DUAL-FRONTIER: WHEN CAN AN AGENT TRUST ITS WORLD MODEL?

Anonymous authors
Paper under double-blind review

ABSTRACT

Learned world models are becoming essential to general-purpose agents: by predicting action consequences, they support planning and decision-making while reducing reliance on costly trial and error. This reliance creates a fundamental ambiguity: when a world-model-guided decision fails, the trajectory alone may not reveal whether the agent's decision rule or the world model caused the loss. We formalize this *failure-attribution problem* as a counterfactual decomposition of return loss and prove that its components are not identifiable from passive interaction, even for finite-horizon planners. This obstruction motivates *Dual-Frontier*, a learning principle that admits a world-model-guided decision only when its predicted advantage exceeds a certified bound on decision-relevant world-model error; otherwise, evidence is allocated to world-model verification. Action-conditioned value bounds and a closed-loop extension guarantee non-decreasing return for admitted decisions. Calibrated gates and simultaneous confidence sequences support adaptive evidence reuse, with sufficient and necessary verification bounds. Controlled learned-model experiments validate the predicted failure modes and certification behavior, while cross-backbone tool-use benchmarks instantiate the same verify-then-promote rule in realistic agent world-model pipelines, consistently improving decision quality and reliability.

1 INTRODUCTION

World models are becoming core infrastructure for agents that act beyond their immediate observations. Consequential action requires anticipating how choices alter future states, information, and opportunities: general multi-step agency entails recoverable knowledge of environmental dynamics (Richens et al., 2025), even under partial observability and stochasticity (Cifuentes, 2026). Learned world models operationalize this knowledge as action-conditioned predictors rolled forward before execution. Whether expressed in latent states, video, language, or internal dynamics, they augment the present observation with forecasts used to compare actions, plan farther ahead, and reduce costly online trial and error.

This predictive interface now spans latent-imagination control in DreamerV3 (Hafner et al., 2025) and scalable model-predictive control in TD-MPC2 (Hansen et al., 2024), as well as navigation (Bar et al., 2025), manipulation (Assran et al., 2025), driving (Russell et al., 2025), and web interaction (Chae et al., 2025). It is increasingly adaptive: WorldEvolver revises predictive memory at test time (Zhang et al., 2026b), CoMAP alternates world-model adaptation with agent reflection (Liu et al., 2026), and recent systems co-train predictive knowledge and policies (Lu et al., 2026) or co-evolve simulators with agents (Guo et al., 2025). Reliability is thus part of the decision mechanism. Short rollouts limit model exploitation (Janner et al., 2019); horizon-calibrated uncertainty addresses compounding error (Wan et al., 2026); safe-improvement methods constrain policy changes (Delgrange et al., 2026).

Yet a poor world-model-guided outcome poses an unresolved question: *what should improve next—the agent's decision rule or the world model on which it relied?* The same trajectory can arise because an inadequate rule ignored an accurate forecast or because an inaccurate forecast misled an otherwise sound rule. Passive interaction records only their composition. We call separating these causes *failure attribution*. The distinction is operational: learning against a misleading world model can reinforce a bad decision, while gathering more world-model data after the relevant forecast is adequate wastes evidence. Nor can global prediction accuracy decide the issue. Perceptual quality and closed-loop

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.

success can diverge (Zhang et al., 2026a); an arbitrarily small transition error may reverse nearly tied actions, whereas a large error in an irrelevant coordinate may alter none. The missing object is decision-specific: does current evidence establish that a world-model-proposed behavior improves upon an explicit reference?

To solve this attribution-and-trust problem, we develop **Dual-Frontier**, a theory of decision-specific trust for learned-world-model agents. It evaluates the true-return contrast between a candidate behavior proposed with the world model and a reference behavior. The world-model frontier retains promising comparisons lacking evidence; the agent frontier admits those remaining beneficial after world-model and estimation uncertainty. This *verify-then-promote* rule directs unresolved comparisons to targeted verification and certified ones to agent improvement, without interpreting rejection as proof of model failure. Our contributions are:

- **Failure attribution.** A fixed-operator counterfactual decomposition separates agent deficiency from world-model effect; a two-step construction proves passive non-identifiability for exact and Monte Carlo planners.
- **Decision reliability.** An exact Bellman-residual identity converts action-conditioned world-model error into comparison-specific radii, sharp separations, and a closed-loop condition guaranteeing nonnegative expected improvement.
- **Dual-frontier learning.** Calibrated promotion rules reuse one simultaneous certificate under adaptive world-model and request selection, with progress and matching-order sufficient/necessary evidence bounds.
- **Empirical validation.** Learned-model experiments test non-identifiability, harmful imagined improvements, and qualification; matched agent–world-model evaluations separate reliability, realized quality, and verification cost.

2 Related Work

General-purpose agents in interactive environments. Modern agents transact with websites (Zhou et al., 2024), operate desktops (Xie et al., 2024), repair repositories (Jimenez et al., 2024), and act in embodied environments (Assran et al., 2025). AgentBench spans eight interactive settings (Liu et al., 2023), while continual agents face changing objectives (Liu et al., 2025). Across domains, multi-step attainment requires recoverable predictive knowledge (Richens et al., 2025), including under partial observability and stochasticity (Cifuentes, 2026). Learned world models expose such knowledge for planning, from navigation (Bar et al., 2025) to web interaction (Chae et al., 2025). Prior work primarily measures end-to-end success or builds domain-specific predictors, leaving forecasts embedded in the system. Dual-Frontier instead isolates the world-model–decision-rule interface and asks whether evidence supports one proposed behavior comparison.

Learned world models for agent planning and adaptation. We use *learned world model* for an action-conditioned predictor that compares future courses of action, whether latent, visual, textual, or internalized. The lineage extends from recurrent simulators (Ha & Schmidhuber, 2018) and value-equivalent models (Schrittwieser et al., 2020) to DreamerV3 (Hafner et al., 2025), TD-MPC2 (Hansen et al., 2024), and multi-task policy learning (Georgiev et al., 2025). Current systems forecast navigation (Yao et al., 2025), manipulation (Assran et al., 2025), driving (Russell et al., 2025), and web transitions (Chae et al., 2025). WorldEvolver adapts predictive memory online (Zhang et al., 2026b); CoEx updates persistent beliefs during exploration (Kim & Hwang, 2025). WebEvolver combines synthetic trajectories with look-ahead planning (Fang et al., 2025), while CoMAP alternates world-model adaptation and reflection (Liu et al., 2026). PaW co-trains policy and world model (Lu et al., 2026); GenEnv co-evolves agents and simulators (Guo et al., 2025); DreamGym and Agent World Model scale synthesized interaction (Chen et al., 2026; Wang et al., 2026). These systems optimize or exploit prediction to improve the agent, but do not identify whether a failed decision implicates its rule or its forecast. Dual-Frontier formalizes that ambiguity and qualifies a behavior comparison rather than a world model globally.

Reliable model-based learning and adaptation. World-model exploitation motivates short rollouts (Janner et al., 2019) and pessimism (Yu et al., 2020); policy-aware learning targets downstream

gradient error (Abachi et al., 2021; D'Oro* et al., 2020). Recent work further studies horizon-dependent uncertainty (Wan et al., 2026), local safe improvement (Delgrange et al., 2026), and confidence-filtered foresight (Zhang et al., 2026b). WAKER collects data where estimated world-model error is high (Rigter et al., 2024); AdaWM separates dynamics and policy mismatch under transfer, then fine-tunes the indicated component (Wang et al., 2025). These methods mainly limit error or respond to an assumed diagnostic. By contrast, Dual-Frontier first proves that passive failure need not identify its source, then links action-conditioned error to value and imagined-gradient distortion. WAKER targets accuracy across environments, whereas we certify a behavior comparison; AdaWM selects adaptation under shift, whereas we establish when a world-model-guided change is actually justified in practice. Split conformal calibration (Angelopoulos & Bates, 2022) and simultaneous confidence yield certificates valid under adaptive evidence reuse, connecting attribution, decision reliability, and allocation in one formulation.

3 SET UP

For task z, let $\mathcal{M}_z = (\mathcal{S}, \mathcal{A}, P_z, r_z, \rho_z, H)$ and $\widehat{\mathcal{M}}_z = (\mathcal{S}, \mathcal{A}, \widehat{P}_z, \widehat{r}_z, \rho_z, H)$ denote the true world and its learned world model. The spaces are standard Borel, $H \geq 1$, rewards lie in $[0, R_b]$, and both systems share an information interface (Appendix A). Suppressing z, let $J, \widehat{J}$ be undiscounted returns, μ_t^π the true law of (s_t, a_t), and $\mathrm{TV}(P, Q) = \sup_B |P(B) - Q(B)|$.

