Define

$$R_{(a,O),e} = \mathbf{1}\{e \in H_a(O)\}, \qquad \Delta_d = \{v \in \mathbb{R}^d : v \geq 0,\ \mathbf{1}^\top v = 1\}.$$

The incidence matrix is fixed before choosing the query law. It encodes which observations break the budget obstructions of the safe resolution game. Sampling laws below are fixed i.i.d. laws, independent of the true world. Classical linear-programming duality and convex optimality conditions are used in their usual sense (Boyd & Vandenberghe, 2004, Chapters 4–5).

Proposition A.6 (T6: primal and dual designs)**.** *Both optima in*

$$\begin{aligned} z_B^* &= \max_{\nu,z}\{z : R\nu \geq z\mathbf{1},\ \nu \in \Delta_K\} \\ &= \min_{\lambda,t}\{t : R^\top\lambda \leq t\mathbf{1},\ \lambda \in \Delta_m\} \end{aligned} \tag{A.1}$$

are attained and equal. They lie in $[0,1]$*. For every feasible* ν, λ*,* $\ell(\nu) \leq z_B^* \leq u(\lambda)$*, with the definitions in T6. For optimal* ν^*, λ^**,*

$$\begin{aligned} \lambda_j^* > 0 &\implies (R\nu^*)_j = z_B^*, \\ \nu_e^* > 0 &\implies (R^\top\lambda^*)_e = z_B^*. \end{aligned} \tag{A.2}$$

If $z_B^* > 0$*, the covering and packing programs satisfy*

$$\tau_B := \min_{w\geq 0,\ Rw\geq\mathbf{1}} \mathbf{1}^\top w = \max_{v\geq 0,\ R^\top v\leq\mathbf{1}} \mathbf{1}^\top v = \frac{1}{z_B^*}. \tag{A.3}$$

If $z_B^* = 0$*, the covering problem is infeasible and* $\tau_B = +\infty$*.*

Proof. For the maximization program, use multipliers $\lambda \geq 0$ for $R\nu - z\mathbf{1} \geq 0$ and t for $\mathbf{1}^\top\nu = 1$. With z free and $\nu \geq 0$, the Lagrangian is

$$L = z(1 - \mathbf{1}^\top\lambda) + \nu^\top(R^\top\lambda - t\mathbf{1}) + t.$$

Its supremum is finite precisely when $\mathbf{1}^\top\lambda = 1$ and $R^\top\lambda \leq t\mathbf{1}$, in which case it is t. This is the stated dual. The primal is feasible and bounded; finite-dimensional LP strong duality gives equal attained optima. Equivalently, both reduced objectives are continuous on compact simplexes. Since R is binary, the values belong to $[0,1]$. Weak duality also follows directly from

$$\ell(\nu) \leq \lambda^\top R\nu \leq u(\lambda).$$

At an optimal pair both bounds equal z_B^*. Equality in these weighted averages gives (A.2).

For positive z_B^*, $w = \nu^*/z_B^*$ is a fractional cover of total weight $1/z_B^*$. Conversely, any cover w has total weight $T > 0$, and $\nu = w/T$ has every hit probability at least $1/T$. Thus $T \geq 1/z_B^*$, proving the covering identity. Its LP dual is the displayed packing program; alternatively $v = \lambda^*/z_B^*$ attains the same value and is feasible. If a finite cover existed when $z_B^* = 0$, normalizing it would give a strictly positive hit rate, a contradiction. □

How the dual bounds data requirements. Let $M_{\max} = \max_a |\mathcal{O}_B(a)| \geq 1$ and $P_N^{\max}(\nu) = \max_a P_N(B \mid a, \nu)$. For any $N \geq 1$, T5 gives

$$(1 - \ell(\nu))^N \leq P_N^{\max}(\nu) \leq \min\{1, M_{\max}(1 - \ell(\nu))^N\}.$$

The lower bound uses a least-hit obstruction and its associated true world. Therefore any candidate primal–dual pair yields

$$(1 - u(\lambda))^N \leq \inf_{\tilde{\nu}\in\Delta_K} P_N^{\max}(\tilde{\nu}) \leq P_N^{\max}(\nu) \leq \min\{1, M_{\max}(1 - \ell(\nu))^N\}. \tag{A.4}$$

For $0 < \ell(\nu) \leq u(\lambda) < 1$ and target $\eta \in (0,1)$,

$$N \geq \left\lceil \frac{\log(1/\eta)}{-\log(1 - u(\lambda))} \right\rceil \quad \text{is necessary for any fixed i.i.d. design,}$$

whereas

$$N \geq \left\lceil \frac{\log(M_{\max}/\eta)}{-\log(1 - \ell(\nu))} \right\rceil \quad \text{suffices for the candidate } \nu.$$

In particular, optimizing the fixed-design worst-world exponent gives $-\log(1 - z_B^*) = -\log(1 - 1/\tau_B)$ when $0 < z_B^* < 1$. At $z_B^* = 1$ an optimal design breaks every obstruction in one sample. At $z_B^* = 0$ some obstruction has no distinguishing query, so a corresponding true world fails budget certification at every sample size. If $\mathcal{J}_B$ is empty, every world already resolves without passive data and the optimization is unnecessary.

The computable gap $u(\lambda) - \ell(\nu)$ certifies the remaining error in the hit-rate optimization. Its primal side supplies an achievable sample bound; its dual side limits every admissible allocation. The guarantee concerns budget certification in the supplied game, with the same truth-containment assumptions as T5.

An exact budget-dependent design example. Use the preceding three-world construction and queries e_0, e_1, e_2. The query e_0 is uninformative; e_1 distinguishes a_1 from the other worlds and e_2 distinguishes a_2. At $B = 0$, the distinct obstruction-hit rows are

$$R_0 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}.$$

Repeated rows across true worlds do not change the LP. The primal $\nu^* = (0, 1/2, 1/2)$ and dual weights $(1/2, 1/2, 0)$ attain $z_0^* = 1/2$, hence $\tau_0 = 2$ and rate $\log 2$. At $B = 1$ every pair is safely resolvable, leaving only the triple; its hit row is $(0, 1, 1)$ for every true world. Consequently $z_1^* = 1$ and $\tau_1 = 1$: any one informative sample leaves a safely resolvable pair. This analytic example changes the obstruction geometry by changing online budget, with the world class and passive menu fixed.

Proposition A.7 (Convex finite-sample design and an optimality certificate)**.** *For an integer $N \geq 1$, define*

$$F_{a,N}(\nu) = \sum_{O \in \mathcal{O}_B(a)} [1 - (R\nu)_{a,O}]^N, \qquad U_N(\nu) = \max_a F_{a,N}(\nu).$$

The function U_N is convex and continuous on Δ_K, attains its minimum, and bounds $P_N^{\max}(\nu)$ from above. Choose any active world $a \in \arg\max_b F_{b,N}(\nu)$. Then

$$g = -N \sum_{O \in \mathcal{O}_B(a)} [1 - (R\nu)_{a,O}]^{N-1} R^\top_{(a,O),:} \tag{A.5}$$

is a subgradient, as is any convex combination over active worlds. For any such g, the computable first-order gap satisfies

$$0 \leq U_N(\nu) - \min_{\tilde{\nu} \in \Delta_K} U_N(\tilde{\nu}) \leq \Gamma_N(\nu, g) := g^\top \nu - \min_e g_e. \tag{A.6}$$

Proof. For $N = 1$, $(1 - s)^N$ is affine. For $N \geq 2$, its second derivative is $N(N-1)(1-s)^{N-2} \geq 0$ on $[0, 1]$. Composition with each affine hit probability, summation and finite maximization preserve convexity. Continuity and compactness give attainment. The union bound in T5 gives $P_N^{\max}(\nu) \leq U_N(\nu)$.

An active $F_{a,N}$ agrees with U_N at ν and lies below it everywhere. Its gradient therefore supports U_N and is a subgradient. Differentiation gives (A.5), using coefficient one for the zeroth power when $N = 1$. For any $\tilde{\nu} \in \Delta_K$,

$$U_N(\tilde{\nu}) \geq U_N(\nu) + g^\top(\tilde{\nu} - \nu) \geq U_N(\nu) + \min_e g_e - g^\top \nu.$$

Taking the minimum proves (A.6); the gap is nonnegative because $g^\top \nu$ is a convex combination of the coordinates of g. □

The negative gradient weights each query by the obstruction survival mass it would remove. The power $N - 1$ adapts this weighting to the sample budget, whereas the asymptotic LP retains only the least-hit combinations. Balancing tied active worlds through an optimal subgradient tightens the first-order gap certificate. Optimization targets the uncapped union bound U_N; the reported probability bound is $\min\{1, U_N\}$. The union bound applies to arbitrarily dependent obstruction-survival events.

The convex programs take an explicit obstruction matrix as input. Matrix construction enumerates minimal infeasible candidate sets through resolution queries. The resulting allocation and dual witness certify the query distribution for that continuation game.

Budget-resolvable envelopes. Fix a deployment budget $B \geq 0$. For a true decision coordinate a, call a deterministic set $E_a \subseteq \Theta_{\mathrm{dec}}$ a B-resolvable envelope if

$$a \in E_a, \qquad C^*_{\mathrm{dec}}(x_0, E_a) \leq B.$$

The envelope is an analysis witness fixed independently of the data. The deployment algorithm solves its data-consistent game. An envelope may contain incompatible immediate choices that become resolvable through an online continuation of cost at most B. Define

$$p_\nu(a, a') = \Pr_{e \sim \nu}\big[h_{\mathrm{dec}}(a, e) \neq h_{\mathrm{dec}}(a', e)\big].$$

Proposition A.8 (Finite passive-data bound for budgeted resolution)**.** *Under the preceding assumptions, fix a true world (a, b) and any B-resolvable envelope E_a. For every integer $N \geq 0$,*

$$\Pr[C^*_{\mathcal{I}}(x_0, V_N) > B] = \Pr[C^*_{\mathrm{dec}}(x_0, A_N) > B] \leq \min\left\{1,\ \sum_{a' \notin E_a} (1 - p_\nu(a, a'))^N\right\}. \tag{PB}$$

The probability is over the offline queries, conditional on the fixed true world. If $K_a = |\Theta_{\mathrm{dec}} \setminus E_a| > 0$ and $p_a = \min_{a' \notin E_a} p_\nu(a, a') > 0$, then for $\eta \in (0, 1)$,

$$N \geq \left\lceil \frac{\log K_a + \log(1/\eta)}{p_a} \right\rceil \tag{PS}$$

suffices for budget-feasible resolution with probability at least $1 - \eta$. If $K_a = 0$, this holds already at $N = 0$ with probability one. When the exact budgeted gate proceeds, all its probes and its eventual commitment are valid for the true world on every allowed online branch.

Proof. Lemma A.6 gives the equality in (PB). If $A_N \subseteq E_a$, then Lemma A.5 yields

$$C^*_{\mathrm{dec}}(x_0, A_N) \leq C^*_{\mathrm{dec}}(x_0, E_a) \leq B.$$

Therefore failure of budget-feasible resolution implies survival of at least one coordinate $a' \notin E_a$. Such a coordinate survives exactly when all N queries fail to separate it from a. The queries are i.i.d., so its survival probability is $(1 - p_\nu(a, a'))^N$. The union bound proves (PB); it needs no independence between different competitors' survival events. With $p_a > 0$, the sum is at most $K_a e^{-N p_a}$, giving (PS). An empty competitor set has sum zero, giving the zero-data case. For $N = 0$, each survival probability is one, including when $p_\nu(a, a') = 1$.

Realizability ensures that V_N contains the true world. The exact gate's admissibility condition and consistency updates preserve that invariant on every possible true-world branch. A commonly accepted terminal token is therefore accepted by the true world, and the cost-to-authorization certificate bounds its online cost by B. The gate's structural invariant establishes commitment validity; η bounds the probability that the offline data leave resolution cost above B. □

Proposition A.9 (The passive-information boundary)**.** *Define the ν-indistinguishable decision class*

$$L_\nu(a) = \{a' \in \Theta_{\mathrm{dec}} : p_\nu(a, a') = 0\}.$$

For the same fixed true world, almost surely there exists a finite random N_0 such that $A_N = L_\nu(a)$ for every $N \geq N_0$. Consequently:

1. *If $C^*_{\mathrm{dec}}(x_0, L_\nu(a)) \leq B$, the exact gate becomes budget-feasible after finitely many passive samples almost surely, and its probability of budget feasibility tends to one.*
2. *If $C^*_{\mathrm{dec}}(x_0, L_\nu(a)) > B$, then almost surely it is not budget-feasible at any sample size N.*

Proof. For $a' \in L_\nu(a)$, the event that one query separates a and a' has probability zero. By countability of the query sequence and finiteness of the decision class, none of these coordinates is ever eliminated, almost surely. For each $a' \notin L_\nu(a)$, the separation probability is positive, so

$$\Pr[a' \text{ survives every sample}] = \lim_{N \to \infty} (1 - p_\nu(a, a'))^N = 0.$$

There are finitely many such coordinates. Each has an almost surely finite elimination time, and their maximum is almost surely finite. This proves the asserted eventual equality. The first consequence then follows from Lemma A.6; its probability statement also follows from (PB) with $E_a = L_\nu(a)$. For the second, $L_\nu(a) \subseteq A_N$ for all N almost surely, and monotonicity gives

$$C^*_{\mathrm{dec}}(x_0, A_N) \geq C^*_{\mathrm{dec}}(x_0, L_\nu(a)) > B.$$

□

Remark A.3 (Budget-dependent sample requirements). The envelope need not be a singleton. At $B = 0$ it can be any common authorization region; for $B > 0$ it can include worlds separated by safe online probes. If the full decision class is B-resolvable, no offline samples are needed, even if full-world identification remains impossible. Conversely, a nonresolvable $L_\nu(a)$ identifies an observation or online-budget barrier that additional samples from the same passive channel cannot remove.

At fixed h_{dec} and ν, the bound has no factor $|\Theta_{\mathrm{nui}}|$. If $|\Theta_{\mathrm{dec}}| \le 2^{d_{\mathrm{dec}}}$ and the relevant separation probabilities are bounded below by $p_0 > 0$, a sufficient sample count is $\lceil (d_{\mathrm{dec}} \log 2 + \log(1/\eta))/p_0 \rceil$ whenever a suitable envelope exists. The qualification matters: expanding a uniform query menu with nuisance-only queries can reduce p_0 and increase sample requirements. The obstruction lower bound concerns exact budget certification under the specified passive channel and realizable world class. The envelope construction quantifies how online resolving capacity changes the required passive evidence in this setting.