A fixed operator A_ϕ maps a supplied world to a policy, with ϕ fixing its search, information access, budget, and randomization. Replacing only the world defines

$$\widehat{\pi} = \mathsf{A}_\phi(\widehat{\mathcal{M}}), \qquad \pi^\circ = \mathsf{A}_\phi(\mathcal{M}). \tag{1}$$

Let $J^* = \sup_{\pi \in \Pi} J(\pi)$, where Π contains both outputs.

Definition 1 (Counterfactual attribution)**.** *Let*

$$\begin{aligned} A &:= J^* - J(\pi^\circ), \qquad W := J(\pi^\circ) - J(\widehat{\pi}), \\ R &:= J^* - J(\widehat{\pi}) = [J^* - J(\pi^\circ)] + [J(\pi^\circ) - J(\widehat{\pi})] = A + W. \end{aligned} \tag{2}$$

Here $R, A \geq 0$, whereas W is a signed world-model effect.

The decomposition is relative to one fixed predictive decision procedure; because model error may accidentally help a limited agent, W is signed.

Theorem 1 (Passive non-identifiability)**.** *For every $c \in (0, R_b]$, a fixed known operator and supplied world model admit a family of two-step true worlds, indexed by $\lambda \in [0, c]$, with $(A, W) = (\lambda, c - \lambda)$ and identical laws of arbitrarily many deployed episodes. Any attribution estimator based on these episodes, the known operator, and the supplied world model satisfies*

$$\begin{aligned} \sup_{\lambda \in [0,c]} \mathbb{E}_\lambda |\widehat{A} - \lambda| &\geq \frac{1}{2} \left(\mathbb{E}_0 |\widehat{A}| + \mathbb{E}_c |\widehat{A} - c| \right) \\ &= \frac{1}{2} \mathbb{E} \left[|\widehat{A}| + |\widehat{A} - c| \right] \geq \frac{c}{2}. \end{aligned} \tag{3}$$

where the middle expectation is under the common observational law. Endpoint classification has minimax error $1/2$. Both bounds are attained, for exact and finite-sample Monte Carlo planning.

Appendix B gives the construction and minimax proof. Notably, the supplied world model may be exact on the deployed occupancy; the ambiguity can reside entirely in untried actions. The obstruction is therefore passive rather than a prohibition on verification: intervention can recover missing information, whereas an unstable operator can amplify small predictive error (Proposition 1). We therefore certify a specified next use rather than infer ownership from passive failure.

4 WORLD-MODEL RELIABILITY

4.1 CERTIFYING A PROPOSED USE

A request $x = (z, \widehat{\mathcal{M}}, \pi_0, \pi_1)$ specifies a learned world model, a reference behavior π_0, and a candidate π_1, including their continuation rules. Define

$$\Gamma_{\mathcal{M}}(x) = J_{\mathcal{M}}(\pi_1) - J_{\mathcal{M}}(\pi_0), \qquad \ell_{\mathcal{C}}(x) = \inf_{\mathcal{N} \in \mathcal{C}} \Gamma_{\mathcal{N}}(x). \tag{4}$$

On $\mathcal{M} \in \mathcal{C}$, $\ell_{\mathcal{C}} > 0$ certifies improvement; failure to certify does not imply harm. The policies remain fixed while the evaluation world varies (Appendix C.1).

For bounded f, write $Pf = \int f(s')P(\mathrm{d}s' \mid s, a)$ and $\mathrm{span}(f) = \sup f - \inf f$. With learned value $\widehat{V}_t^\pi$ and $\widehat{V}_H^\pi = 0$, set

$$\Delta_t^\pi = r - \widehat{r} + (P - \widehat{P})\widehat{V}_{t+1}^\pi, \tag{5}$$

$$\varepsilon_{\mathrm{I}}(\pi) = \sum_{t=0}^{H-1} \mathbb{E}_{\mu_t^\pi}[|r - \widehat{r}| + \mathrm{span}(\widehat{V}_{t+1}^\pi)\,\mathrm{TV}(P, \widehat{P})]. \tag{6}$$

Theorem 2 (Decision reliability)**.** *For every policy and every policy comparison,*

$$\begin{aligned} J(\pi) - \widehat{J}(\pi) &= \int_{\mathcal{S}} \big(V_0^\pi - \widehat{V}_0^\pi\big)(s)\,\rho(\mathrm{d}s) \\ &= \sum_{t=0}^{H-1} \int_{\mathcal{S}\times\mathcal{A}} \Delta_t^\pi(s,a)\,\mu_t^\pi(\mathrm{d}s, \mathrm{d}a) \\ &= \sum_{t=0}^{H-1} \mathbb{E}_{\mu_t^\pi} \Delta_t^\pi. \end{aligned} \tag{7}$$

Moreover,

$$\begin{aligned} |J(\pi) - \widehat{J}(\pi)| &\le \sum_{t=0}^{H-1} \int \left(|r - \widehat{r}| + \left| \int \widehat{V}_{t+1}^\pi \,\mathrm{d}(P - \widehat{P}) \right| \right) \mathrm{d}\mu_t^\pi \\ &\le \sum_{t=0}^{H-1} \int \Big(|r - \widehat{r}| + \mathrm{span}(\widehat{V}_{t+1}^\pi)\,\mathrm{TV}(P, \widehat{P}) \Big)\,\mathrm{d}\mu_t^\pi \\ &= \varepsilon_{\mathrm{I}}(\pi). \end{aligned} \tag{8}$$

Consequently,

$$\begin{aligned} J(\pi_1) - J(\pi_0) &= \widehat{J}(\pi_1) - \widehat{J}(\pi_0) + [J(\pi_1) - \widehat{J}(\pi_1)] - [J(\pi_0) - \widehat{J}(\pi_0)] \\ &\ge \widehat{J}(\pi_1) - \widehat{J}(\pi_0) - |J(\pi_1) - \widehat{J}(\pi_1)| - |J(\pi_0) - \widehat{J}(\pi_0)| \\ &\ge \widehat{J}(\pi_1) - \widehat{J}(\pi_0) - \varepsilon_{\mathrm{I}}(\pi_1) - \varepsilon_{\mathrm{I}}(\pi_0). \end{aligned} \tag{9}$$

Corollary 1 (Planning regret)**.** *If $\widehat{J}(\widehat{\pi}) \ge \sup_{\pi\in\Pi} \widehat{J}(\pi) - \delta_p$, then*

$$J^* - J(\widehat{\pi}) \le \sup_{\pi\in\Pi} \varepsilon_{\mathrm{I}}(\pi) + \delta_p + \varepsilon_{\mathrm{I}}(\widehat{\pi}). \tag{10}$$

Appendix C gives the Bellman telescope and full regret derivation; Appendix C.4 treats shifts beyond verified occupancy.

Corollary 2 (Predictive accuracy does not certify a decision)**.** *Arbitrarily small uniform transition error can reverse world-model-greedy action rankings; total variation one can coexist with exact values for every policy.*

Appendix D.4 gives both constructions: qualification must compare prediction error with the advantage at stake.

4.2 From a certified comparison to closed-loop planning

Replanning changes a certified continuation. Fix a reference policy π_0, with $Q_t^0 = r + PV_{t+1}^{\pi_0}$ and $\widehat{Q}_t^0 = \widehat{r} + \widehat{P}\widehat{V}_{t+1}^{\pi_0}$. For candidate action kernel κ_t, let $|Q_t^0 - \widehat{Q}_t^0| \le b_t$ and define

$$d_t(s) = \int Q_t^0(s,a)(\kappa_t - \pi_{0,t})(\mathrm{d}a \mid s), \tag{11}$$

$$\widehat{d}_t(s) = \int \widehat{Q}_t^0(s,a)(\kappa_t - \pi_{0,t})(\mathrm{d}a \mid s), \tag{12}$$

$$B_t(s) = \int b_t(s,a)(\kappa_t + \pi_{0,t})(\mathrm{d}a \mid s). \tag{13}$$

Theorem 3 (Closed-loop reliability)**.** *Suppose simultaneously at all admissible states that* $\widehat{d}_t \geq S_t - \xi_t$*, where* $\xi_t \geq 0$*. Put*

$$\begin{aligned} T_t &= S_t - B_t - \xi_t, \qquad g_t = \mathbb{I}\{T_t > 0\}, \\ \nu_t &= g_t \kappa_t + (1 - g_t)\pi_{0,t}. \end{aligned} \tag{14}$$

If ρ_t^ν *is the true state law under* ν*, then*

$$\begin{aligned} J(\nu) - J(\pi_0) &= \sum_{t=0}^{H-1} \int_{\mathcal{S}} \rho_t^\nu(\mathrm{d}s) \int_{\mathcal{A}} Q_t^0(s,a)[\nu_t - \pi_{0,t}](\mathrm{d}a \mid s) \\ &= \sum_{t=0}^{H-1} \int_{\mathcal{S}} g_t(s) d_t(s) \rho_t^\nu(\mathrm{d}s) \\ &\geq \sum_{t=0}^{H-1} \int_{\mathcal{S}} g_t(s)[\widehat{d}_t(s) - B_t(s)] \rho_t^\nu(\mathrm{d}s) \\ &\geq \sum_{t=0}^{H-1} \int_{\mathcal{S}} g_t(s) T_t(s) \rho_t^\nu(\mathrm{d}s) \geq 0. \end{aligned} \tag{15}$$

Strict improvement holds if a positive-margin intervention is reached with positive probability.

Appendix C.6 proves the result under repeated replanning. The guarantee remains decision-specific: it compares local predicted advantage against continuation-value uncertainty under the state distribution induced by the gated policy. Under $|r - \widehat{r}| \leq \epsilon_r$ and $\mathrm{TV}(P, \widehat{P}) \leq \epsilon_p$, Theorem 2 gives

$$b_t = (H-t)\epsilon_r + \tfrac{1}{2}(H-t)(H-t-1)R_b\epsilon_p. \tag{16}$$

Appendix C.6 gives tighter and truncated-rollout variants; Appendix D gives the corresponding differential certificates.

5 Dual-Frontier Learning

5.1 Qualification and promotion

Let $S(x)$ estimate $\widehat{\Gamma}(x) = \widehat{J}(\pi_1) - \widehat{J}(\pi_0)$, $B(x)$ bound world-model distortion, and $\xi(x)$ bound estimation error. Define

$$T(x) = S(x) - B(x) - \xi(x), \tag{17}$$

$$\mathcal{F}_{\mathrm{W}} = \{x : S(x) > \xi(x),\ B(x) \geq S(x) - \xi(x)\}, \tag{18}$$

$$\mathcal{F}_{\mathrm{A}} = \{x : T(x) > 0\}. \tag{19}$$

$\mathcal{F}_{\mathrm{W}}$ retains promising but uncertified comparisons for verification, whereas $\mathcal{F}_{\mathrm{A}}$ contains certified improvements; $S \leq \xi$ is deferred.

Calibrated uncertainty. For $e(x) = |\widehat{\Gamma}(x) - \Gamma_{\mathcal{M}}(x)|$, fit a nonnegative score u independently of n exchangeable calibration requests and set

$$q_j = e_j - u(x_j), \qquad k = \lceil (n+1)(1-\alpha) \rceil, \qquad B(x) = [u(x) + q_{(k)}]_+. \tag{20}$$

Proposition 7 gives $\mathbb{P}\{e(x) > B(x)\} \leq \alpha$. If $\widehat{\Gamma} \geq S - \xi$ except with probability δ, then

$$\begin{aligned} &\{T(x) > 0, \Gamma_{\mathcal{M}}(x) \leq 0\} \subseteq \{e(x) > B(x)\} \cup \{\widehat{\Gamma}(x) < S(x) - \xi(x)\}, \\ &\mathbb{P}\{T(x) > 0, \Gamma_{\mathcal{M}}(x) \leq 0\} \leq \alpha + \delta. \end{aligned} \tag{21}$$

Appendix E gives the rank proof, noisy-label extension, and selection-conditional distinction.

5.2 Reusable evidence and adaptive planning

For a stationary finite task with D states, m state–action rows, shared known rewards, and n samples per row, let $\overline{P}$ be empirical and define

$$a_n = \sqrt{\frac{D\log 2 + \log(4m/\alpha)}{2n}}, \qquad \mathcal{C} = \left\{\mathcal{N} : \max_{s,a} \mathrm{TV}(P_{\mathcal{N}}, \overline{P}) \le a_n\right\}. \tag{22}$$

Appendix E.4 gives $\mathbb{P}\{\mathcal{M} \in \mathcal{C}\} \ge 1-\alpha$. Put $K = R_b H(H-1)$ and, for selected $\widehat{P}_i$,

$$u_i = \max_{s,a} \min\{1, \mathrm{TV}(\overline{P}, \widehat{P}_i) + a_n\}. \tag{23}$$

Theorem 4 (Adaptive reuse of a world certificate)**.** *Let $x_i, \widehat{P}_i$ depend on the audit and all previous observations. Conditional on their selection, assume $\mathbb{P}(F_i^c \mid \mathcal{H}_i) \le \delta_i$, where $F_i = \{\widehat{\Gamma}_i \ge S_i - \xi_i\}$. Accept only*

$$T_i = S_i - Ku_i - \xi_i > 0. \tag{24}$$

For $E = \{\mathcal{M} \in \mathcal{C}\}$ and $\mathcal{B} = \bigcup_{i\ge 1}\{T_i > 0, \Gamma_{\mathcal{M}}(x_i) \le 0\}$,

$$\begin{aligned} \mathcal{B} &\subseteq E^c \cup \bigcup_{i\ge 1} F_i^c, \\ \mathbb{P}(\mathcal{B}) &\le \mathbb{P}(E^c) + \sum_{i\ge 1} \mathbb{E}[\mathbb{P}(F_i^c \mid \mathcal{H}_i)] \le \alpha + \sum_{i\ge 1} \delta_i. \end{aligned} \tag{25}$$

The audit costs mn queries, independently of subsequent comparisons.

Appendix E.5 gives the full adaptive proof and extensions.

5.3 Progress and evidence cost

For successive accepted behaviors with certified margins $T_i > 0$, simultaneous validity gives

$$J(\pi_n) - J(\pi_0) \ge \sum_{i=0}^{n-1} T_i, \tag{26}$$

$$|\{i \in \{0,\dots,n-1\} : T_i \ge \tau\}| \le \frac{HR_b - J(\pi_0)}{\tau}, \qquad \tau > 0. \tag{27}$$

Thus a bounded objective cannot sustain a fixed positive certified margin indefinitely (Appendix F.2).

With shared rewards and $H \ge 2$, let $q_i = \max_{s,a} \mathrm{TV}(P, \widehat{P}_i)$ and $a_i = \widehat{\Gamma}_i - Kq_i$. For exact world-model evaluation, on E,

$$T_i \ge a_i - 2Ka_n, \tag{28}$$

$$n > \frac{2}{s^2}[D\log 2 + \log(4m/\alpha)] \tag{29}$$

certifies every request with $a_i/K \ge s$ simultaneously (Appendix E.6).

Conversely, distinguishing true advantages $\pm c\gamma$, $\gamma \in (0, 1/4]$, with promotion probabilities at least $1-\delta$ and at most δ requires

$$n \ge \frac{\mathrm{kl}(1-\delta, \delta)}{16\gamma^2}, \qquad 0 < \delta < \tfrac{1}{2}. \tag{30}$$

Appendices E.8 and F give the lower bound and adaptive accounting. Together, the bounds establish the inverse-square statistical price of reliable promotion with reusable evidence.

6 Experiments

Our experiments progressively test whether the theory identifies when a learned world model can be used reliably and whether Dual-Frontier converts that diagnosis into better decisions. We proceed in two stages: controlled finite-world experiments directly isolate and validate attribution, qualification, and evidence allocation under auditable conditions, after which a cross-backbone evaluation on public agent benchmarks directly instantiates the same verify-then-promote rule in modern tool-use pipelines across heterogeneous model-task settings.

Table 1: Agent–world-model evaluation on BFCL v4, API-Bank, and NexusRaven. We report Task Success (Success) and Parameter Accuracy (Acc.); Avg. is the uniform mean across benchmarks.