A.7 Convex continuation responses on general history spaces

This extension concerns bounded finite-horizon decision costs. Histories may lie in an arbitrary measurable space, and transitions may be stochastic. Let $z = \phi(h)$ be a measurable representation. At stage t, the true and represented problems have the same admissible actions, bounded costs $c_t(h, a)$ and $\widehat{c}_t(z, a)$, and nonempty sets of transition kernels $\mathcal{P}_t(h, a)$ and $\widehat{\mathcal{P}}_t(z, a)$. Nature selects a kernel separately at each visited history, as in the robust Bellman recursion below. All functions used in expectations are bounded and measurable; assume the indicated minima have measurable minimizing selectors. Finite candidate-action sets satisfy this selection requirement directly.

For a bounded test vector $\psi_{t+1} : \mathcal{Z} \to \mathbb{R}^{k_{t+1}}$, define the nonempty compact convex response sets

$$K_t(h, a) = \overline{\mathrm{conv}} \left\{ \mathbb{E}_P[\psi_{t+1}(\phi(h'))] : P \in \mathcal{P}_t(h, a) \right\}, \tag{A.7}$$

$$\widehat{K}_t(z, a) = \overline{\mathrm{conv}} \left\{ \mathbb{E}_{\widehat{P}}[\psi_{t+1}(z')] : \widehat{P} \in \widehat{\mathcal{P}}_t(z, a) \right\}. \tag{A.8}$$

Boundedness of the tests gives compactness after closure in the finite dimensional response space, for an arbitrary measurable world class. Let $\sigma_K(v) = \sup_{x \in K} v^\top x$, and let $d_{H,\infty}$ be Hausdorff distance for the ℓ_∞ norm.

Lemma A.7 (Dual characterization of response error)**.** *For nonempty compact convex sets $K, L \subset \mathbb{R}^k$,*

$$d_{H,\infty}(K, L) = \sup_{\|v\|_1 \le 1} |\sigma_K(v) - \sigma_L(v)|.$$

Thus equality of the response sets is necessary and sufficient for equality of every bounded linear continuation evaluation in this test space. If only one true and one modeled kernel are specified, this distance is the maximum coordinate error between their mean responses.

Proof. For $\epsilon \ge 0$, the inclusion $K \subseteq L + \epsilon B_\infty$ implies $\sigma_K(v) \le \sigma_L(v) + \epsilon \|v\|_1$ for every v. Conversely, if this inclusion fails, separation of a point of K from the compact convex set $L + \epsilon B_\infty$ yields a vector violating the displayed inequality. We used $\sigma_{L+\epsilon B_\infty}(v) = \sigma_L(v) + \epsilon \|v\|_1$. Apply the same argument to the reverse inclusion and take the least ϵ. Positive homogeneity restricts the separating vectors to the unit ℓ_1 ball. A singleton kernel set gives a singleton mean-response set, whose Hausdorff distance is its pointwise norm. These are standard support-function and separation tools (Boyd & Vandenberghe, 2004). □

The true recursion is

$$V_t(h) = \min_a \left\{ c_t(h, a) + \sup_{P \in \mathcal{P}_t(h,a)} \mathbb{E}_P[V_{t+1}(h')] \right\}, \qquad V_H = c_H,$$

with the corresponding hatted recursion on z. Write $\widehat{Q}_t(z, a)$ for the quantity inside the hatted minimum.

Proposition A.10 (T7: continuation-response transfer)**.** *Uniformly over the histories and admissible actions in these recursions, suppose that*

$$|c_t(h,a) - \widehat{c}_t(\phi(h),a)| \le \epsilon_{c,t}, \qquad d_{H,\infty}(K_t(h,a), \widehat{K}_t(\phi(h),a)) \le \delta_t,$$
$$|c_H(h) - \widehat{c}_H(\phi(h))| \le \epsilon_H.$$

Suppose also that the represented optimal continuation values satisfy

$$\widehat{V}_{t+1}(z) = b_{t+1} + v_{t+1}^\top \psi_{t+1}(z) + r_{t+1}(z), \quad \|v_{t+1}\|_1 \le L_{t+1}, \quad \|r_{t+1}\|_\infty \le \rho_{t+1}.$$

Set $e_t = \epsilon_{c,t} + L_{t+1}\delta_t + 2\rho_{t+1}$ *and* $E_t = \epsilon_H + \sum_{j=t}^{H-1} e_j$*. Then*

$$|V_t(h) - \widehat{V}_t(\phi(h))| \le E_t.$$

If π *selects an action with* $\widehat{Q}_t(\phi(h), \pi_t(h)) \le \widehat{V}_t(\phi(h)) + \eta_t$ *at every history, its true robust cost obeys*

$$V_t^\pi(h) - V_t(h) \le 2E_t + \sum_{j=t}^{H-1} \eta_j.$$

In particular, if the true best action is unique and its Q_t *gap exceeds* $2E_t + \eta_t$*, this selected action is the true best action.*

Proof. For any bounded f, g and any nonempty set of probability kernels, the difference of their worst expected values is at most $\|f - g\|_\infty$. By Lemma A.7, the response assumption gives

$$\left| \sup_P \mathbb{E}_P[\widehat{V}_{t+1}(\phi(h'))] - \sup_{\widehat{P}} \mathbb{E}_{\widehat{P}}[\widehat{V}_{t+1}(z')] \right| \le L_{t+1}\delta_t + 2\rho_{t+1}.$$

The constant b_{t+1} cancels; the two residual expectations each contribute at most ρ_{t+1}. Inductively comparing the true Q_t to $\widehat{Q}_t$ therefore gives an error at most $E_{t+1} + e_t = E_t$. Taking minima is nonexpansive, which proves the value bound starting with the terminal error $E_H = \epsilon_H$.

For policy performance, compare V_t^π directly with $\widehat{V}_t$, rather than accumulating the preceding Q errors. If $V_{t+1}^\pi(h') \le \widehat{V}_{t+1}(\phi(h')) + D_{t+1}$, the same one-step response comparison and the action-selection inequality give

$$V_t^\pi(h) \le \widehat{V}_t(\phi(h)) + D_{t+1} + e_t + \eta_t.$$

With $D_H = \epsilon_H$, this yields $D_t = E_t + \sum_{j=t}^{H-1} \eta_j$. Combining it with $\widehat{V}_t(\phi(h)) - V_t(h) \le E_t$ proves the performance bound. The approximation condition concerns the model's optimal continuation value. Finally, the uniform Q comparison gives $Q_t(h, \pi_t(h)) - \min_a Q_t(h,a) \le 2E_t + \eta_t$. A larger unique-action gap excludes every suboptimal action. □

Corollary A.5 (Linear programs for response distance)**.** *Let* $K = \operatorname{conv}\{p_1, \dots, p_n\}$ *and* $L = \operatorname{conv}\{q_1, \dots, q_m\}$ *be nonempty response polytopes in* $\mathbb{R}^k$*. With* $Q = [q_1 \ \cdots \ q_m]$*, set*

$$d(p, L) = \min_{\lambda,\tau} \left\{ \tau : -\tau \mathbf{1} \le p - Q\lambda \le \tau \mathbf{1}, \quad \lambda \ge 0, \quad \mathbf{1}^\top \lambda = 1, \quad \tau \ge 0 \right\}.$$

Then

$$d_{H,\infty}(K, L) = \max \left\{ \max_i d(p_i, L), \ \max_j d(q_j, K) \right\}.$$

Thus $n + m$ *linear programs compute the response error exactly for the supplied polytopes, without constructing their faces.*

Proof. The linear program is precisely the distance from p to the convex hull of the columns of Q in the ℓ_∞ norm. Distance to a nonempty convex set is convex: convex combinations of two candidate nearest points remain feasible, and the norm satisfies its convexity inequality. Hence for $p = \sum_i \alpha_i p_i$ in K, $d(p, L) \le \sum_i \alpha_i d(p_i, L) \le \max_i d(p_i, L)$. Since every p_i belongs to K, this maximum equals the directed Hausdorff distance. Repeat in the reverse direction. □

For empirical or ensemble response polytopes, the LP computes their exact geometric distance. A coverage bound relating each supplied polytope to its true response set adds directly to this distance through the Hausdorff triangle inequality.

Relation to exact resolution and to learned models. For a finite support game, use histories containing the current consistency set and allow all point masses on its feasible successors. Worst expected continuation then equals the original worst-successor operator. A finite-budget version uses at most $\lfloor B/c_{\min} \rfloor$ probes and a terminal noncommitment penalty larger than B, with common-token commitment absorbing at zero cost. In this worst-support construction the truncated value is at most B exactly when the original game admits commitment within B. For a single stochastic kernel on a continuous history space, T7 bounds expected decision cost. Support-sound admissibility establishes the zero-violation contract; T8 supplies calibrated trajectory-risk contracts from observed outcomes.

The approximation budget consists of response error δ, test-space closure error ρ, terminal readout error ϵ_H, and planner optimization error η. Observed response fitting addresses the first term: for a fixed bounded test vector and fixed prediction m, conditional Jensen gives

$$\|m - \mathbb{E}[Y \mid h, a]\|_\infty^2 \le \mathbb{E}[\|m - Y\|_\infty^2 \mid h, a], \qquad Y = \psi_{t+1}(\phi(h')).$$

For a test map fixed before the future observation, this inequality relates response fitting to conditional prediction error. Deployment coverage transfers that error to the required histories. DeepMDP (Gelada et al., 2019) matches latent transitions; response fitting targets the continuation test space, with ρ quantifying its approximation of decision values.

A.8 Continuation error under the deployment distribution

A finite-horizon performance-difference argument (Kakade & Langford, 2002) weights continuation accuracy by the histories visited by the compared policies under a specified deployment law. Consider horizon H, actual history kernels P_t, a common initial law D_0, bounded stage costs $c_t(h, a)$, and bounded terminal cost $c_H(h)$. All functions below are measurable. Let $z = \phi(h)$ and suppose that the nonempty actual and modeled feasible action sets agree, $\mathcal{A}_t(h) = \widehat{\mathcal{A}}_t(\phi(h))$, with measurable minimizing selectors. Policies are feasible and nonanticipating. Any algorithmic randomization is included in the history with the same initial law for the compared policies.

Fix bounded modeled scores $\widehat{Q}_t$ and set $\widehat{V}_t(z) = \min_{a \in \widehat{\mathcal{A}}_t(z)} \widehat{Q}_t(z, a)$ for $t < H$, with a bounded modeled terminal value $\widehat{V}_H$. Write $q_t(h, a) = \widehat{Q}_t(\phi(h), a)$, $f_t(h) = \widehat{V}_t(\phi(h))$, and $J(\pi) = \mathbb{E}_\pi[\sum_{t<H} c_t(h_t, a_t) + c_H(h_H)]$. Let d_t^π be the true law of (h_t, a_t) and d_H^π its terminal-history law. Define

$$b_t(h, a) = c_t(h, a) + \mathbb{E}_{P_t}[f_{t+1}(h') \mid h, a] - q_t(h, a), \qquad g_t(h, a) = q_t(h, a) - f_t(h) \ge 0,$$
$$b_H(h) = c_H(h) - f_H(h), \qquad \mathcal{G}(\pi) = \sum_{t<H} \mathbb{E}_{d_t^\pi} g_t,$$
$$\mathcal{E}(\pi) = \sum_{t<H} \mathbb{E}_{d_t^\pi} |b_t| + \mathbb{E}_{d_H^\pi} |b_H|.$$

Lemma A.8 (Occupancy performance identity)**.** *Every feasible policy satisfies*

$$J(\pi) - \mathbb{E}_{D_0} f_0 = \sum_{t<H} \mathbb{E}_{d_t^\pi} b_t + \mathbb{E}_{d_H^\pi} b_H + \mathcal{G}(\pi).$$

Proof. Since $b_t + g_t = c_t + P_t f_{t+1} - f_t$, conditional expectation gives $\mathbb{E}_{d_t^\pi}(b_t + g_t) = \mathbb{E}_\pi c_t + \mathbb{E}_\pi f_{t+1} - \mathbb{E}_\pi f_t$. Summation telescopes; adding $\mathbb{E}(c_H - f_H)$ proves the identity. Boundedness ensures integrability. The representation need not be Markov. □

Proposition A.11 (Deployment-weighted policy comparison)**.** *For an evaluated policy $\widehat{\pi}$ and any feasible comparator σ,*

$$J(\widehat{\pi}) - J(\sigma) \le \mathcal{E}(\widehat{\pi}) + \mathcal{E}(\sigma) + \mathcal{G}(\widehat{\pi}) - \mathcal{G}(\sigma). \tag{A.9}$$

Exact modeled greediness gives $\mathcal{G}(\widehat{\pi}) = 0$. If its selected actions satisfy $g_t(h, a) \le \eta_t(h)$ almost surely, then $\mathcal{G}(\widehat{\pi}) \le \sum_{t<H} \mathbb{E}_{d_t^{\widehat{\pi}}} \eta_t(h_t)$.

Proof. Subtract the two identities in Lemma A.8. The common initial value cancels. Bounding the signed discrepancies by their absolute expectations gives Eq. (A.9). The nonnegative comparator gap may also be omitted for an upper bound. The remaining statements follow from the definition of g_t. □

Choosing an optimal feasible comparator, when one exists, gives expected regret. The error budget concerns both policies' trajectories; it permits arbitrary errors outside those occupancies.

Response decomposition. Let $\widehat{P}_t$ and $\widehat{c}_t$ be the modeled history-response kernel and cost. Define the modeled Bellman discrepancy

$$\xi_t(h,a) = \left| q_t(h,a) - \widehat{c}_t(\phi(h),a) - \mathbb{E}_{\widehat{P}_t} \widehat{V}_{t+1}(z') \right|.$$

For a fixed bounded test vector ψ_{t+1}, suppose $\widehat{V}_{t+1}(z) = \beta_{t+1} + v_{t+1}^\top \psi_{t+1}(z) + r_{t+1}(z)$, where $\|v_{t+1}\|_1 \le L_{t+1}$. This is the same minimized-score value defining f_{t+1} above. Write

$$\delta_t(h,a) = \left\| \mathbb{E}_{P_t} \psi_{t+1}(\phi(h')) - \mathbb{E}_{\widehat{P}_t} \psi_{t+1}(z') \right\|_\infty,$$
$$R_t(h,a) = \mathbb{E}_{P_t} |r_{t+1}(\phi(h'))| + \mathbb{E}_{\widehat{P}_t} |r_{t+1}(z')|.$$

All kernel expectations in these expressions are conditional on the current history or represented state and the selected action.