Models	Methods	BFCL v4		API-Bank		NexusRaven		Avg.	
		Success	Acc.	Success	Acc.	Success	Acc.	Success	Acc.
Llama-3.1-8B Instruct	Agent-only	69.12	85.31	64.96	78.76	56.29	72.36	63.46	78.81
	Always-WM	22.50	60.79	66.73	79.20	63.21	80.33	50.81	73.44
	Confidence	35.12	69.51	88.58	92.55	82.39	87.89	68.70	83.32
	Consistency	46.38	74.52	97.24	98.00	96.54	96.83	80.05	89.78
	Pessimistic	41.62	72.14	94.29	95.67	90.88	92.91	75.60	86.91
	Dual-Frontier	**78.12**	**90.30**	**98.62**	**98.83**	**98.11**	**98.01**	**91.62**	**95.71**
Qwen3-8B	Agent-only	76.38	90.78	71.26	80.00	74.21	81.31	73.95	84.03
	Always-WM	21.88	60.38	67.52	80.68	62.58	80.68	50.66	73.91
	Confidence	35.00	69.41	91.34	94.06	91.19	93.02	72.51	85.50
	Consistency	47.75	75.42	98.03	98.10	97.11	97.80	80.96	90.44
	Pessimistic	41.62	73.18	93.11	95.24	94.97	96.27	76.57	88.23
	Dual-Frontier	**79.38**	**92.20**	**98.92**	**98.64**	**97.80**	**97.96**	**92.03**	**96.27**
Frontier Models	GPT-6 Astra	82.19	89.56	84.24	85.14	87.40	91.35	84.61	88.68
	Claude Opus 5	84.47	93.96	64.72	93.35	67.62	96.85	72.27	94.72
	GLM-5.3-Flash	85.78	80.86	67.00	79.08	79.53	86.46	77.44	82.13

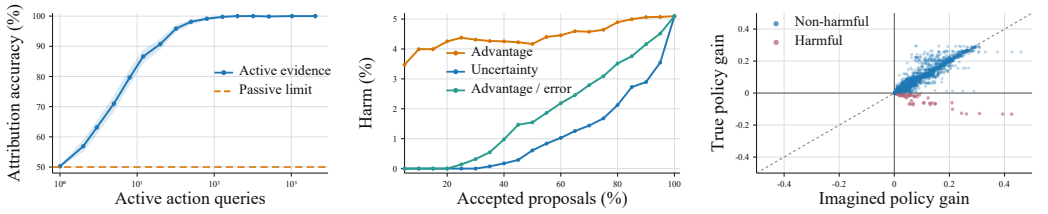

Figure 1: Controlled finite-world validation of intervention attribution, planning qualification, and imagined-update transfer.

6.1 Controlled validation with learned world models

We train action-conditioned learned world models in sparse, chain, and grid finite-horizon worlds and test held-out decisions. Given a reference action and candidate intervention, the model predicts the counterfactual consequence and the router decides whether to alter the action; exact returns are evaluation-only. All methods share the world model, reference behavior, proposals, calibration split, and evidence budget, varying only structure, horizon, margin, and evidence. Appendix G.1 gives the construction and protocol. Figure 1 shows that targeted evidence raises attribution accuracy from 50.25% to 100% and cuts effect error fivefold (0.0900 to 0.0189); over 11,520 requests, qualification reduces harmful use from 5.10% to 0.30%. Certified imagined updates preserve beneficial transfer while rejecting harmful revisions. Figure 2 yields a log–log exponent of 2.014 when varying only the true decision margin, matching $n = \Theta(\gamma^{-2})$; a shared certificate handles 200 adaptive requests with 664 versus 350,400 real queries ($527.7\times$ fewer) while keeping familywise false-promotion risk below 5%. These results validate counterfactual attribution, decision-relative qualification, inverse-margin scaling, and reusable certification. Appendix G.2 reports full curves, ablations, and per-environment results.

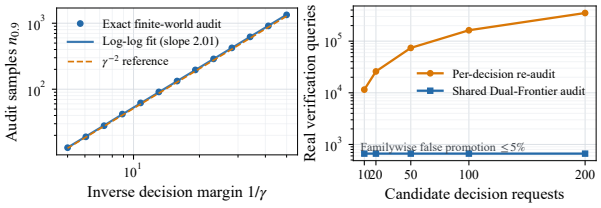

Figure 2: Evidence complexity and audit reuse in finite worlds. Left, certification cost follows the inverse-square law; right, a shared certificate amortizes verification under familywise control.

Table 2: Dual-Frontier component ablation with Llama-3.1-8B-Instruct on three benchmarks. We report Task Success (Success), Parameter Accuracy (Acc.), and uniform benchmark averages.

Models	Methods	BFCL v4		API-Bank		NexusRaven		Avg.	
		Success	Acc.	Success	Acc.	Success	Acc.	Success	Acc.
Llama-3.1-8B Instruct	Advantage-only	65.41	82.60	94.60	96.12	93.71	95.40	84.57	91.37
	w/o world-model error	69.82	85.57	96.49	97.40	95.61	96.61	87.31	93.19
	w/o estimation error	76.41	89.33	98.02	98.23	97.39	97.70	90.61	95.09
	Dual-Frontier	**78.12**	**90.30**	**98.62**	**98.83**	**98.11**	**98.01**	**91.62**	**95.71**

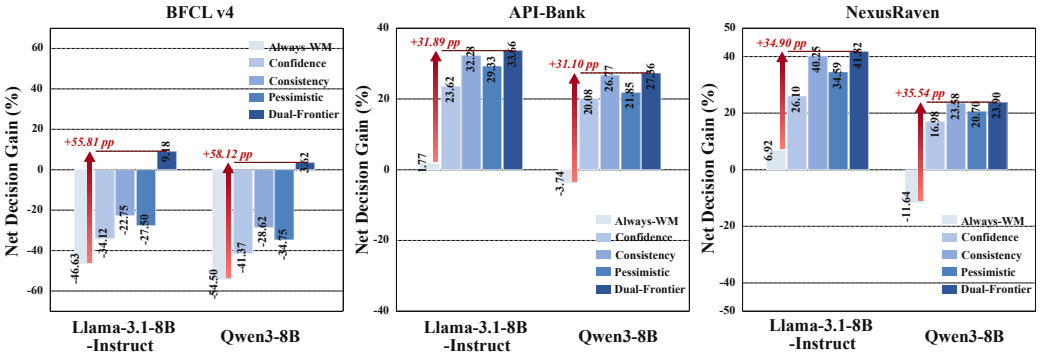

Figure 3: Net decision gain across benchmarks and backbones. Red arrows show Dual-Frontier's NDG improvement over Always-WM.

6.2 Agent-world-model evaluation across benchmarks

We next instantiate the same Dual-Frontier decision rule from controlled certification in realistic agent–world-model interaction. We fix Qwen-AgentWorld (Zuo et al., 2026) as the learned world model and evaluate Llama-3.1-8B-Instruct (Grattafiori et al., 2024) and Qwen3-8B (Yang et al., 2025) on BFCL v4 (Patil et al., 2025), API-Bank (Li et al., 2023), and NexusRaven (Srinivasan et al., 2023), covering heterogeneous tool-use requests and action schemas. For each request, the agent first emits its base action, after which Qwen-AgentWorld proposes candidate revisions together with predicted consequences. To isolate the trust decision itself, all routing rules evaluate the same fixed candidate and share identical prompts, schemas, candidate records, and inference budgets; they differ only in whether and how that proposal is admitted. This matched protocol separates the benefit of selective qualification from differences in proposal quality or generation cost. We also test frontier models (OpenAI, 2026; Anthropic, 2026; GLM-5-Team et al., 2026). Appendix H.1 details the calibrated routing rule and Appendix H.2 the metrics. Table 1 shows that Dual-Frontier is best on every benchmark–metric pair for both backbones. Against the strongest competing rule, average Success improves by 11.57/11.07 points and Acc. by 5.93/5.83 for Llama-3.1-8B-Instruct/Qwen3-8B. Table 2 further validates the frontier structure: performance recovers monotonically as the world-model and estimation-error coordinates are restored; Appendix H.3 reports the Qwen3-8B ablation. Figure 3 shows positive net gains on all three benchmarks, exceeding Always-WM by 31.10–58.12 points, while Figure 4 shows sharply lower harmful revisions and selective risk on BFCL v4 for both backbones. Together with the coverage–risk results in Appendix H.3, these results show that gains arise from selectively admitting revisions that pass the frontier checks rather than using the world model more aggressively, consistent with decision-specific reliability. Finally, under three independent world-model generation seeds with otherwise fixed settings, Table 3 shows small variation across both backbones and metrics, with mean performance close to the main held-out results.

Table 3: Robustness of Dual-Frontier across three independent world-model generation seeds. Results are mean ± standard deviation over a fixed randomly sampled benchmark subset.

Backbone	Success	Acc.
Llama-3.1-8B-Instruct	91.48 ± 0.63	95.57 ± 0.41
Qwen3-8B	91.91 ± 0.57	96.18 ± 0.35

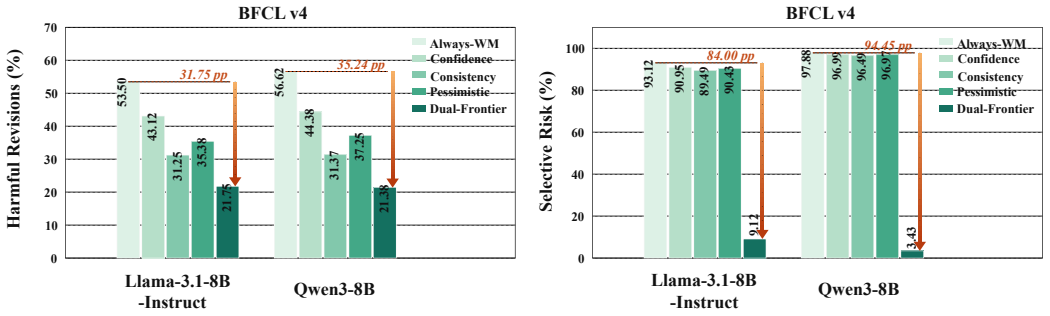

Figure 4: BFCL v4 reliability. Dual-Frontier reduces harmful revisions and selective risk across both agent backbones.

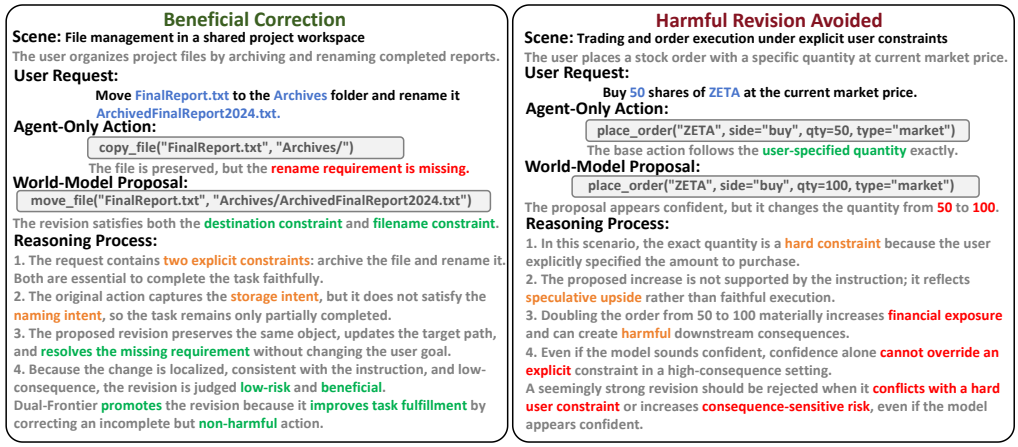

Figure 5: Representative BFCL v4 cases. Dual-Frontier promotes a reliable correction while deferring a confident but high-risk revision.

6.3 In-depth Analysis

We finally examine the mechanism behind Dual-Frontier. Figure 6 shows that beneficial revisions cluster in the positive-benefit, positive-verification region, whereas many harmful proposals remain outside the promotion frontier despite favorable predicted benefit. This indicates that predicted benefit alone is insufficient and verification is needed to screen unreliable interventions. Figure 5 provides complementary BFCL v4 cases, where Dual-Frontier promotes a supported correction but defers a confident, high-risk revision. Together, these observations align with Sections 3–5: decision-level qualification separates promising predictions from those sufficiently supported for action. Controlled finite-world experiments and public agent benchmarks consistently validate the same frontier mechanism under formal certification and realistic tool use, confirming its robustness across both settings.

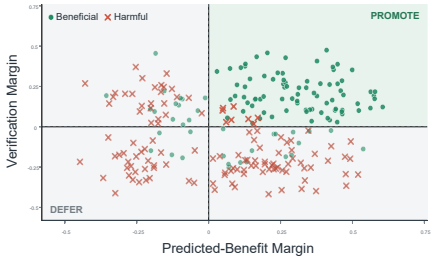

Figure 6: Frontier geometry over 240 sampled requests.

7 Conclusion

Dual-Frontier establishes a decision-specific view of world-model reliability. Passive failures cannot identify their source, motivating qualification against model and estimation uncertainty. Theory and experiments support closed-loop, reusable verification. Trust should be established at the decision boundary rather than inferred from prediction quality alone.

REFERENCES

Romina Abachi, Mohammad Ghavamzadeh, and Amir massoud Farahmand. Policy-aware model learning for policy gradient methods, 2021. URL https://arxiv.org/abs/2003.00030.

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification, 2022. URL https://arxiv.org/abs/2107.07511.

Anthropic. Introducing Claude Opus 5. https://www.anthropic.com/news/claude-opus-5, 2026.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba, Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-jepa 2: Self-supervised video models enable understanding, prediction and planning, 2025. URL https://arxiv.org/abs/2506.09985.

Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15791–15801, June 2025.

Hyungjoo Chae, Namyoung Kim, Kai Tzu iunn Ong, Minju Gwak, Gwanwoo Song, Jihoon Kim, Sunghwan Kim, Dongha Lee 0003, and Jinyoung Yeo. Web agents with world models: Learning and leveraging environment dynamics in web navigation. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=moWiYJuSGF.

Zhaorun Chen, Zhuokai Zhao, Kai Zhang, Bo Liu, Qi Qi, Yifan Wu, Tarun Kalluri, Xuefei Cao, Yuanhao Xiong, Haibo Tong, Huaxiu Yao, Hengduo Li, Jiacheng Zhu, Xian Li, Dawn Song, Bo Li, Jason E Weston, and Dat Huynh. Scaling agent learning via experience synthesis. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 121394–121420, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/c53793534833d30a40f3d18352735519-Paper-Conference.pdf.

Santiago Cifuentes. General agents contain world models, even under partial observability and stochasticity. *ArXiv*, abs/2602.03146, 2026. URL https://api.semanticscholar.org/CorpusID:285275520.

Florent Delgrange, Raphael Avalos, and Willem Röpke. Deep spi: Safe policy improvement via world models. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 788–824, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/0212b0dff7b4aa3bd8b8cc4ce24fd384-Paper-Conference.pdf.

Pierluca D'Oro*, Alberto Maria Metelli*, Andrea Tirinzoni, Matteo Papini, and Marcello Restelli. Gradient-aware model-based policy search. In *AAAI Conference on Artificial Intelligence (AAAI)*, pp. 3801–3808. 2020. URL https://doi.org/10.1609/aaai.v34i04.5791.

Tianqing Fang, Hongming Zhang, Zhisong Zhang, Kaixin Ma, Wenhao Yu, Haitao Mi, and Dong Yu. WebEvolver: Enhancing web agent self-improvement with co-evolving world model. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 8959–8975, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.454. URL https://aclanthology.org/2025.emnlp-main.454/.

Ignat Georgiev, Varun Giridhar, Nick Hansen, and Animesh Garg. Pwm: Policy learning with multi-task world models. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu

(eds.), *International Conference on Learning Representations*, volume 2025, pp. 13737–13757, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/240ea1741b205ea295721d55184ac43b-Paper-Conference.pdf.