Proposition A.12 (Local continuation-response error)**.** *On every feasible history–action pair,*

$$|b_t(h,a)| \le |c_t(h,a) - \widehat{c}_t(\phi(h),a)| + \xi_t(h,a) + L_{t+1}\delta_t(h,a) + R_t(h,a).$$

Integrating this inequality under the two policy occupancies supplies the stagewise terms of Eq. (A.9)*.*

Proof. Insert the modeled cost and expected continuation into b_t. The triangle inequality separates the cost error, ξ_t, and the difference of continuation expectations. In the last difference, β_{t+1} cancels. Hölder's inequality bounds the linear term by $L_{t+1}\delta_t$; the two residual expectations contribute at most R_t. □

Exact modeled Bellman backups give $\xi_t = 0$, and $\|r_{t+1}\|_\infty \le \rho_{t+1}$ gives $R_t \le 2\rho_{t+1}$, recovering T7's local error expression for single kernels. The occupancy form instead allows expected closure error to be measured on actual and modeled successor distributions. For fixed scores and an observed cost C_t with conditional mean c_t, the real-transition innovation $C_t + f_{t+1}(h_{t+1}) - q_t(h_t, a_t)$ has conditional mean b_t; conditional Jensen bounds $|b_t|$ by its expected absolute value. Independent rollouts supply samples of these dominating losses under each compared policy. Estimation from an offline law μ_t additionally requires coverage of those occupancies: a bound $\mathrm{d}d_t^\pi / \mathrm{d}\mu_t \le C_{\pi,t}$ implies $\mathbb{E}_{d_t^\pi}|b_t| \le C_{\pi,t}\mathbb{E}_{\mu_t}|b_t|$. Including the remaining probe budget in h makes these quantities specific to the available continuation problem. The conclusion concerns expected decision cost under the specified deployment law.

A.9 Calibrated decision frontiers on stochastic history spaces

The finite support game describes exact resolution. Here we preserve the attainable cost and risk of complete continuations under an unknown deployment law. Histories, observations, actions and latent variables may be continuous. Their joint law may be dependent and non-Markovian. Controllers act on observation histories, with arbitrary measurable representations of those histories.

Continuation policies and contracts. Fix a finite horizon and a common initial-history distribution. Let Π_B be a measurable class of history-dependent continuation policies, each enforcing the probe budget B on every trajectory. A policy includes its observation requests, stopping rule and terminal decision. For a trajectory τ, let $L_0(\tau) \in [0,1]$ be normalized decision cost and $L_j(\tau) \in [0,1]$, $1 \le j \le m$, be contract losses. For example, an indicator of any probe violation or unacceptable commitment gives the probability of that event, including all stages of the continuation. Contract losses enter calibration as observations or measurements with a supplied error bound. These inputs determine the response certificate.

Write $r_j(\pi) = \mathbb{E}_\pi L_j(\tau)$ under the true law. An episode mixture q selects one complete policy from Π_B before the trajectory starts, and follows it throughout. Thus $r_j(q) = \int r_j(\pi)\, q(\mathrm{d}\pi)$, and the probe budget remains pathwise valid. Let $\mathcal{Q}_B$ be the set of these mixtures and

$$J_B^*(b) = \inf_{q \in \mathcal{Q}_B} \{ r_0(q) : r_j(q) \le b_j,\ 1 \le j \le m \}.$$

This is an expected-cost frontier with explicit trajectory-risk contracts. The probability $1-\alpha$ below concerns the calibration sample; each b_j bounds a probability or expected loss in deployment.

A response confidence event. Let $\widehat{r}_j(\pi)$ be a data-based response and $s_j(\pi) \geq 0$ its radius. Suppose, simultaneously over the controller class and all channels, that

$$\mathcal{E}_\alpha = \{\, |\widehat{r}_j(\pi) - r_j(\pi)| \leq s_j(\pi) \text{ for every } \pi \in \Pi_B,\ 0 \leq j \leq m \,\} \quad \text{has probability at least } 1-\alpha. \tag{A.10}$$

Define $u_j = \widehat{r}_j + s_j$, $l_j = \widehat{r}_j - s_j$ and extend them linearly to mixtures. A feasible η-optimal solution of

$$\min_{q \in \mathcal{Q}_B} u_0(q) \quad \text{subject to} \quad u_j(q) \leq b_j \quad (1 \leq j \leq m) \tag{A.11}$$

is denoted by $\widehat{q}$. With K supplied controllers, this is a linear program on Δ_K. The class is fixed before calibration; the mixture is optimized afterward. The confidence event therefore covers the selected mixture as well as every individual controller.

Proposition A.13 (T8: stochastic feasibility, performance and a dual certificate)**.** *On $\mathcal{E}_\alpha$, every feasible $\widehat{q}$ in Eq.* (A.11) *satisfies $r_j(\widehat{q}) \leq b_j$ for every contract. For any $\lambda \geq 0$, define*

$$d(\lambda) = \inf_{\pi \in \Pi_B} \left[l_0(\pi) + \sum_{j=1}^{m} \lambda_j l_j(\pi) \right] - \lambda^\top b. \tag{A.12}$$

Then

$$0 \leq r_0(\widehat{q}) - J_B^*(b) \leq u_0(\widehat{q}) - d(\lambda). \tag{A.13}$$

Moreover, suppose $s_j(\pi) \leq e_j$ and there exists $q^{\mathrm{s}} \in \mathcal{Q}_B$ with $r_j(q^{\mathrm{s}}) \leq b_j - \gamma_j$, where $\gamma_j > 0$ and $2e_j \leq \gamma_j$ for all $j \geq 1$. An η-optimal solution exists whenever the indicated optimization has a minimizer, and obeys

$$r_0(\widehat{q}) - J_B^*(b) \leq 2e_0 + \max_{1 \leq j \leq m} \frac{2e_j}{\gamma_j} + \eta. \tag{A.14}$$

For problems without attained minima, the same conclusion holds for feasible approximate minimizers with their optimization error included in η.

Proof. On the confidence event, integration gives $l_j(q) \leq r_j(q) \leq u_j(q)$ and $u_j(q) \leq r_j(q) + 2e_j$ for every mixture. The upper constraints therefore imply true feasibility. Every truly feasible q is feasible for the optimistic problem with costs l_0 and constraints $l_j \leq b_j$. For $\lambda \geq 0$, its Lagrangian at such a q is at most $l_0(q)$, which is at most $r_0(q)$. Minimizing that Lagrangian over mixtures is the infimum over individual policies in Eq. (A.12). Hence $d(\lambda) \leq J_B^*(b)$. Combining this with $r_0(\widehat{q}) \leq u_0(\widehat{q})$ proves the gap certificate. For a finite library, maximizing d is the ordinary dual of the optimistic LP, so both sides of the certificate are computable (Boyd & Vandenberghe, 2004).

For the rate, take any feasible comparator q^* and put $\beta = \max_j 2e_j/\gamma_j \leq 1$ and $q_\beta = (1-\beta)q^* + \beta q^{\mathrm{s}}$. Linearity gives $r_j(q_\beta) \leq b_j - \beta\gamma_j \leq b_j - 2e_j$. Thus q_β is feasible for the upper program, and

$$r_0(\widehat{q}) \leq u_0(\widehat{q}) \leq u_0(q_\beta) + \eta \leq r_0(q^*) + \beta + 2e_0 + \eta,$$

because all true decision costs lie in $[0, 1]$. Take feasible comparators approaching the infimum. This proof also establishes nonemptiness of the upper feasible set. Feasibility and the computable gap certificate hold on $\mathcal{E}_\alpha$; the strict-feasibility margin yields the explicit rate. □

Compression without a product support. The relevant geometric object is the attainable response set

$$\mathcal{C}_B = \overline{\mathrm{conv}}\{(r_0(\pi), \ldots, r_m(\pi)) : \pi \in \Pi_B\}.$$

Let $\mathcal{C}_B^\phi$ be the corresponding set when controllers use a specified learned history representation. For closed attainable response sets, as with finite controller libraries, equality preserves every cost–risk frontier regardless of the latent factorization. Equality of all their linear support functions is equivalent to this set equality by Lemma A.7. Closed convex hulls describe the closure of attainable mixtures; finite libraries attain every point. For other classes, the frontier of the closure equals the attainable infimum at thresholds with a strictly feasible mixture: mix an approximating response with that mixture by a weight tending to zero to correct any constraint error.

Corollary A.6 (Quantitative compression of the stochastic frontier)**.** *Suppose every attainable full-controller response is approximated by an attainable represented mixture to coordinatewise error at most $a_j \geq 0$. Suppose the represented class contains a mixture with risk slack $\gamma_j > 0$. Then, whenever calibration satisfies $s_j \leq e_j$ and $2e_j \leq \gamma_j$, a represented upper-program solution has, on $\mathcal{E}_\alpha$,*

$$r_0(\widehat{q}) - J^*_{B,\mathrm{full}}(b) \leq a_0 + 2e_0 + \max_{1\leq j\leq m} \frac{a_j + 2e_j}{a_j + \gamma_j} + \eta.$$

Proof. Approximate a full feasible comparator by a represented mixture q'. Its cost is at most the comparator's cost plus a_0, and its risks are at most $b_j + a_j$. Mix it with the represented slack policy with weight $\beta = \max_j(a_j + 2e_j)/(a_j + \gamma_j) \leq 1$. The mixed risks are at most $b_j + (1-\beta)a_j - \beta\gamma_j \leq b_j - 2e_j$. Its cost increases by at most β, since both mixtures have costs in $[0, 1]$. Apply the upper-program argument from T8 and take infima. □

The quantities a_j describe controller expressivity; e_j describe statistical response uncertainty. They permit continuous, correlated representations and nonzero approximation error. The computable certificate in Eq. (A.13) compares against the supplied controller class, while this corollary states the additional condition for comparison with a larger class.

Coupled environment uncertainty. The slack-mixture proof also applies to convex response functionals $r_j(q) = \sup_{P\in\mathcal{U}} \mathbb{E}_{P,q} L_j$, for an arbitrary fixed set $\mathcal{U}$ of complete environment laws, provided the estimated upper functionals are convex and their uniform errors are bounded by e_j. Convexity gives $r_j((1-\beta)q+\beta q^{\mathrm{s}}) \leq (1-\beta)r_j(q)+\beta r_j(q^{\mathrm{s}})$, which supplies the mixture step of the slack proof. Each member of $\mathcal{U}$ specifies one complete environment law throughout the continuation. Calibration below provides coverage for the actual deployment law; coverage uniformly over a supplied $\mathcal{U}$ is the corresponding condition for this robust variant.

A.10 Learning the response certificate from noisy trajectories

Condition on a training or pilot sample that fixes the neural models, controller class and evaluation scores. Calibration uses independent episodes. Within each episode the history can be arbitrarily dependent. Two constructions supply Eq. (A.10).

Observed rollout responses. For a fixed controller, each directly observed $L_j(\tau)$ is an unbiased $[0, 1]$-valued sample of $r_j(\pi)$. This applies to deterministic and stochastic controllers alike. It estimates the actual contract outcomes, without fitting a transition support.

Response correction from a shared logged source. For off-policy reuse, let μ be the logged behavior policy, with the same initial law and conditional environment law as deployment. Assume action-density coverage for the candidate policies. The prefix likelihood ratio is

$$w^\pi_t = \prod_{s=0}^{t} \frac{\pi_s(a_s \mid h_s)}{\mu_s(a_s \mid h_s)}.$$

The ratio is computed from logged and candidate action probabilities. Write a contract loss as $L_j = \sum_{t=0}^{H-1} \ell_{j,t}$, with nonnegative terms and total at most one. A trajectory-level event indicator can be assigned to the last stage. Let $Q^\pi_{j,t}(h, a)$ be any fixed bounded neural score, including a misspecified one, and put $V^\pi_{j,t}(h) = \mathbb{E}_{a\sim\pi_t} Q^\pi_{j,t}(h, a)$, $V^\pi_{j,H} = 0$. Define the observed score

$$Z^\pi_j = V^\pi_{j,0}(h_0) + \sum_{t=0}^{H-1} w^\pi_t [\ell_{j,t} + V^\pi_{j,t+1}(h_{t+1}) - Q^\pi_{j,t}(h_t, a_t)]. \tag{A.15}$$

This is a history-space form of sequential doubly robust evaluation (Jiang & Li, 2016). Its role here is to supply confidence bounds for the entire cost–risk frontier.

Lemma A.9 (Unbiased response correction)**.** *With the preceding coverage and integrability conditions, $\mathbb{E}_\mu Z^\pi_j = r_j(\pi)$ for every fixed neural score.*

Proof. For any integrable function of (h_t, a_t, h_{t+1}), multiplication by w_t^π changes its expectation under μ to that under π. Also $\mathbb{E}_\pi Q_{j,t}^\pi(h_t, a_t) = \mathbb{E}_\pi V_{j,t}^\pi(h_t)$. Taking expectations in Eq. (A.15) therefore telescopes the value terms, leaving $\mathbb{E}_\pi \sum_t \ell_{j,t}$. The score enters only through these telescoping terms, so misspecification does not bias Z_j^π. □

If scores are clipped to $[0, 1]$ and $0 \le w_t^\pi \le W_t$, then Z_j^π lies in an interval of width at most $D = 1 + 3\sum_t W_t$. Tighter valid range bounds can be used. Exact coverage is only needed for logged-source reuse; direct rollout responses above have range width one. Estimated likelihood ratios or noisy contract labels with a justified mean-bias bound a_j^{obs} add that bound to the radius. The bias bound is separate from sampling uncertainty.