GLM-5-Team, :, Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, Chenzheng Zhu, Congfeng Yin, Cunxiang Wang, Gengzheng Pan, Hao Zeng, Haoke Zhang, Haoran Wang, Huilong Chen, Jiajie Zhang, Jian Jiao, Jiaqi Guo, Jingsen Wang, Jingzhao Du, Jinzhu Wu, Kedong Wang, Lei Li, Lin Fan, Lucen Zhong, Mingdao Liu, Mingming Zhao, Pengfan Du, Qian Dong, Rui Lu, Shuang-Li, Shulin Cao, Song Liu, Ting Jiang, Xiaodong Chen, Xiaohan Zhang, Xuancheng Huang, Xuezhen Dong, Yabo Xu, Yao Wei, Yifan An, Yilin Niu, Yitong Zhu, Yuanhao Wen, Yukuo Cen, Yushi Bai, Zhongpei Qiao, Zihan Wang, Zikang Wang, Zilin Zhu, Ziqiang Liu, Zixuan Li, Bojie Wang, Bosi Wen, Can Huang, Changpeng Cai, Chao Yu, Chen Li, Chengwei Hu, Chenhui Zhang, Dan Zhang, Daoyan Lin, Dayong Yang, Di Wang, Ding Ai, Erle Zhu, Fangzhou Yi, Feiyu Chen, Guohong Wen, Hailong Sun, Haisha Zhao, Haiyi Hu, Hanchen Zhang, Hanrui Liu, Hanyu Zhang, Hao Peng, Hao Tai, Haobo Zhang, He Liu, Hongwei Wang, Hongxi Yan, Hongyu Ge, Huan Liu, Huanpeng Chu, Jia'ni Zhao, Jiachen Wang, Jiajing Zhao, Jiamin Ren, Jiapeng Wang, Jiaxin Zhang, Jiayi Gui, Jiayue Zhao, Jijie Li, Jing An, Jing Li, Jingwei Yuan, Jinhua Du, Jinxin Liu, Junkai Zhi, Junwen Duan, Kaiyue Zhou, Kangjian Wei, Ke Wang, Keyun Luo, Laiqiang Zhang, Leigang Sha, Liang Xu, Lindong Wu, Lintao Ding, Lu Chen, Minghao Li, Nianyi Lin, Pan Ta, Qiang Zou, Rongjun Song, Ruiqi Yang, Shangqing Tu, Shangtong Yang, Shaoxiang Wu, Shengyan Zhang, Shijie Li, Shuang Li, Shuyi Fan, Wei Qin, Wei Tian, Weining Zhang, Wenbo Yu, Wenjie Liang, Xiang Kuang, Xiangmeng Cheng, Xiangyang Li, Xiaoquan Yan, Xiaowei Hu, Xiaoying Ling, Xing Fan, Xingye Xia, Xinyuan Zhang, Xinze Zhang, Xirui Pan, Xu Zou, Xunkai Zhang, Yadi Liu, Yandong Wu, Yanfu Li, Yidong Wang, Yifan Zhu, Yijun Tan, Yilin Zhou, Yiming Pan, Ying Zhang, Yinpei Su, Yipeng Geng, Yong Yan, Yonglin Tan, Yuean Bi, Yuhan Shen, Yuhao Yang, Yujiang Li, Yunan Liu, Yunqing Wang, Yuntao Li, Yurong Wu, Yutao Zhang, Yuxi Duan, Yuxuan Zhang, Zezhen Liu, Zhengtao Jiang, Zhenhe Yan, Zheyu Zhang, Zhixiang Wei, Zhuo Chen, Zhuoer Feng, Zijun Yao, Ziwei Chai, Ziyuan Wang, Zuzhou Zhang, Bin Xu, Minlie Huang, Hongning Wang, Juanzi Li, Yuxiao Dong, and Jie Tang. Glm-5: from vibe coding to agentic engineering, 2026. URL https://arxiv.org/abs/2602.15763.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Jiacheng Guo, Ling Yang, Peter Chen, Qixin Xiao, Yinjie Wang, Xinzhe Juan, Jiahao Qiu, Ke Shen, and Mengdi Wang. Genenv: Difficulty-aligned co-evolution between llm agents and environment simulators, 2025. URL https://arxiv.org/abs/2512.19682.

David Ha and Jürgen Schmidhuber. World models. 2018. doi: 10.5281/ZENODO.1207631. URL https://zenodo.org/record/1207631.

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.

Nick Hansen, Hao Su, and Xiaolong Wang. Td-mpc2: Scalable, robust world models for continuous control. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 47376–47405, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/cf73d57b6dcda32b293df7c2d5341f49-Paper-Conference.pdf.

Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-based policy optimization. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/5faf461eff3099671ad63c6f3f094f7f-Paper.pdf.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. Swe-bench: Can language models resolve real-world github issues? In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.),

International Conference on Learning Representations, volume 2024, pp. 54107–54157, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/edac78c3e300629acfe6cbe9ca88fb84-Paper-Conference.pdf.

Minsoo Kim and Seung-won Hwang. CoEx – co-evolving world-model and exploration. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 21629–21651, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.1179. URL https://aclanthology.org/2025.findings-emnlp.1179/.

Minghao Li, Yingxiu Zhao, Bowen Yu, Feifan Song, Hangyu Li, Haiyang Yu, Zhoujun Li, Fei Huang, and Yongbin Li. API-bank: A comprehensive benchmark for tool-augmented LLMs. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 3102–3116, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.187. URL https://aclanthology.org/2023.emnlp-main.187/.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, and Jie Tang. Agentbench: Evaluating llms as agents, 08 2023.

Youwei Liu, Jian Wang, Hanlin Wang, and Wenjie Li. Comap: Co-evolving world models and agent policies for llm agents, 2026. URL https://arxiv.org/abs/2606.02372.

Zichen Liu, Guoji Fu, Chao Du, Wee Sun Lee, and Min Lin. Continual reinforcement learning by planning with online world models. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 38397–38423. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/liu25p.html.

Ning Lu, Baijiong Lin, Shengcai Liu, Jiahao Wu, Haoze Lv, Yanbin Wei, Lingting Zhu, Shengju Qian, Xin Wang, Ying-Cong Chen, Qi Wang, and Ke Tang. Policy and world modeling co-training for language agents, 2026. URL https://arxiv.org/abs/2606.02388.

OpenAI. GPT-6 Astra System Card. https://deploymentsafety.openai.com/gpt-6-astra, 2026.

Shishir G Patil, Huanzhi Mao, Fanjia Yan, Charlie Cheng-Jie Ji, Vishnu Suresh, Ion Stoica, and Joseph E. Gonzalez. The berkeley function calling leaderboard (BFCL): From tool use to agentic evaluation of large language models. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 48371–48392. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/patil25a.html.

Jonathan Richens, Tom Everitt, and David Abel. General agents need world models. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 51659–51687. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/richens25a.html.

Marc Rigter, Minqi Jiang, and Ingmar Posner. Reward-free curricula for training robust world models. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 2369–2396, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/0a2b3e9107efed3d361b29f300a903ff-Paper-Conference.pdf.

Lloyd Russell, Anthony Hu, Lorenzo Bertoni, George Fedoseev, Jamie Shotton, Elahe Arani, and Gianluca Corrado. Gaia-2: A controllable multi-view generative world model for autonomous driving, 2025. URL https://arxiv.org/abs/2503.20523.

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

Venkat Krishna Srinivasan, Zhen Dong, Banghua Zhu, Brian Yu, Hanzi Mao, Damon Mosk-Aoyama, Kurt Keutzer, Jiantao Jiao, and Jian Zhang. Nexusraven: a commercially-permissive language model for function calling. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*, 2023. URL https://openreview.net/forum?id=Md6RUrGz67.

Shenghua Wan, Le Gan, and De-Chuan Zhan. Learning to be uncertain: Pre-training world models with horizon-calibrated uncertainty. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 106107–106127, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/ad222e6e38e26e3f00d85829b9e02f43-Paper-Conference.pdf.

Hang Wang, Xin Ye, Feng Tao, Chenbin Pan, Abhirup Mallik, Burhan Yaman, Liu Ren, and Junshan Zhang. Adawm: Adaptive world model based planning for autonomous driving. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 85591–85615, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/d4c745dcbaf8ef0d7e145754e31b1516-Paper-Conference.pdf.

Zhaoyang Wang, Canwen Xu, Boyi Liu, Yite Wang, Siwei Han, Zhewei Yao, Huaxiu Yao, and Yuxiong He. Agent world model: Infinity synthetic environments for agentic reinforcement learning. 2026. URL https://arxiv.org/abs/2602.10090.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caiming Xiong, Victor Zhong, and Tao Yu. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 52040–52094. Curran Associates, Inc., 2024. doi: 10.52202/079017-1650. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/5d413e48f84dc61244b6be550f1cd8f5-Paper-Datasets_and_Benchmarks_Track.pdf.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Xuan Yao, Junyu Gao, and Changsheng Xu. Navmorph: A self-evolving world model for vision-and-language navigation in continuous environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5536–5546, October 2025.

Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. Mopo: Model-based offline policy optimization. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 14129–14142. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/a322852ce0df73e204b7e67cbbef0d0a-Paper.pdf.

Jiahan Zhang, Muqing Jiang, Nanru Dai, Taiming Lu, Arda Uzunoglu, Shunchi Zhang, Yana Wei, Jiahao Wang, Vishal Patel, Paul Liang, Daniel Khashabi, Cheng Peng, Rama Chellappa, Tianmin Shu, Alan Yuille, Yilun Du, and Jieneng Chen. World-in-world: World models in a closed-loop world. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust

(eds.), *International Conference on Learning Representations*, volume 2026, pp. 55660–55699, 2026a. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/5b4263be85820683d78675cc18d2efc7-Paper-Conference.pdf.

Xuan Zhang, Wenxuan Zhang, See-Kiong Ng, and Yang Deng. Self-evolving world models for llm agent planning, 2026b. URL https://arxiv.org/abs/2606.30639.

Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. Webarena: A realistic web environment for building autonomous agents. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 15585–15606, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/4410c0711e9154a7a2d26f9b3816d1ef-Paper-Conference.pdf.

Yuxin Zuo, Zikai Xiao, Li Sheng, Fei Huang, Jianhong Tu, Yuxuan Liu, Tianyi Tang, Xiaomeng Hu, Yang Su, Qingfeng Lan, Yantao Liu, Qin Zhu, Yinger Zhang, Bowen Yu, Haiquan Zhao, Haiyang Xu, Jianxin Yang, Jiayang Cheng, Junyang Wang, Lianghao Deng, Mingfeng Xue, Tianyi Bai, Yang Fan, Yubo Ma, Yucheng Li, Zeyu Cui, Zhihai Wang, Zhihui Xie, Zhuorui Ye, An Yang, Dayiheng Liu, Jingren Zhou, and Ning Ding. Qwen-agentworld: Language world models for general agents, 2026. URL https://arxiv.org/abs/2606.24597.

A Mathematical Interface and Notation

For reference, the paired world notation is

$$\mathcal{M}_z = (\mathcal{S}, \mathcal{A}, P_z, r_z, \rho_z, H), \qquad \widehat{\mathcal{M}}_z = (\mathcal{S}, \mathcal{A}, \widehat{P}_z, \widehat{r}_z, \rho_z, H).$$

A.1 Objects held fixed in an intervention

The task index z identifies a reference environment, not merely a generated description. Once z is fixed, both worlds share state space $\mathcal{S}$, action space $\mathcal{A}$, initial law ρ, and horizon H. Their transition kernels are $P, \widehat{P}$ and bounded rewards are $r, \widehat{r}$. All are measurable. An episode contains states and actions at times $0, \dots, H-1$; the transition following the final reward is omitted. This convention explains why transition error is summed only through $H-2$.

The kernel construction on standard Borel spaces defines a unique trajectory law. For a measurable policy π, let p^π and $\widehat{p}^\pi$ denote these two laws; μ_t^π and $\widehat{\mu}_t^\pi$ are their state–action marginals. Policies may depend on time, already included in the state. $V_t^\pi(s)$ and $\widehat{V}_t^\pi(s)$ denote expected rewards from time t onward. Returns are finite and lie in $[0, HR_b]$. Suprema over policy classes are used when maximizers are not guaranteed to exist.

A_ϕ is a measurable, possibly randomized model-to-policy operator. Returns of randomized outputs average the return of each realized policy over a fixed seed law. This is not generally the return of the pointwise mean of Markov policy kernels. Oracle replacement changes the model argument only: the operator, available actions, budget, and seed law remain fixed. An oracle that also changes the optimizer or observation interface defines a different intervention.

In the optional differential analysis, $\theta \in \Theta \subseteq \mathbb{R}^d$ parameterizes a behavior evaluated by predictive policy search. It is distinct from the fixed competence description ϕ used to define counterfactual attribution. At an update, the kernels and rewards do not depend on the differentiation variable. Models may be refitted between updates; every certificate must then be checked for the new snapshot.

A.2 Partial observability and learned representations

An observable history can serve as the state: at time t take the full observation–action history, including any observed rewards or tool outputs. The true kernel is the conditional law of the next observable history under an action, and the learned kernel predicts the same object. This is a common measurable interface on which the finite-horizon arguments apply. It does not assume access to hidden physical states or equality of internal representations.

For a latent model supporting a history-dependent policy, one route is to specify a history-space kernel, for example through a decoder. Alternatively, when the policy consumes only $f(h)$, Proposition 5 compares true history-conditional latent transitions directly with the learned latent kernel. Its uniform error includes representation aliasing; neither a decoder nor exact Markov sufficiency is then assumed. Passive latent prediction loss alone supplies neither guarantee. Likewise, a generated executable environment that defines a new task is not automatically an approximation to a reference environment. The evaluation correspondence must be specified.

A.3 Predictive decision procedures

A planner with specification ϕ maps the supplied world $\mathcal{N}$ to a behavior $\mathsf{A}_\phi(\mathcal{N})$. It may select a plan once or replan after each observation. With a fixed internal supplied world, the latter procedure defines a history-based policy. Randomized search, finite imagined rollouts and a fixed rollout budget are part of the operator; no parameter update is needed.

Attribution recomputes the operator with the oracle world and compares $J_{\mathcal{M}}(\mathsf{A}_\phi(\mathcal{M}))$ with $J_{\mathcal{M}}(\mathsf{A}_\phi(\widehat{\mathcal{M}}))$. Certification instead first constructs π_0, π_1 and evaluates these same policies in every plausible world. In particular, evaluating $J_{\mathcal{N}}(\pi_1)$ does not replace the internal model or rerun the candidate search. A realized randomized output can be conditioned on before fresh evaluation; reuse of the data that selected it requires a simultaneous certificate.

For repeated action selection, a policy-pair certificate applies to the specified complete continuation. Theorem 3 instead fixes the reference continuation in each local action comparison and controls the resulting policy through the performance-difference identity. This distinction connects a frozen predictive model to an interactive agent without assuming that a short imagined plan will actually be followed.

A.4 NOTATION LEDGER

Table 4: Notation for attribution and use-dependent reliability.

Symbol	Meaning
z, t, i, k	Task, within-episode time, sample or accepted-update index, verification round
$\mathcal{M}, \widehat{\mathcal{M}}$	True environment and its learned approximation on a common interface
$P, \widehat{P}, r, \widehat{r}$	Transition kernels and reward functions
ρ, H, R_b	Shared initial law, horizon, and common nonnegative reward bound
Π, A_ϕ	Reference policy class and fixed model-to-policy operator
$J, \widehat{J}, J^*$	True return, model return, supremal true return in Π
$\widehat{\pi}, \pi^\circ$	Deployed policy and oracle-model counterfactual
R, A, W	Observed regret, intrinsic agent regret, signed model effect
$\mu_t^\pi, \widehat{\mu}_t^\pi$	True and learned state–action occupancies
$\Delta_t^\pi, \varepsilon_{\mathrm{I}}(\pi)$	Cross-model Bellman residual and value-error certificate
ℓ_t, δ	Policy-independent error envelope and policy-TV radius; δ denotes a sampling error probability where stated
ψ_θ, G, S_t	Policy score, uniform score bound, accumulated score through t
d_t, e, D_t	Expected local transition error, its sum, and prefix coupling bound
$g, \widehat{g}, \widetilde{g}$	True gradient, exact imagined gradient, sampled imagined gradient
C_H, b_θ, ξ	Trajectory-score bound, model gradient-bias bound, sampling radius
b, h, L, η	Total gradient radius, sampled gradient norm, smoothness, step size
x, u, q_i, U_α, T	Snapshot, uncertainty score, residual, calibrated upper bound, promotion margin
$\mathcal{H}_k, \varepsilon_k, N_K$	Verification history, round risk budget, count of non-improving executed steps
x, π_0, π_1	Decision request, reference policy, candidate policy
$\Gamma_{\mathcal{M}}, \mathcal{C}, \ell_{\mathcal{C}}$	True policy contrast, world confidence set, robust contrast
$\mathcal{F}_{\mathrm{W}}, \mathcal{F}_{\mathrm{A}}$	Model-evidence frontier and certified agent frontier
S, B, ξ, T	Estimated contrast, model-error radius, estimation radius, certified margin (return units)

Local symbols are defined when first introduced. $\|\cdot\|_2$ is Euclidean norm, $\langle\cdot,\cdot\rangle$ its inner product, and the norm of a matrix is its induced operator norm. $[a]_+ = \max\{a, 0\}$. $\mathbb{E}$ and $\mathbb{P}$ denote expectation and probability under the law specified in context.