Proposition A.14 (Finite-sample and variance-adaptive calibration)**.** *For K fixed controllers, let $n \ge 2$ i.i.d. calibration scores per controller and channel have mean $r_j(\pi)$ and known range width at most D. Shared episodes across controllers and channels are allowed. Their empirical means satisfy Eq.* (A.10) *with either the Hoeffding radius*

$$e_n = D\sqrt{\frac{\log(2K(m+1)/\alpha)}{2n}}, \tag{A.16}$$

or the policy-specific empirical Bernstein radius

$$s_j(\pi) = \sqrt{\frac{2\widehat{v}_j(\pi)\,\log(4K(m+1)/\alpha)}{n}} + \frac{7D\log(4K(m+1)/\alpha)}{3(n-1)}, \tag{A.17}$$

where $\widehat{v}_j(\pi)$ is the unbiased sample variance.

Proof. Apply the bounded-variable concentration inequality to each of the $K(m+1)$ score means and both signs. Hoeffding's inequality gives Eq. (A.16). For Eq. (A.17), rescale each known range to $[0, 1]$ and apply Theorem 4 of Maurer & Pontil (2009) to both tails with failure probability $\alpha/[2K(m+1)]$. A union bound proves simultaneous coverage. Episodes are independent for each score; dependence across controllers, channels and within-episode time steps is unrestricted. Either radius supplies the stated confidence event. □

Corollary A.7 (A noisy-data requirement for a decision frontier)**.** *If the true frontier has a slack policy with $\gamma_j \ge \gamma > 0$, then for $0 < \Delta \le 1$, the sufficient count*

$$n \ge \frac{2D^2(1+\gamma^{-1})^2}{\Delta^2}\log\frac{2K(m+1)}{\alpha}$$

gives true feasibility and excess decision cost at most $\Delta + \eta$ with probability at least $1 - \alpha$. This count is per controller for direct rollouts; with shared logged episodes it is the common episode count, with the corresponding off-policy range D.

Proof. The displayed count ensures $e_n \le \Delta/[2(1+\gamma^{-1})]$. Since $\Delta \le 1$, it also ensures $2e_n \le \gamma$. Substitute $e_j = e_n$ into Eq. (A.14). □

This replaces exact obstruction survival by statistical resolution of a cost–risk frontier. Its complexity depends on the controller response class, coverage and contract slack, rather than the number of latent worlds. The variance-adaptive radius can give a tighter certificate for predictable continuation responses.

Infinite controller families. Finiteness of the controller library is also optional. On a common logged-trajectory space, suppose each bounded score class $\{Z_j^\pi : \pi \in \Pi_B\}$ admits a fixed uniform ξ-net of cardinality at most $N(\xi)$, with range width at most D for its members. Then, with probability at least $1 - \alpha$,

$$\sup_{\pi,j} |\overline{Z}_j^\pi - r_j(\pi)| \le 2\xi + D\sqrt{\frac{\log(2(m+1)N(\xi)/\alpha)}{2n}}.$$

Indeed, empirical and population means each change by at most ξ on moving to a net element; apply the preceding union bound on the net. This supplies T8 for continuous controller families with controlled response complexity and measurable optimization. The covering condition applies to response scores; the physical state space and world class may remain continuous.

A.11 Convex acquisition with stochastic response uncertainty

The confidence contribution to a decision certificate can itself be optimized. Freeze a pilot mixture q and nonnegative dual vector λ, then allocate fresh calibration episodes $n_k > 0$ among the K controllers. With known channel ranges and a fixed union-bound level, write their Hoeffding radii as $s_{jk}(n_k) = a_{jk}/\sqrt{n_k}$, where $a_{jk} \geq 0$ is fixed. The difference between the interval-based gap $u_0(q) - d(\lambda)$ and its zero-radius expression at the same empirical centers is at most

$$F(n) = \sum_k q_k \frac{a_{0k}}{\sqrt{n_k}} + \max_k \frac{a_{0k} + \sum_j \lambda_j a_{jk}}{\sqrt{n_k}}.$$

Consequently

$$\min_{n_k \geq n_{\min}} F(n) \quad \text{subject to} \quad \sum_k c_k n_k \leq N \tag{A.18}$$

is a convex allocation of the statistical part of the certificate, where $c_k > 0$ is the collection cost and N the calibration budget. For the bound on F, use $\min_k(x_k - y_k) \geq \min_k x_k - \max_k y_k$. Convexity follows from that of $n \mapsto n^{-1/2}$, nonnegative weights, and pointwise maxima. Integer allocations use the resulting actual counts in their confidence radii. Pilot information is held fixed before these episodes, preserving calibration coverage. The program minimizes confidence width; the empirical centers determine the final cost and feasibility gap. As in T6, dual prices identify the distinctions on which additional evidence matters; here these distinctions are noisy cost and risk responses of complete continuations.

Dependence on the online budget. If $B' \geq B$ permits every continuation in Π_B and possibly more, then $\mathcal{C}_B \subseteq \mathcal{C}_{B'}$ and $J^*_{B'}(b) \leq J^*_B(b)$. This follows by retaining every old mixture. Confidence intervals on retained controllers can also be reused when the joint class and confidence allocation were specified in advance. The statistical rate tracks the complexity and contract slack of the resulting response class.

A.12 A continuous noisy continuation with an explicit certificate

Let the hidden target be $\theta \sim \mathcal{N}(0, 1)$. One probe observes $Y = \theta + \zeta$, where $\zeta \sim \mathcal{N}(0, 0.8^2)$ is independent. A probed controller uses action $a = kY$; without a probe it uses $a = 0$. Histories and observations are continuous, and the observation couples the hidden target with noise. Define the contract violation as $|\theta - a| > 0.9$. Decision cost is $(1 - \cos(\theta - a))/2$ without a probe, and $0.02 + 0.98(1 - \cos(\theta - a))/2$ with a probe. Abstention has cost one, risk zero, and no interaction. For a gain k,

$$v(k) = (1-k)^2 + 0.64k^2, \qquad r_0(k) = 0.02 + 0.49(1 - e^{-v(k)/2}), \qquad r_1(k) = 2\Phi(-0.9/\sqrt{v(k)}),$$

where Φ is the standard normal cumulative distribution function. These expressions supply an analytic comparison independent of the calibration estimates.

With risk cap $b = 0.2$, the zero-probe frontier over the fixed-action and abstention policies has cost 0.563585. Allowing one probe and gains $\{0.2, 0.4, 0.6, 0.8, 1.0\}$ lowers the exact library frontier to 0.106892. The controller class and confidence level are fixed before sampling. We collect $4,096$ independent episodes separately for each of the six interacting policies, for $24,576$ episodes in total; abstention uses its known contract. At joint confidence 95%, the empirical Bernstein program selects $k = 0.6$. Its analytic risk is $0.149750 < 0.2$, and its cost is 0.106892. The computed upper-cost minus optimistic-dual certificate is 0.021266. All calibrated responses in this fixed run contain their analytic means. Thus the upper program selects an optimal member of this library while reporting a data-based upper bound on its possible suboptimality.

The same calculation also checks the rate and dual inequalities on 100 perturbed response LPs. A separate two-stage action-sum calculation uses fixed nonlinear neural scores unrelated to the true response, continuous Gaussian inputs, and known logging probabilities. The response-correction identity agrees to floating-point precision. The recorded output contains the analytic responses, episode counts and LP certificates for this continuous stochastic interface.

A.13 CONVEX FACTOR INFERENCE AND ITS DETERMINISTIC ERROR BOUND

The inference layer combines a learned role transition with a source-calibrated readout of the recurrent world-model state. For Reacher, the four coordinates are the endpoint and target image positions, divided by 63 for a 64×64 image. Let $C = [0, 1]^4$. The role transition predicts the endpoint from its two previous estimates and the previous action; the target follows the within-episode identity transition. The recurrent state, actor, critic, reward and continuation models retain their trained weights.

A strictly convex source calibration. Let $\widetilde{y}_{ig} \in \mathbb{R}^2$ be the decoded recurrent-state readout for role g, and let y_{ig}^{RGB} be its label from the frozen, source-fitted RGB extractor. Fit the endpoint and target blocks separately:

$$W_g^* = \arg\min_W \sum_{i \in I} \|[\widetilde{y}_{ig}^\top, 1]W - (y_{ig}^{\mathrm{RGB}})^\top\|_2^2 + \|W - W_{g,0}\|_F^2, \qquad W_{g,0} = \begin{bmatrix} I_2 \\ 0\ 0 \end{bmatrix}.$$

Thus $W_g^* = (X_g^\top X_g + I_3)^{-1}(X_g^\top Y_g + W_{g,0})$. The ridge term makes each problem strictly convex with a unique solution. Calibration uses the original 20-episode preload: 2,580 already acquired RGB frames, of which 1,244 yield valid readout labels and 889 yield complete transition-label triples. The mask determines the regression sample; evaluation executes every episode. The calibration fits twelve affine coefficients using the existing source, with all neural weights fixed.

Let R contain the four mean squared source-calibration residuals. The endpoint entries of Q contain source transition residuals; the target entries are zero because the role model fixes the target. Positive entries have a floor $(0.5/63)^2$. These empirical scales specify the quadratic metric. Initialize $P_0 = R$ and, at each step, form $P_t^- = P_{t-1} + Q$, the box-projected role prediction p_t and the box-projected affine readout m_t. With diagonal matrices,

$$\widehat{z}_t = \arg\min_{z \in C} \tfrac{1}{2}\|z - p_t\|^2_{(P_t^-)^{-1}} + \tfrac{1}{2}\|z - m_t\|^2_{R^{-1}}, \tag{A.19}$$

$$K_t = P_t^-(P_t^- + R)^{-1}, \qquad \widehat{z}_t = \Pi_C\big(p_t + K_t(m_t - p_t)\big), \qquad P_t = (I - K_t)P_t^-. \tag{A.20}$$

The box projection is coordinatewise clipping. At episode reset, the estimate is initialized from the current readout and the scale is reset to R. The actor receives its original recurrent features concatenated with the renormalized estimate. Inference therefore uses the learned role transition and the current image without refitting the world model.

For $P, R \succ 0$, the general fusion objective is $\min_{z \in C} \tfrac{1}{2}\|z - p\|^2_{P^{-1}} + \tfrac{1}{2}\|z - m\|^2_{R^{-1}}$, with $\|v\|_M^2 = v^\top M v$.

Proposition 1 (Convex factor inference)**.** *For nonempty closed convex C and $P, R \succ 0$, the minimizer is unique. Put $G = P^{-1} + R^{-1}$ and $q = G^{-1}(P^{-1}p + R^{-1}m)$. For every $z \in C$,*

$$\|\widehat{z} - z\|_G^2 \le \|q - z\|_G^2 - \|q - \widehat{z}\|_G^2.$$

Proof of Proposition 1. For arbitrary positive-definite P, R, put $G = P^{-1} + R^{-1}$ and $q = G^{-1}(P^{-1}p + R^{-1}m)$. Completing the square gives

$$\|z - p\|^2_{P^{-1}} + \|z - m\|^2_{R^{-1}} = \|z - q\|_G^2 + \|p - m\|^2_{(P+R)^{-1}}.$$

The objective is strictly convex and coercive. On any nonempty closed convex C, its unique minimizer is the metric projection $\widehat{z} = \Pi_C^G(q)$. Its variational inequality is $(q - \widehat{z})^\top G(z - \widehat{z}) \le 0$ for every $z \in C$. Expanding $\|q - z\|_G^2$ yields

$$\|\widehat{z} - z\|_G^2 \le \|q - z\|_G^2 - \|q - \widehat{z}\|_G^2.$$

Consequently,

$$\|\widehat{z} - z\|_G^2 \le \|p - z\|^2_{P^{-1}} + \|m - z\|^2_{R^{-1}} - \|p - m\|^2_{(P+R)^{-1}} - \|q - \widehat{z}\|_G^2.$$

For the implemented box and diagonal metrics, projecting each raw input before fusion also decreases its corresponding squared error to any $z \in C$. These deterministic projection inequalities hold for arbitrary error distributions and dependence. If a continuation readout is L-Lipschitz in this metric, a factor error bound ζ gives a readout error bound $L\zeta$, which can enter T7's continuation-response budget. □

A.14 CONVEX READOUTS FROM POINT AND SET ANNOTATIONS

Let $\phi_i \in \mathbb{R}^p$ include a constant coordinate, let $W \in \mathbb{R}^{q\times p}$, and let $C_i \subseteq \mathbb{R}^q$ be nonempty closed convex image-space label sets. A visible position provides a singleton label. An occluded endpoint associated with a contact observation provides the target-support ball from the source image. Available coordinates contribute independent annotation records, so a visible target remains supervised when the endpoint is occluded.

$$F(W) = \sum_i w_i\, d^2(W\phi_i, C_i) + \lambda\|W - W_0\|_F^2, \qquad w_i \geq 0, \quad \lambda > 0. \tag{A.21}$$

Proposition 2 (Set-supervised readout)**.** *The objective in Eq.* (A.21) *is* 2λ*-strongly convex with a unique minimizer. Its gradient is Lipschitz with constant at most* $2\lambda + 2\sum_i w_i\|\phi_i\|_2^2$.

Proof of Proposition 2. The objective has gradient

$$\nabla F(W) = 2\sum_i w_i\big(W\phi_i - \Pi_{C_i}(W\phi_i)\big)\phi_i^\top + 2\lambda(W - W_0).$$

The squared distance to a nonempty closed convex set is convex and continuously differentiable, with gradient $2(u - \Pi_C u)$. The residual map $I - \Pi_C$ is nonexpansive. Composition with the linear map $W \mapsto W\phi_i$ preserves convexity and gives gradient $2(W\phi_i - \Pi_{C_i}(W\phi_i))\phi_i^\top$. Its Lipschitz constant in Frobenius norm is at most $2\|\phi_i\|_2^2$. Summing with nonnegative weights gives the stated gradient and smoothness bound. The ridge term is 2λ-strongly convex and makes the objective coercive. A unique minimizer therefore exists. □

For a ball $C_i = \{u : \|u - g_i\|_2 \leq r_i\}$,

$$d^2(W\phi_i, C_i) = \big(\|W\phi_i - g_i\|_2 - r_i\big)_+^2.$$

Singleton labels yield ordinary squared-error regression; balls yield a differentiable convex set loss. The objective combines both with the same affine parameterization.

Source-label intervention. The readout uses 128 fixed role-hidden features with source-fitted centering and scaling, plus an intercept. The original linear layer supplies W_0. The source pool combines 64 policy episodes and 32 random-source episodes, yielding 36,128 aligned feature frames. The charged source budget is 1,090,384 physical steps. Readout fitting uses 48 policy and 24 random episodes; the remaining 16 and eight episodes supply the source-label holdout. Neural parameters are fixed.

Static target labels are formed from visible target pixels within each source episode. Endpoint points use visible endpoint annotations; contact frames with an occluded endpoint use image-support balls expanded by the half-pixel convention. The pool includes 2,731 visible endpoint points and 30,268 occluded contact-set labels. The endpoint objective adds the mean point-coordinate squared loss, mean squared distance to contact balls, and a 10^{-4} ridge penalty to the original readout. Target coordinates use their independently available point labels with the same ridge coefficient. The point problems use the normal equations; the mixed point/set problem uses L-BFGS on this convex objective. Table 3 gives the held-out errors.

Readout	Target MSE	Visible endpoint MSE	Contact-set discrepancy
Original	0.072408	0.009113	0.125804
Point/set source refinement	**0.001024**	**0.005087**	**0.007996**

Table 3: Readout prediction on the 24 source episodes held out from label fitting. Target and visible-endpoint columns average coordinate squared errors; the contact column averages squared distance to the image-support label ball.

B COMPLETE EXPERIMENTAL FIGURES

Table 4 collects the three exact-support comparisons.

Scientific question	Data comparison	Result
Less passive evidence?	Tabular: $144/14$	**10.3**×; 95.33% / 94.50% success versus 100%; retention lower bounds 0.938 / 0.930; exact interface
Discoverable structure?	Opaque tables: $144/18$	**8**×; 583/600 and 592/600 successes; all 24 opaque tables yield certified factors; all refusals have full-game explanations
External exact transfer?	Boxoban: $1024/32$	**32**× versus the strongest competing full-data planner: 81% versus 79%; no irreversible failures in 1,200 episodes

Table 4: **Decision resolution across three data settings.** These controlled and external exact-support experiments use distinct data currencies and references. The main text reports LeWorldModel separately; complete protocols, controls, and replication results are in the corresponding appendices.

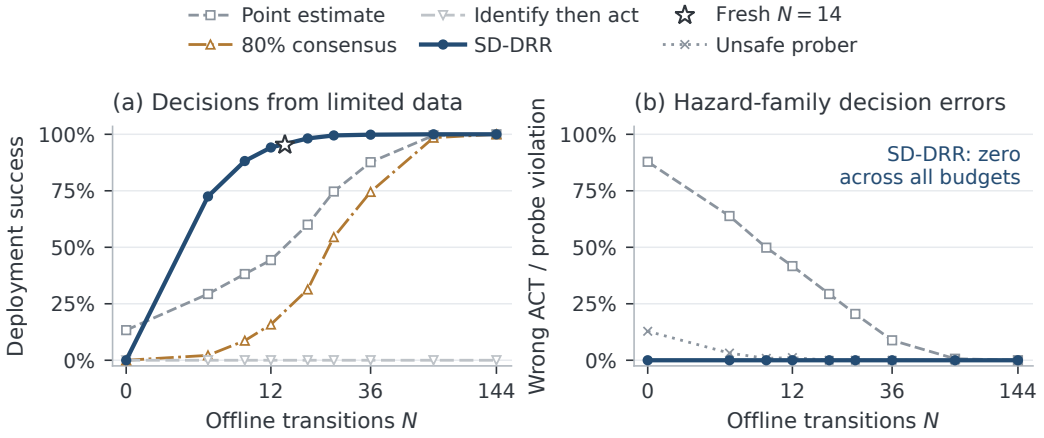

Figure 4: **Decision quality across observation budgets.** (a) Original tabular source-budget comparison; the star marks the separate fresh $N = 14$ replication (95.33% success, $144/14 \approx 10.3\times$ fewer observations). (b) Hazard-family errors: false ACTs for point estimation, probe violations for the unsafe prober, and their sum for SD-DRR. Each original cell has 600 trials. Methods share data and the online budget except the explicitly unsafe control. The N axes are linear through 6 and logarithmic thereafter.

For T1–T4, independently coded checks supplement the written proofs (Table 5). Figure 6 connects T4's interaction counts to complete-table discovery costs. The counterexample statistics quoted after Proposition A.2 come from the 300-instance DNF pool in that table.

Figure conventions. Empirical points report saved cell means with the trial counts specified in their captions. Paired success and return intervals use the stated fixed cohorts. The finite-catalogue admission bounds use exact count inversion. The F5 figure distinguishes analytic interaction counts from measured discovery and certification times.

Claim checked	Instances	Result
CE1 values: full $= 2$, profile quotient $= \infty$, fine quotient equal	frozen construction	pass
T1: saturation and state-wise value equality	500 random (240 with $k_f < \lvert\mathcal{H}\rvert$)	0 violations
T2: one-sided soundness	500 random (one strict-gap instance)	0 violations
T3: fiber-quotient equality	300 random DNF	0 violations; mean compression $3.01\times$
CE2 values: full $= 2$ / profile $= \infty$ / fiber $= 2$	frozen construction	pass
Certificate self-checks (losing-set closure, policy value replay)	every solver run	pass
Budget-recursion verifier sharing no solver code	200 random + all frozen	0 mismatches
Static degeneration vs. brute-force test tree	100 instances	0 mismatches

Table 5: Independently coded verification checks. Every theorem statement is exercised on randomized instances by checkers written separately from the reference solver.

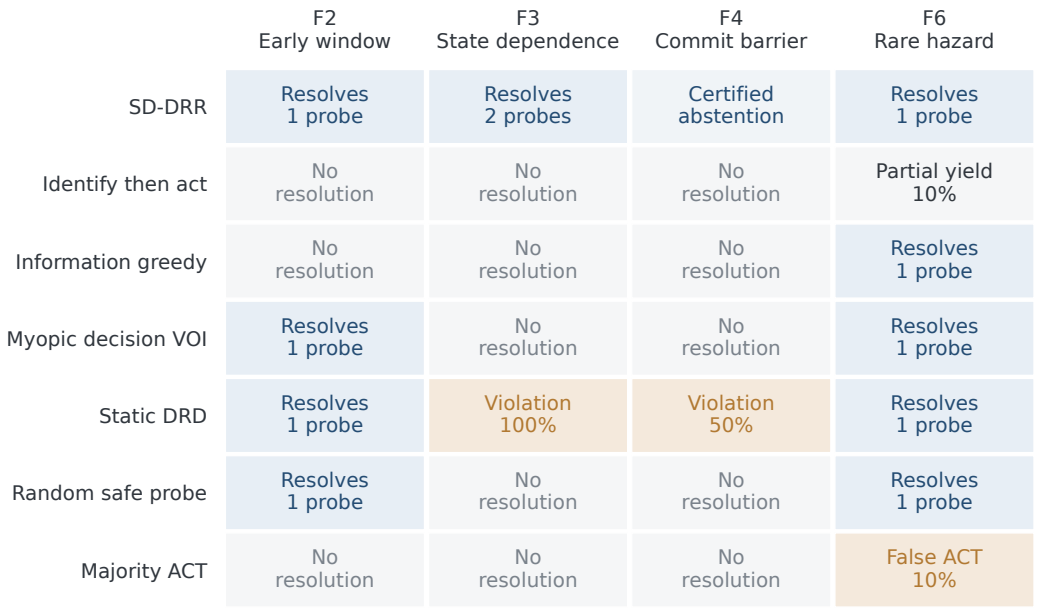

Figure 5: Outcomes on the four finite dynamic families. F2, F3, F4, and F6 enumerate 8, 4, 2, and 10 worlds respectively. Cells distinguish resolution, certified structural abstention, unresolved execution, and invalid actions. Costs are worst-case probes among resolved worlds; percentages are fractions of the enumerated worlds.

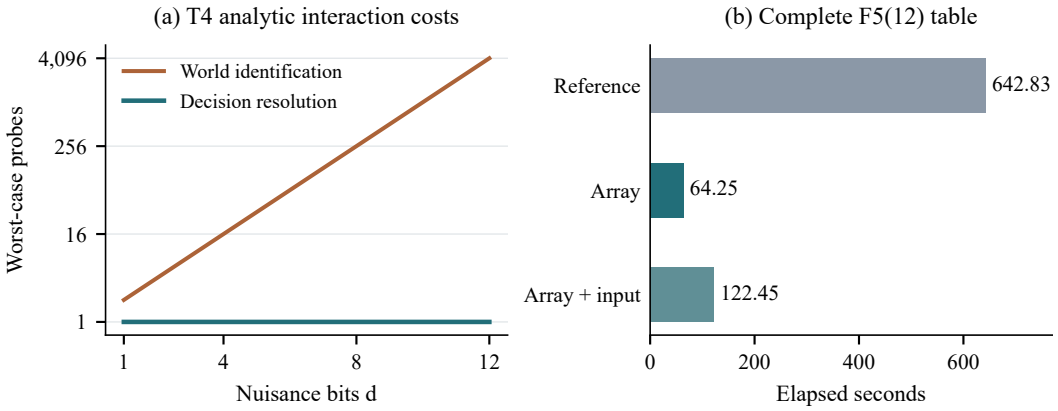

Figure 6: F5 interaction complexity and complete-table discovery. (a) The analytic worst-case probe counts from T4: one decision probe versus 2^d for world identification. (b) Completed runs on the same 33.56-million-entry $F_5(12)$ table: 642.83 s for reference discovery and certification, 64.25 s for the array implementation, and 122.45 s including input generation.

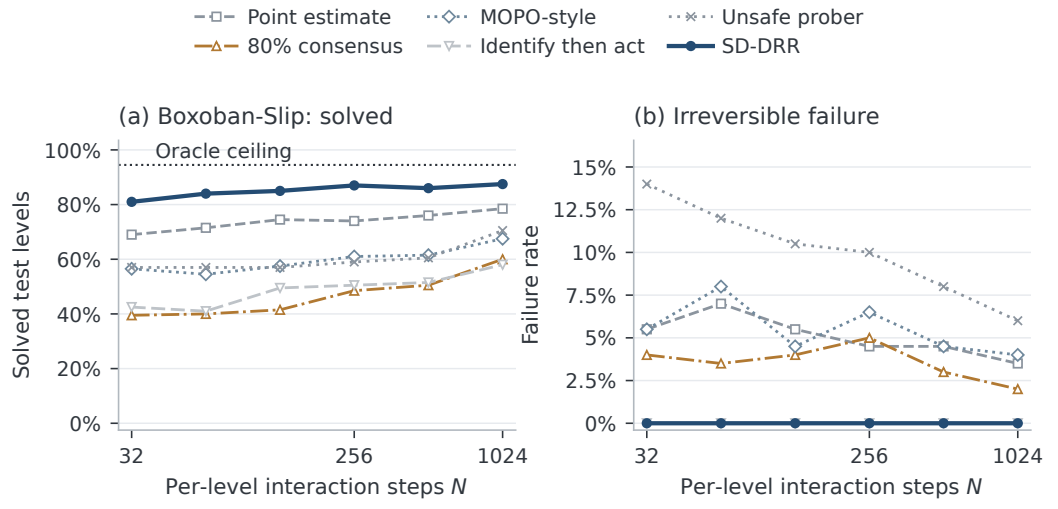

Figure 7: Boxoban-Slip at six per-level interaction budgets, with 200 fixed test levels per cell. (a) Fraction solved; the dotted line is the oracle's 189/200 ceiling. (b) Irreversible-failure rate. SD-DRR solves 81% at $N = 32$, versus 79% for the strongest unconstrained comparator at $N = 1024$, with zero irreversible failures in its 1,200 trials.

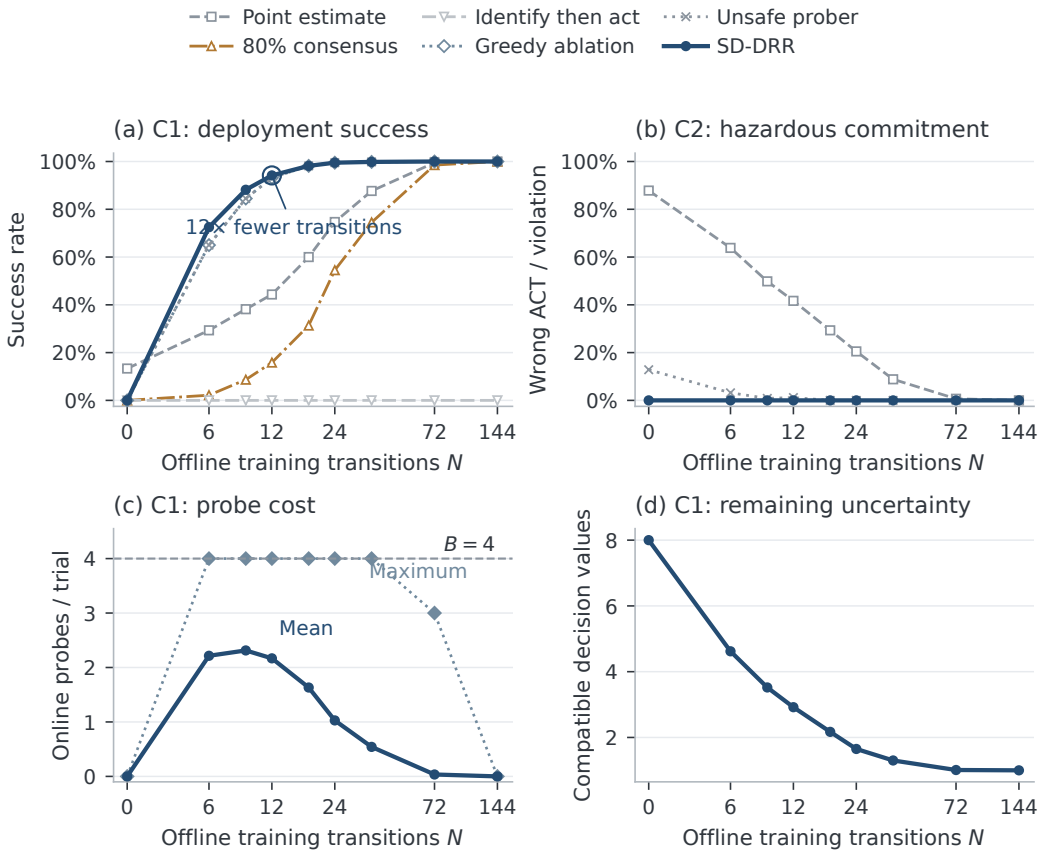

Figure 8: Tabular data reduction (600 trials per family/budget). (a) Success on the resolvable family. (b) On the hazard family: point-estimate false ACTs, unsafe-prober violations, and SD-DRR's zero error. (c) Mean and maximum probes under $B = 4$. (d) Mean surviving decision coordinates. The N axes are linear through 6, logarithmic thereafter.

C AUTODNF: DISCOVERING THE DECLARATION (PROPOSE-THEN-CERTIFY)

AutoDNF constructs candidate decision–nuisance grids from flat tables and certifies proposals under the three structural conditions of T3: fiber-wise acceptability/safety constancy, and a shared successor with a rectangular observation grid. Discovery is therefore split into a **proposal stage** and an independent exact **certifier** that checks those conditions. Certified candidates define the quotient; other proposals retain a finer partition during bounded search. The implemented certifier operates on singleton transition supports and checks the observation product pattern separately at each (x, q). The discovery protocols and measurements are specified below. Structural-signature timings and shared-symbol alignment are reported in App. C.1; the separate opaque-table confirmation appears in App. C.2.

Formal statement. An AutoDNF instance is a flat singleton-support presentation of an SD-DRR instance. Worlds and observations are opaque identifiers. A candidate grid g assigns each world to a fiber-row and nuisance-column pair, with rows of equal width. The certifier checks row-constant acceptability and safety, then checks each (x, q) for a shared successor and an observation product pattern. A symbol determines and is determined by its row and column classes.

Proposition A.15 (AutoDNF soundness: certification *is* declaration)**.** *If the certifier accepts g, the instance satisfies the per-(x, q) DNF condition with Θ_{dec} the rows and Θ_{nui} the columns of g. T3 then applies verbatim: the closure quotient over g's rows is sound and complete with the usual compression. Soundness is unconditional in the proposer:* any *heuristic may propose g.*

Proof. The accepted conditions are hypotheses (i)–(iii) of the per-(x, q) DNF condition used in T3's proof (App. A.4); they imply product invariance, y_1-survival, and the y_2-repair property. □

Candidate grids and certification cost. Candidate fiber widths divide the gcd of profile-class sizes, a necessary condition for row-uniform acceptability. The symbol-grid certifier checks singleton-support tables at cost $O(|\mathcal{H}| \cdot |X| \cdot |Q|)$ per proposed grid. Certified candidates determine the quotient used by the planner; the proposal procedure combines structural signatures, shared-symbol alignment, and bounded enumeration.

Exact tabular recovery. On the 300-instance DNF pool used for Table 5, AutoDNF recovers a certified decomposition on all 300 instances. In 294 cases its compression matches the planted decomposition; in six it discovers additional certified symmetry. Every accepted quotient preserves the full-game values in the paired solver checks. The array implementation below evaluates discovery and certification on complete input tables using the same structural conditions.

C.1 STRUCTURAL SIGNATURES AND SHARED-SYMBOL ALIGNMENT

A cardinality-signature proposal for column alignment. The frozen AutoDNF search aligns nuisance columns by bounded backtracking. We add a deterministic proposal before this search, keeping its independent exact DNF certifier unchanged. Let $F_1, \ldots, F_k$ be the coarse profile classes, and suppose each has cardinality $m \geq 2$. For a channel $r = (x, q)$ safe in every class and a world $w \in F_i$, define

$$b_r(w) = \big|\{v \in F_i : y_r(v) = y_r(w)\}\big|.$$

Starting with one color, refine $\chi_{r+1}(w) = \mathrm{intern}(\chi_r(w), b_r(w))$, using a shared interning dictionary across all classes at that channel. The resulting integer colors encode each world's sequence of observation-equality block cardinalities. They require neither planted metadata nor an interpretation of world or observation names. Matching colors propose column alignment; ambiguous colors use deterministic ordering only as a proposal. Every proposal is passed to the original exact certifier. Unequal class sizes, incompatible color multiplicities, or failed certification invoke bounded search. The timing comparison below retains the original fallback; the shared-symbol extension described next is enabled separately without changing its search budgets.

Proposition A.16 (Recovery from separating structural signatures)**.** *Suppose the equal-size profile classes are the fibers of a valid singleton-support DNF grid. If the block-cardinality signatures over their shared safe channels distinguish all m nuisance columns, the structural-signature proposal recovers a certified grid up to a common column permutation. Recovery under this condition is invariant to bijective renaming of worlds and observation symbols.*

Proof. In a valid DNF grid, equality of observations within a row is exactly equality of its nuisance observations. Thus corresponding columns have identical block cardinalities at every shared safe channel. Induction over the shared integer refinements gives identical colors for corresponding columns. Separating signatures make these colors unique in each row, so matching colors identifies the common column alignment. A common permutation preserves the rectangular grid and passes the certifier. Bijective renaming preserves observation equality and block sizes; with unique colors, world-name tie-breaking is never used. □

This is a sufficient recovery condition. Profile classes can merge distinct fibers, and repeated signatures can remain ambiguous. Such proposals still require exact certification. An equal-size profile partition has the largest admissible fiber width, so accepting it does not sacrifice compression to a narrower grid. A refusal retains the bounded search.

Cross-row constraints from shared public symbols. Block cardinalities do not resolve all alignments: balanced observation blocks can have identical signatures, while their public symbols impose additional relations across profile rows. Let g_{ij} denote the world placed in row i, column j, and let $n_r(j)$ label the nuisance observation block of column j at channel r. For rows safe at r, every certified grid satisfies

$$y_r(g_{ij}) = y_r(g_{i'j'}) \quad \Longrightarrow \quad n_r(j) = n_r(j').$$

The implementation maintains a per-channel map from each encountered public symbol to its current column-block label. A placement that gives an already encountered symbol a different block label is rejected; the map is restored with the other constraints on backtracking. This propagates equality across rows without interpreting world ids or supplying a planted factorization.

Why the pruning is sound. The certifier's singleton product condition makes an observation symbol identify its row-class and column-class pair. Equal symbols therefore imply equal column classes, which is precisely the displayed constraint. A placement violating it cannot extend to a valid grid. The constraint prunes candidate placements; the unchanged certifier establishes validity for each completed grid. Bijective renaming preserves the constraint. Recovery searches the budget-bounded prefix of the naming-dependent candidate order. The implementation enables this extension through the flag `shared_symbols=True`.

Algorithm: AutoDNF with signatures and shared-symbol propagation

1. Partition worlds by their acceptability and safety profiles.
2. For equally sized nontrivial classes, refine shared colors using within-class observation-block sizes at common safe channels.
3. If color multiplicities match, align equal colors into a grid. Use deterministic tie-breaking to propose candidate alignments.
4. Certify row uniformity, shared successors, and the rectangular symbol pattern over all safe channels. Return the grid if accepted.
5. Otherwise enumerate row partitions and column placements within the fixed search budgets. In shared-symbol mode, propagate both within-row equality partitions and the per-channel symbol-to-column-block constraint. Backtrack on a conflict.
6. Apply the same exact certifier to every completed candidate; return a certified grid or the original refusal status.

Computational cost. For $N = km$ worlds and $R = |X||Q|$ channels, the proposal after profile construction costs $O(NR + N \log m)$ time and $O(N)$ additional working storage, with constant-time table operations. The implementation's profile-construction cost is accounted for separately as T_{profile}; the original certifier is also included in the measured discovery time. On F_5, each equality probe has one singleton nuisance block, so these signatures separate all columns and remove alignment backtracking. Discovery complexity is measured in the explicit input size: the $F_5(d)$ table contains $2^{d+1}(2^d + 1)$ world–action entries, exponential in the succinct parameter d.

Correctness of the structural-signature proposal. On 300 random DNF instances, 285 use the proposal and 15 use the unchanged fallback. All retain the reference method's compression: 294 match the planted partition and six find a coarser certified partition. CE1 and CE2 return trivial

partitions, and all 60 general random controls retain the reference refusal status. On 303 tractable instances, exact original-versus-quotient game comparisons find zero value mismatches for both the reference and proposal implementations. The complete-input comparison below evaluates the final array implementation and reference on the same machine.

An exact array certificate. The scalar certifier repeatedly reconstructs observation equality partitions. The following equivalent criterion permits array operations on a supplied grid without relaxing any product condition.

Proposition A.17 (Rectangular symbols from row and column counts)**.** *Let M be a nonempty finite matrix of public observation symbols. Let $r(M)$ and $c(M)$ be the numbers of distinct complete row and column vectors, and let $s(M)$ count its distinct symbols. There are row labels a_i, column labels b_j, and a bijection ι from their occurring label pairs to the symbols such that $M_{ij} = \iota(a_i, b_j)$ if and only if*

$$s(M) = r(M)c(M).$$

Consequently, applying this test at every safe channel, together with row-uniform acceptability and safety and a shared successor, is exactly the singleton-support DNF certificate for that proposed grid.

Proof. Identify identical complete rows and identical complete columns. The symbol in a cell is constant on the resulting class pair, so there is a well-defined surjection from the $r(M)c(M)$ class pairs onto the $s(M)$ symbols. Equality of these finite cardinalities makes the map a bijection, proving sufficiency. Conversely, under a bijective product representation, two complete rows agree exactly when their row labels agree, and likewise for columns. Every occurring row label is paired with every occurring column label; injectivity therefore gives $s(M) = r(M)c(M)$. The remaining DNF conditions are checked separately and unchanged. □

Observation symbols are interned by equality and stored with successor indices and safety bits in a $W \times R$ table. Sorting-based signature refinement and the displayed certificate cost $O(WR \log W)$ time and $O(WR)$ storage, including table conversion. This bound is in the explicit input size. Under the separating-signature condition of Proposition A.16, it covers discovery and certification; ambiguous proposals retain bounded search. Neither symbol interning nor the certificate interprets the world names.

Complete-input scaling. We additionally measured the reference and array paths on identical complete dictionaries in one Linux/Python 3.10 process. Reusing equal immutable support values reduces allocation while preserving every world–action entry. No planted declaration is supplied to discovery. The reference runs first at $d = 10$ and the array path first at $d = 12$; there is one measurement per method and size. Discovery times below include table conversion, proposal, and exact certification.

d	Entries	Generation (s)	Reference (s)	Array (s)	Total (s)
10	2,099,200	2.34	23.69	3.46	5.80
12	33,562,624	58.21	642.83	64.25	122.45

Total denotes generation plus the array path. Both methods recover the same two fibers, with widths 1,024 and 4,096 respectively. At $d = 12$, the array path spends 57.34 s on conversion, 3.40 s on the proposal and 3.50 s on certification, so generation, discovery and certification of the complete input finish in 122.45 s. The explicit F_5 family satisfies the separating-signature hypothesis used by the discovery algorithm.

C.2 Opaque-table recovery and joint deployment confirmation

The joint study tests automatic recovery, reduced passive data, and safe deployment on the same instances. Shared-symbol propagation uses fixed partition, alignment, and candidate limits. After development, we froze a confirmation batch of 24 fresh opaque presentations, 12 each from the deterministic, realizable C1 and C2 families. Each table has 64 worlds, three states, and five actions. Worlds, states, actions, observations, and tokens are independently renamed; discovery receives the flat tables and produces the certified grid.

For each table, 50 fresh seeds choose a true world and an iid sequence of uniform state–action table observations. The low-data set contains the first 18 observations of the full 144-observation sequence. These passive table observations are distinct from online physical probes. The primary online budget is fixed at $B = 4$; $B = 0$ and $B = 5$ are separately specified budget boundaries. Discovery precedes these deployments and returns a certified 64-to-8-world quotient on all 24 confirmation tables (Table 6). The original certifier is unchanged.

Confirmation batch	C1	C2
Tables with certified $8\times$ compression	12/12	12/12
Success, $N = 18$, $B = 4$	583/600	592/600
Success, $N = 144$, $B = 4$	600/600	600/600
Low-data online probes	939	727
Low-data erroneous ACT / unsafe probes	0 / 0	0 / 0
Low-data budget / structural refusals	17 / 0	0 / 8

Table 6: Automatic compression and deployment on the same 24 confirmation tables. All 1,200 low-data trials contribute to task success; certified refusals remain in the task denominator.

Certificates for finite-budget decisions. For all 25 low-data refusals, solving the original fine game at the saved consistency set gives the same value as the discovered quotient: $C^* = 5$ for all 17 C1 cases and $C^* = \infty$ for all eight C2 cases. The shortest safe path to any ACT state has length 5 in those C1 cases; the corresponding C2 reachable graph contains no ACT state. C1 requires three decision probes and two state moves, while C2 retains a hazardous alternative that blocks the informative probes. The certificates attribute each refusal to its original budget or safety constraint and establish value preservation by the abstraction.

Online budget controls the offline evidence requirement. C1 reaches 600/600 success at $B = 5$ with both 18 and 144 samples, while low-data C2 reaches 592/600. A zero-data calculation on the same 600 C1 true worlds also gives 600/600 success using five probes per trial. Mean actual probe use is 1.7067 with 18 samples and zero with 144. This is the envelope criterion of Proposition A.8: the entire C1 uncertainty set is resolvable at $B = 5$. Here offline evidence reduces online probe use, while $B = 4$ yields the preceding data-versus-success comparison.

C.3 Sound aggregation and structural refinement

The pessimistic construction applies to a general supplied partition. Fix a partition $\mathcal{P} = \{C_1, \ldots, C_k\}$ of $\mathcal{H}$ (declared, discovered, or wrong) and define the **pessimistic quotient** $\mathcal{I}/\mathcal{P}$: class signature = union of member signatures, class safety = conjunction (q is safe for C iff safe for every $w \in C$), class acceptability = intersection ($A_C(x) = \bigcap_{w \in C} A_w(x)$). Classes may contain arbitrary acceptability and safety profiles.

Proposition A.18 (unconditional one-sided soundness of any grouping)**.** *Run the gate on $\mathcal{I}/\mathcal{P}$ and execute its verdicts in any true world $w^* \in \mathcal{H}$. Then (i) the class $[w^*]$ is never eliminated by any observation; (ii) every executed probe is safe in w^*; (iii) every committed token is acceptable to w^*. Wrong partitions cost* completeness *(extra probes or abstentions), never validity.*

Proof. (i) An observation (x', y) of w^* lies in $\mathrm{sig}_{w^*}(x, q) \subseteq \mathrm{sig}_{[w^*]}(x, q)$ by the union semantics, so consistency retains $[w^*]$. (ii) A quotient-admissible probe is safe for every alive class, in particular for $[w^*]$, hence for every member by the conjunction semantics. (iii) An ACT token is common to all alive classes; by (i) $[w^*]$ is alive, and by the intersection semantics its tokens are acceptable to every member, including w^*. The shadow-quotient strategy follows the same action and cost sequence, so a finite worst-case resolution-cost bound also transfers. This extends T2's soundness argument to arbitrary partitions. Adding T3's row uniformity and certified product pattern establishes the converse direction, preserving completeness and optimal resolution cost. CE2 exhibits a strict increase in resolution cost under coarsening. □

The proposition supplies a sound interface for every partition. T3 identifies the product structure under which the quotient also preserves completeness and optimal resolution cost. Refining the

partition preserves soundness while restoring distinctions required by the full game's continuation strategies.

USING THE INTERFACE

An exact application specifies hypotheses, transition supports, safety predicates, and deployment tokens. AutoDNF may compress that instance when its structure passes certification; otherwise one can retain a finer representation and accept the associated planning cost. The gate then solves for a common policy or a safe resolution plan within budget. A learned application specifies its observation map, calibration source, commitment scores, and task outcome predicate. The LeWorldModel implementation specifies its learned descriptors, source budgets, planning configuration, and task-risk protocol below.

D EXPERIMENTAL PROTOCOLS AND IMPLEMENTATION

Object	Assumptions and protocol	Result
Exact resolution	Finite truth-containing world set; sound transition supports, safety predicates, and acceptable policy tokens.	Certified ACT/PROBE/ABSTAIN verdicts and preserved safe continuation cost (T1–T3).
Continuation responses	Bounded finite horizon; uniform response, cost, and value-closure error bounds; common admissible actions.	Value and policy-loss bounds, with exact action preservation under a sufficient margin (T7).
Deployment-weighted responses	Fixed bounded scores; common initial law and feasible actions; actual policy occupancies.	Expected cost comparison controlled by occupied Bellman residuals and modeled action gaps (Prop. A.11).
Calibrated stochastic frontier	Bounded observed contract losses; independent episode calibration; coverage for off-policy reuse; fixed response class or controlled covering number.	High-probability trajectory-risk feasibility and a computable cost gap; explicit rates under contract slack (T8).
Visual planning	Declared training source, fixed controller and task pairs, and specified planning and observation budgets.	Measured success at each data budget and paired component contributions.
Frozen-policy admission	Controller and admission rule fixed before uniform sampling from a finite task catalogue.	Confidence bound on task-failure risk among remaining accepted catalogue items.

Table 7: **Decision objects, assumptions, and evaluation.** Each result is stated for its corresponding mathematical interface or experimental protocol.

Data and evaluation units. The LeWorldModel source-frame ratio is $2{,}336{,}736/291{,}969 = 8.003$. Protocols also record auxiliary fitting, development feedback, calibration, online RGB observations, and inference. Frame selection reuses the source through a lookup. The finite and Boxoban studies report their respective passive-table and own-level interaction counts.

Task construction and comparisons. The standard-start pool contains 640 pairs from training-source trajectories. The primary comparison and score intervention report the 560 pairs outside the fixed 80 development starts. The role comparison uses the same exclusion after selecting each representation's score weight on the development set. This split separates controller feedback from evaluation while retaining source-trajectory task construction. Checkpoints and planning caps are paired within each source; each component intervention fixes its specified image frame and controller configuration. Individual paired intervals resample task items conditional on these frozen models. The admission study below uses its own frozen rules and finite-catalogue sampling protocol.

Environment and outcome definitions. Finite families specify exact world tables, supports, and admissibility predicates. Boxoban-Slip augments external Boxoban levels with the stated ice dynamics and compatible-map model. These studies measure safe commitments and probes. The visual study measures PushT task success, and admission reports accepted-task risk together with coverage on the specified catalogue.

D.1 AUTOMATIC VISUAL FACTORS AND SOURCE-SUPPORTED PLANNING

The source budgets are 291,969 and 2,336,736 frames from 2,328 and 18,685 episodes. Each source fits its own visual mapper, factor predictor and proposal from its own frames, and uses its own frozen LeWorldModel members.

Learning role membership. Palette components are grouped by their response to observed actions. A maximum response-gap split distinguishes controlled from passive moving components; fragments assigned the same role are merged, and learned static occupancy removes background. The output vocabulary contains robot position, object position and two orientation coordinates. Thus the learned assignment operates within a declared six-coordinate interface. Its RGB descriptors feed the factor dynamics, proposal, and geometric scoring modules. The residual dynamics head and context-conditioned proposal each receive 10,000 fitting updates on their own source.

A frame supported by the training source. Before the first action, the current and goal images are evaluated in the four right-angle coordinate frames. Write g_k for the kth rotation and ϕ for the frozen RGB role map. With D containing source-only feature scales, the selected frame is

$$k^* = \arg \min_{k\in\{0,1,2,3\}} \min_{(z_i,z_i')\in\mathcal{S}_{\mathrm{src}}} \left\| D^{-1}\left(\begin{bmatrix} \phi(g_k I_0) \\ \phi(g_k I_g) \end{bmatrix} - \begin{bmatrix} z_i \\ z_i' \end{bmatrix} \right) \right\|_2 . \tag{D.1}$$

The support contains role pairs from training-source frames separated by the goal offset. The selected frame remains fixed throughout the episode; the complete supplied image history is rotated, and actions are mapped back before execution. Before planning, selection uses the current and goal images together with training-source support. Observation histories and relative actions then remain in the selected coordinate system.

The source-supported frame. Equation D.1 uses only training-source frame pairs separated by the fixed 25-step goal offset. The retained pair counts are 209,939 in the reduced source and 1,676,231 in the full source. Source feature means and scales are fixed before evaluation. The minimum-distance frame is selected once from the four right-angle rotations, with a fixed tie order. Views with unavailable role readouts are omitted; if none is valid, the original frame is retained. The entire supplied image history and goal are transformed; relative actions are inverse-transformed before they reach the environment. Frame selection is a pre-planning source lookup over fixed models and previously acquired training frames.

Planning and accounting. The main comparison retains three source-specific backbone members, 300 CEM candidates and 30 optimization rounds per solve, with the original six-plan schedule and 50 physical steps. Frame selection, latent prediction, factor prediction, proposal fitting and online RGB are reported as different parts of the pipeline. Realized model calls and trajectory lengths are recorded alongside the nominal planning caps. Start, environment and planner seeds are paired within each fixed controller comparison.

Module and feedback budgets. Table 8 separates acquired source frames, derived training records, fitting updates and task-feedback labels. The source-frame count is the union reused by the backbone, mapper, factor dynamics, proposal and frame selector. Acquisition counts each source frame once; fitting updates are counted separately. Feedback counts describe their respective stages, whose episode sets can overlap.

Each executed task supplies its initial and goal images, then RGB observations along its physical rollout. Result records retain physical steps, CEM calls and additional terminal-feedback reads for each arm; the common upper bound is 50 physical steps. Synthetic rotations reuse the supplied images. The admission catalogue excludes the union of 7,667 previously used feedback episodes before drawing its 640 items.

D.2 LEARNING BOUNDED CONTINUATION RESPONSES ON TWOROOM

Source data, initialization, and evaluation. The full source contains 1,200 native TwoRoom episodes of 51 frames; the fixed reduced subset contains 75 episodes. Each arm uses its own source action normalization and its own frozen PushT initialization. The reduced initialization acquired 144,853 source frames and the full initialization acquired 2,336,736. Thus task-only counts are 3,825 and 61,200, while the respective unions including initialization are 148,678 and 2,397,936.

Stage and counting unit	Reduced source	Full source
Acquired source RGB/action records	291,969	2,336,736
Source episodes	2,328	18,685
Valid RGB-derived factor records	291,418	2,332,495
Source pairs retained for frame lookup	209,939	1,676,231
Factor-dynamics fitting updates	10,000	10,000
Proposal fitting updates	10,000	10,000
Fixed score-selection development starts	80	80
Reported primary comparison starts	560	560
Admission-critic feedback labels used	2,400	720
Admission-threshold selection feedback	1,200	1,200
New frozen-rule catalogue evaluation starts	640	640

Table 8: Visual pipeline accounting. The last three rows concern the separate admission experiment; full-source counts are per frozen rule. The reduced-source critic uses 1,676 fitting and 724 validation labels; the retained full-source RGB critics are fitted on all 720 own-controller labels. The main comparison evaluates every task under its unfiltered controller.

Initialization data are charged even though their model weights are reused. New evaluation episodes are separate from source collection. Each fixed trained model executes all 200 evaluation episodes; the same environment and planner seeds are used across models. Images and actions are model inputs. Simulator state is used for task initialization and the native success predicate.

Bounded response objective. Let $z_0, \dots, z_5$ be embeddings of six source images spaced by five physical steps, and let $\widehat{z}_t$ be the five-step free rollout conditioned on the recorded action blocks. Goals comprise the five observed future embeddings and the final embedding of another source window in the same batch. With stop-gradient targets, define

$$d(x,g) = \tfrac{1}{p}\|x-g\|_2^2, \qquad s_g = d(z_0,g) + 10^{-4}, \qquad r(x,g) = \frac{d(x,g)}{d(x,g)+s_g}.$$

The auxiliary objective is

$$\mathcal{L}_{\text{resp}} = \tfrac{1}{2}\mathbb{E}_{t,g}\big(r(\widehat{z}_t,g) - r(z_t,g)\big)^2 + \tfrac{1}{2}\mathbb{E}_{\text{window}} \max_{t,g}\big(r(\widehat{z}_t,g) - r(z_t,g)\big)^2.$$

Both terms operate on observed source continuations. Bounded responses make goal discrepancies comparable across the source windows; the maximum term emphasizes the largest horizon/goal discrepancy in each window. These sampled, bounded responses operationalize T7's continuation-value geometry; its transfer bound is stated under uniform response and value-approximation conditions.

Matched learning and planning. For each source and each of three seeds, we fit the original one-step objective, the five-step rollout control, and that same rollout control with $0.5\mathcal{L}_{\text{resp}}$. All fits use 4,000 updates, batch size 128, learning rate $5 \cdot 10^{-5}$ and the same initialization within the source/seed pair. The two rollout arms share their six-frame windows, source order, rotation augmentation, and loss weights on all common objectives. Their encoder and predictor are learned jointly. Evaluation uses one frozen backbone, 300 CEM samples, 30 rounds, five action blocks of five steps, and a 50-step physical budget. The action-domain projection is common to every arm. The PushT role mapper, factor dynamics, proposal and terminal servo are absent from every TwoRoom policy. Table 9 compares the three objectives.

Training objective	75 source episodes	1,200 source episodes
One-step prediction	32.00 ± 2.60	35.17 ± 0.58
Five-step rollout control	51.83 ± 3.01	53.33 ± 6.21
Rollout + continuation response	$\mathbf{57.17 \pm 9.93}$	$\mathbf{58.83 \pm 6.29}$

Table 9: **Continuation responses improve the rollout objective.** Success percentages: mean and sample standard deviation across three training seeds, each evaluated on 200 common tasks. Within each source and seed pair, initialization, fitting budget and planner are fixed. Bold marks the highest mean within each source.

The paired response-objective gain over the rollout control is 5.33 ± 6.93 points in the reduced source and 5.50 ± 2.78 at full source. These means and sample standard deviations use the three trained-model pairs, keeping training randomness distinct from resampling the shared evaluation tasks.

Intermediate source and source-subset variation. The intermediate source contains 150 episodes, with 7,650 task frames and 291,969 initialization frames, for 299,619 total source frames. Its subset contains the original 75-episode source and preserves the whole-episode training/validation split. The same three training seeds give 49.00% success for the rollout control and 52.50% for the continuation-response objective (sample standard deviations 1.73 and 0.87 points). Each model retains the 4,000-update fitting recipe and all 200 evaluation tasks.

Two additional 75-episode draws sample 67 training and eight validation episodes from their respective fixed pools, without using evaluation outcomes. Crossing these draws with the same three training seeds adds twelve paired-objective models. Across all three source draws and three training seeds, success averages 52.17% for the rollout control and 55.94% for the response objective, a 3.78-point gain. The nine fits per arm cross three source draws with three training seeds and use one shared panel of 200 evaluation tasks (Table 10).

75-episode source draw	Rollout control	Response objective	Paired gain
Original	51.83	**57.17**	5.33
Additional 1	48.17	**49.33**	1.17
Additional 2	56.50	**61.33**	4.83
Mean across draws	52.17	**55.94**	3.78

Table 10: Success percentages averaged across three training seeds within each fixed source draw. The gains pair the same source, initialization, training seed, and evaluation tasks.

D.3 Independent trajectory-window confirmation

The independent confirmation uses 400 newly generated PushT trajectory windows, disjoint from the 80 development windows. Each trajectory follows the native weak policy, with a starting offset drawn uniformly from 0 through 103 and a goal 25 physical steps after that offset. Prefix replay reconstructs the physical initialization. Policy input begins at the selected offset; the replayed prefix is confined to initialization. Task generation and inclusion follow the fixed native policy, offset distribution and goal-spacing rule. The four source/controller combinations retain their frozen backbones, mappers, proposal modules, source supports, and the 300-candidate, 30-iteration CEM cap within 50 physical steps.

All 400 tasks are shared across the four arms. The reduced-source complete interface succeeds on 184 tasks and its latent-only control on 121. The full-source complete interface succeeds on 214 and its latent-only control on 191. The paired improvements are 15.75 percentage points ($[10.50, 21.00]$) and 5.75 ($[0.75, 10.75]$). Intervals resample tasks paired across frozen checkpoints. Table 11 separates source frames from online interactions.

Source	Controller	Source frames	Physical steps	CEM calls	Extra RGB reads
$1/8$	Complete	291,969	15,345	1,607	4,777
$1/8$	Latent only	291,969	17,461	1,984	0
Full	Complete	2,336,736	14,814	1,519	4,396
Full	Latent only	2,336,736	15,727	1,670	0

Table 11: Data and interaction counts for independent windows. Source acquisition and online execution are distinct units. Extra RGB reads count the interface's additional visual-feedback accesses beyond the latent planner's regular observations. Every listed policy executes all 400 tasks, with early stopping only under the native success rule.

D.4 Frozen-agent inference intervention on pixel Reacher

Agent and acquired source. The study uses a fixed role-augmented agent from the PyTorch DreamerV3 reproduction in the R2-Dreamer codebase (Hafner et al., 2025; Morihira et al., 2026) with 64×64 RGB input, bounded two-dimensional actions, and action repeat two. The charged training budget is 1,010,000 physical environment steps. The model uses batch size 16, sequence length 64, train ratio 512, a 2,048-dimensional deterministic state, and 32 categorical variables with 16 classes each. Encoder, dynamics, reward, continuation, actor, and value parameters are fixed in this intervention.

The 20-episode source preload contains 2,560 transitions and 2,580 images. Its 5,120 physical acquisition steps and 5,120 reward-reconstruction steps are included in the charged budget; 999,760 steps are subsequent interaction. The role interface predicts endpoint and target image positions and a residual endpoint transition at horizons one, two, and four. Actor and critic inputs append four normalized role coordinates to the recurrent features.

Paired controllers and fixed cohorts. The reference uses the original learned role readout. The intervention uses the affine source calibration and convex factor inference of App. A.13. Both controllers use the same neural checkpoint, images, action space, and evaluation seeds. Calibration fits twelve affine coefficients from the acquired source preload. Controller development uses 32 episodes beginning at seed 1957001. The two independent cohorts each contain 200 episodes, beginning at 1969001 and 2019001. All episodes last 1,000 physical steps; consecutive groups of four share the prescribed posterior-sampling generator. The actor uses its native evaluation mode and the recurrent posterior retains stochastic sampling.

Cohort	Original readout	Convex inference	Gain	Paired 95% interval
First 200	912.89	**974.69**	$+61.81$	$[31.50, 95.94]$
Second 200	906.08	**953.24**	$+47.16$	$[12.86, 84.26]$
All 400	909.48	**963.96**	$+54.48$	$[31.31, 79.39]$

Table 12: **Convex inference improves the same frozen controller in both independent cohorts.** Neural parameters and acquired source are fixed. Intervals resample paired tasks; both complete cohorts contribute to the pooled estimate. Means and differences are rounded separately from the underlying paired episode records.

Table 12 gives the paired returns. Because the checkpoint, acquired source and training updates are shared by both arms, the reported gain is the return change caused by the deployed inference layer.

D.5 Actual admission with finite-population risk control

The admission experiment uses the reduced- and full-source controllers and retained full-source reference. On development feedback, each RGB rule maximizes coverage under a pointwise CP90 ceiling of 3%. Rules and controllers are then frozen before catalogue sampling; the new evaluation supplies the risk bounds below.

After excluding 7,667 feedback episodes, we construct a finite catalogue of $M = 11{,}018$ eligible episodes. Before sampling, each catalogue item receives one fixed start, environment seed, and planner seed. We then uniformly sample $n = 640$ items without replacement, shared across the three controllers. Starts follow the same source-trajectory sampling scheme as the standard pool. All 640 sampled items are run, and batch size one with independent planner generators makes each item's control independent of the other sampled items.

Admission uses the RGB image rendered after reset and the goal image, and is decided at the initial start/goal pair before any environment step. For accepted items, admission leaves the frozen action policy unchanged, allowing reuse of its trajectories in the raw comparison. Denied items are executed without admission for that reference comparison; their actual deployment records contain zero steps. The raw comparison contains 1,920 trajectories over 640 common items; denied deployments remain unsuccessful in the deployment denominator.

Frozen rule	Accepted / 640	Successful / ACT	Coverage	$1-U$
Automatic, $1/8$ source	502	488/502	78.44%	95.1301%
Automatic, full source	544	531/544	85.00%	95.7367%
Retained full reference	474	473/474	74.06%	98.6970%

Table 13: Frozen-policy task admission. $1-U$ is the success lower bound among remaining accepted catalogue items. Coverage uses all 640 starts; the success fraction uses accepted starts. Denied deployments contribute zero successes and execute zero backend steps.

For one fixed rule, let N be its unknown accepted catalogue size and K its total accepted failures. The sample reveals m acceptances and k failures. Write $H_{a,b,c} \sim \mathrm{Hypergeom}(a,b,c)$ and set $\eta = 0.1/6$. Exact count inversion gives

$$\begin{aligned} N_U &= M - n + m, \qquad L = \min\{N \in [m, N_U] : \Pr(H_{M,N,n} \geq m) > \eta\}, \\ K_+(N) &= \max\{K \in [k, N-m+k] : \Pr(H_{N,K,m} \leq k) > \eta\}, \\ U &= \max_{N=\max(L,m+1),\ldots,N_U} \frac{K_+(N)-k}{N-m}. \end{aligned} \tag{D.2}$$

Conditioned on m, observed accepted items form a uniform m-subset of the true accepted population. The lower bound for N and the failure bound at the true N each spend at most η. Maximization over compatible N transfers the count bound to the remaining accepted population with the same error allocation. A union bound gives $6\eta = 0.1$ across the three fixed rules, without independence between rules. The simultaneous 90% bounds apply to the three rules' unobserved accepted subsets within the remaining 10,378 catalogue items. For $m = 0$, the risk bound is 1; for $N = m$, the accepted remainder is empty.

All three bounds are below 5% (Table 13), with coverage 78.44%, 85.00% and 74.06% respectively.

Data units across the studies. The finite experiments count passive table observations; their $N = 14$ versus $N = 144$ retention lower bounds are 0.938 and 0.930. The visual experiments count acquired source RGB/action records and compare controllers within each source budget. Boxoban-Slip counts own-level interactions against its specified competing planner. The protocols record online interactions separately.

D.6 Matched-data controls and role-representation comparison

Fixed 560-start comparison. Table 2 compares the frozen source-supported interface against the original latent-only control. All four arms use the same task, environment and planner-seed arrays. The complete low-source controller succeeds on 518 starts, versus 480 for its same-source latent baseline. Full-source counts are 536 and 514. The paired differences are 38 and 22 successes. The individual 95% paired intervals are $[3.93, 9.64]$ and $[1.79, 6.07]$ percentage points. Arms are paired within each source; when an earlier complete run already used the same arrays, its results are reused.

Original-frame score intervention. The LeWM-score control fixes the original image frame, reuses the same 560 standard-start pairs and retains the original source models, factor score, proposal, servo and planning schedule. Each cost call computes the LeWM score and then replaces its contribution by zero. The intervention changes the objective while keeping forward computation fixed, using the existing cohort.

Source, fixed original frame	Score retained	Score removed	Difference (pp)
$1/8$	519/560	397/560	+21.79
Full	533/560	405/560	+22.86

Table 14: LeWorldModel-score contributions conditional on the original frame and retained auxiliary controller. The complete-interface comparison in Table 2 uses source-frame selection.

The score contributes 122 and 128 additional successes at reduced and full source, respectively (Table 14). Their 95% paired intervals are $[18.04, 25.54]$ and $[19.11, 26.61]$. Both source regimes

show a positive contribution from the learned latent prediction score. Realized stopping times can change total compute despite retaining every forward call at a given solve.

Role representations with equal selection budgets. Both arms fix the original image frame and disable visual servo and admission. The role-based arm reuses its source-fitted grouped heads; the role-free arm refits its dynamics and proposal with the same source, eligible windows, split, seeds, widths and 10,000-update recipe. Its 16 binary presence coordinates use natural-unit scales of one. The other 112 continuous coordinates retain their source scales, avoiding amplification of rare absences.

Each arm tests auxiliary-score weights $\{0, 0.1, 1\}$ on the same 80 development starts, choosing the largest success count with a fixed tie order. The selected weights are one for the grouped arm and 0.1 for the role-free arm. The seven reserved 80-start batches exclude that development batch. Grouped success is $526/560$ (93.93%) versus $489/560$ (87.32%), for a 6.61-point difference with 95% paired interval $[4.11, 9.11]$. This comparison replaces the representation and fitted modules within the specified visual vocabulary.

Intervals and reproducibility. Each descriptive interval uses 20,000 paired task resamples conditional on its fixed configuration. Role, score, and complete- interface interventions have their own paired references and are reported as individual contrasts. Numerical values and intervals for the summary figures accompany the manuscript sources (App. E).

D.7 Exact finite-game confirmation of the sample exponent

The staged game has decision worlds $\{0, 1, 2, 3\}$, one physical state, and a distinct acceptable token in each world. Three probes have unit cost. The split probe is safe in every world and returns $\lfloor w/2 \rfloor$. The low probe is safe only in $\{0, 1\}$ and returns w; the high probe is safe only in $\{2, 3\}$ and returns w. Value iteration yields resolution cost 0 on singletons, 1 on all six pairs, and 2 on all four triples and on the full set. Thus at budget 0 the obstructions containing a true world are its three pairs; at budget 1 they are its three triples; at budget 2 or more the obstruction family is empty.

The passive menu has seven deterministic tests. The identity test e_w returns 1 only at world w. The group tests return 1 on $\{0, 1\}$ and on $\{2, 3\}$ respectively, and the constant test returns 0. A sample's effect depends only on the set of distinct tests it contains. Dynamic programming over that set gives the exact probability of each support. The surviving worlds are those matching the true world on every test in the support, and the trial fails when their resolution cost exceeds the budget. On every support, every true world, and budgets 0, 1, and 2, that event agrees with survival of a minimal obstruction, matching the theorem exactly.

At budget 0 the maximin solution puts mass $1/4$ on each identity test. Its hit rate is $1/2$, its covering number is 2, and the primal–dual gap is 0. Each of the seven pure tests has worst-world hit rate 0. Uniform mass on the six non-constant tests has hit rate $1/3$. Weighting tests by the number of obstructions they hit has hit rate 0.30 and reaches failure 0.05 at nine samples, equal to its necessary count. The reported probabilities use $\eta = 0.05$. An independent sampler with $20{,}000$ trials per world at $N = 8$ differs from the maximin probability 0.011673 by 3.8×10^{-4}.

The same program recovers the three-world closed forms (EX1) and (EX0) with maximum absolute error 1.4×10^{-16}, and the design values $z_0^* = 1/2$ and $z_1^* = 1$ with zero duality gap.

The random-channel panel draws 40 menus of six distinct binary tests with seed 20260922, at budgets 0 and 1. Each menu is evaluated at the maximin allocation, the uniform allocation, and two Dirichlet allocations. Of these, 280 have hit rate strictly between 0 and 1 and successive failure probabilities above 10^{-8}. Their median absolute exponent gap is 2.8×10^{-4} and the largest is 0.032. The other 40 are boundary cases of T5: the hit rate is 0 or 1, or the probability has fallen below the numerical floor. Every tested allocation satisfies the T5 band and the necessary/sufficient sample-count sandwich at $\eta = 0.05$. The accompanying code regenerates the figure and the recorded probabilities.

D.8 Exact confirmation of continuation-response transfer

T7 is checked by backward induction on finite layered games. The true and modeled games share states and actions and may differ in costs and successor sets. A successor set responds by the convex hull of its state features. The stage response error is the largest ℓ_∞ Hausdorff distance between those

hulls, from the linear programs in Corollary A.5. The certificate uses the linear fit of the modeled continuation that minimizes $L\delta + 2\rho$, together with the realized cost and terminal gaps.

Five one-step games make that fit exact and move every successor by a known feature gap. Their value errors equal the certificate, with values 1, 2.5, 1.5, 1.4, and 3. On 80 random horizon-3 games with two-dimensional features and seed 20260922, the initial value error is always below the certificate. The median ratio of error to certificate is 0.126 and the largest is 0.524 (Fig. 9).

The action rule is checked on the same recursion. Eight one-step pairs keep the true action at all 28 states whose gap exceeds twice the certificate. The matched pairs with a smaller gap change the selected action at all 28 states. The threshold is therefore sufficient, and a gap below it can move the action.

The deployment identity is evaluated by occupancy recursion on 40 horizon-4 MDPs with four states. The largest absolute discrepancy is 1.3×10^{-15}. On a two-path cost the evaluated policy returns 1 above the comparator, and the comparison bound equals that gap. The accompanying code regenerates the figure and these values.

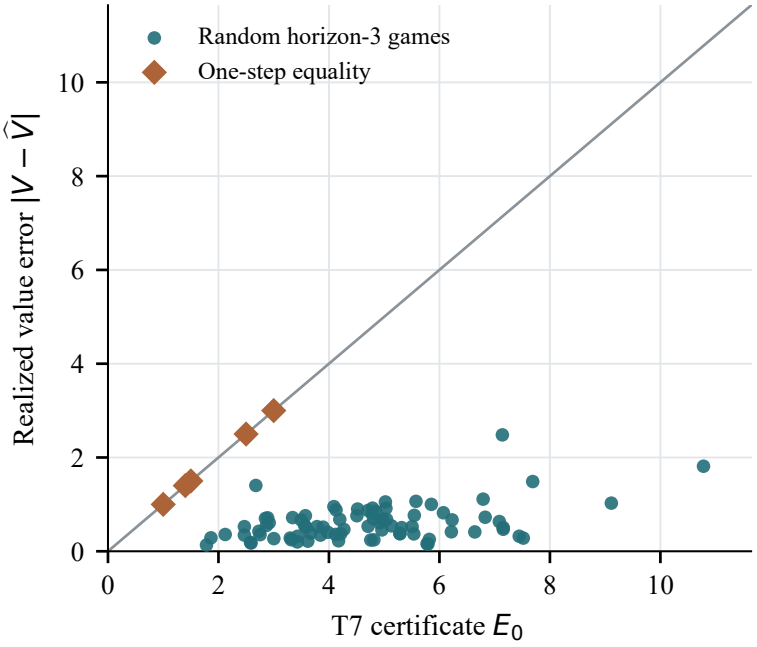

Figure 9: **The continuation certificate is valid and attainable.** Diamonds are one-step games whose value error equals E_0. Points are 80 random horizon-3 games; all lie on or below the diagonal.

E SUPPLEMENTARY FILE MAP

The supplementary archive contains the code for the reported experiments and their per-trial records, covering the finite-game, Boxoban-Slip and AutoDNF studies, the TwoRoom, PushT and Reacher evaluations, and the checks for convex allocation, continuation transfer and the stochastic frontier. One included script recomputes the counts, means and paired differences reported in Sections 4 and 5 from these records. Model weights and training datasets are omitted because of their size; the archive description names the public code releases and the training scripts needed to regenerate them.