A.5 LOGICAL DEPENDENCIES

Table 5: What each guarantee requires and what it does not supply.

Result	Required information or regularity	Not implied
Theorem 1	Passive episodes of a fixed composition	Impossibility after interventions
Theorem 2	Common measurable interface; bounded rewards	An error bound from arbitrary pixel loss
Proposition 4	Uniform policy-TV control	Reliability after unrestricted policy search
Theorem 3	Simultaneous conditional value bounds; fixed reference	Guaranteed improvement from an arbitrary critic
Theorem 5	Frozen worlds; common reward; bounded policy score	Gradient fidelity for shared-parameter model updates
Theorem 6	Gradient-error ball; local smoothness; feasible step	Harmfulness of every rejected step
Proposition 7	Independent score fitting; exchangeable true error labels	Coverage conditional on selection
Proposition 8	Finite spaces; fresh generative row queries	Cheap verification in arbitrary visual domains
Theorem 4	One stationary task; uniform world confidence; valid estimation	Acceptance of all future proposals
Theorem 7	Conditional risk control after adaptive task choice	Improvement of every task when only one task is audited
Appendix F.2	Fixed objective; relative accuracy; accepted updates	A bound on waiting time or real samples
Theorem 8	Explicit no-transfer, monotone precedence response	Universal advantage over scalar curricula

B Attribution: Impossibility, Instability, and Recovery

The attribution continuum in Theorem 1 is

$$(A, W) = (\lambda, c - \lambda).$$

B.1 Complete fixed-operator construction

Proof of Theorem 1. Let $\mathcal{S} = \{s, g, b\}$ and $\mathcal{A} = \{a_0, a_1, a_2\}$. Start at s, use horizon two, and set $r(s, a) = r(b, a) = 0$ and $r(g, a) = c$ for every action. The states g, b are absorbing. For $\lambda \in [0, c]$, define

$$P_\lambda(g \mid s, a_0) = 0, \quad P_\lambda(g \mid s, a_1) = 1 - \lambda/c, \quad P_\lambda(g \mid s, a_2) = 1.$$

The remaining probability goes to b. The supplied world model $\widehat{P}$ sends every action at s to b and agrees with the true kernels elsewhere. The rewards are known and shared.

The operator evaluates only actions a_0, a_1 using its supplied world model and chooses the higher-return action, breaking ties toward a_0. Its terminal action is fixed arbitrarily. This same operator is used for every λ; it is not hidden from the observer. Its restricted search is a concrete competence limitation, while the reference class contains the policy choosing a_2.

Under $\widehat{P}$, deployment always chooses a_0. Under P_λ, this produces $(s, a_0, 0, b, a_0, 0)$ with probability one. For $\lambda < c$, oracle replacement chooses a_1 and achieves $c - \lambda$; at $\lambda = c$ the tie-breaking rule chooses a_0 and still achieves $c - \lambda = 0$. The reference value is always c. Hence $R = c, A = \lambda, W = c - \lambda$.

Let $\mathcal{O}$ be generated by any number of deployed episodes, the supplied world model, the known operator, and independent randomization used by an estimator. All these objects have the same law

for every λ. For any $\mathcal{O}$-measurable real estimate X,

$$\begin{aligned}\sup_{\lambda\in[0,c]} \mathbb{E}_\lambda|X-\lambda| &\geq \max\{\mathbb{E}_0|X|, \mathbb{E}_c|X-c|\} \\ &\geq \frac{1}{2}\big(\mathbb{E}_0|X| + \mathbb{E}_c|X-c|\big) \\ &= \frac{1}{2}\int \big(|x| + |x-c|\big)\, \mathsf{Q}(\mathrm{d}x) \\ &\geq \frac{c}{2},\end{aligned}$$

where Q is the common law of X and the last step is the triangle inequality. Conversely, $X = c/2$ obeys

$$\sup_{\lambda\in[0,c]} |X-\lambda| = c/2,$$

so the lower bound is exact. The endpoint classification argument follows from the same common-law experiment: the two endpoint labels have equal prior probability and identical observations, hence Bayes and minimax error are both $1/2$. This proves Theorem 1 for arbitrary passive sample size, including infinite passive repetition.

Notice that the supplied world model is exact on the deployed occupancy: its error there is zero. The missing information is about untried actions. Direct observation of the world-model parameters does not remove this obstruction. Repeated controlled trials of a_1 identify its success probability asymptotically; a single trial need not determine the attribution. The construction isolates why passive return and on-policy fit cannot alone justify an attribution or a certificate for new interventions.

B.2 Finite-sample Monte Carlo planning

The obstruction persists when action values are estimated from finitely many imagined rollouts. Use the same states, reward, supplied world model, and reference class as above, but let $P(g \mid s, a_1) = p$ range over $[0,1]$. Fix a positive integer N. The planner samples N independent supplied-world episodes starting with each of a_0, a_1, compares the sample return means, and outputs the corresponding deterministic initial action. Ties select a_0; terminal actions are fixed.

Under the supplied world model all $2N$ imagined episodes return zero, so the complete search transcript and output policy are identical for all true worlds. Subsequent real deployment of a_0 is also identical. Under oracle replacement, a_0 still always returns zero, whereas a_1 is selected if and only if at least one of its N imagined episodes succeeds. Its selection probability is $1-(1-p)^N$. Evaluation uses a new episode independent of search, giving

$$J(\pi^\circ) = c f_N(p), \qquad f_N(p) = p[1-(1-p)^N].$$

The function f_N is continuous, has $f_N(0) = 0$, $f_N(1) = 1$, and for $p \in (0,1)$,

$$f_N'(p) = 1-(1-p)^N + Np(1-p)^{N-1} > 0.$$

For every $\lambda \in [0,c]$, there is therefore a unique p with $c f_N(p) = c - \lambda$. Since $J^* = c$ and $J(\widehat{\pi}) = 0$, the same $(A, W) = (\lambda, c-\lambda)$ continuum results. The common-law estimation and classification proofs apply even when the observer also sees every imagined search record.

The planner, rollout budget, and supplied world model are fixed throughout this family. The oracle does not increase compute or give a better optimizer; it changes only the predictive law used for look-ahead. This finite-sample construction requires no change of agent parameters.

B.3 A fixed true-world variant

A different construction keeps the true one-step MDP fixed but varies the agent operator. Use actions with rewards $c, 0$ and one supplied world model distinct from the truth. For each λ, let the operator choose the zero-reward action under the supplied world model and the good action with probability $1-\lambda/c$ under the true model. This gives the same decomposition and minimax bounds already in a one-state MDP. Theorem 1 is stronger with respect to what the observer knows about the operator: there the operator and supplied world model are fixed.

B.4 Predictive reliability does not identify a black-box operator's model effect

Proposition 1 (No universal modulus for failure ownership)**.** *There exist one fixed agent operator and true one-step world such that, for every sufficiently small $\epsilon > 0$, a supplied world model satisfies $\sup_{\pi\in\Pi} |J(\pi) - \widehat{J}(\pi)| \le \epsilon$ but $W = c$ for a constant $c > 0$ independent of ϵ.*

Proof. Use three actions with true rewards $0, 0, c$. The model changes only the first reward from zero to ϵ, with $0 < \epsilon \le R_b$. Define the operator to choose the second action if the supplied first reward is positive, and the third action otherwise. This is one fixed measurable operator. It achieves zero in deployment and c after oracle replacement, so $W = c$. The value of any policy changes by at most ϵ. Thus no function $\omega(\epsilon) \to 0$ can universally bound $|W|$ for arbitrary operators.

A continuity or optimization-residual assumption on the operator can exclude this example, but it cannot be omitted. The value and update comparisons in the main text avoid that extra assumption by certifying the policies actually compared. Their validity is not a claim that they estimate A or W.

The sign of W is equally unrestricted. Reverse the two branches of the operator in the preceding example. The wrong model now induces the good action and oracle replacement the bad action, giving $W = -c$. The decomposition remains an exact signed identity.

B.5 Finite-sample attribution with an oracle intervention

Proposition 2 (Interventional recovery)**.** *Fix $\widehat{\pi}, \pi^\circ$ and a reference policy π_r before evaluation, with $0 \le J^* - J(\pi_r) \le \omega$. Use m independent episodes of each policy and write their sample means as $\overline{J}, \overline{J}^\circ, \overline{J}_r$. For $\delta \in (0, 1)$, put*

$$t = HR_b\sqrt{\frac{\log(6/\delta)}{2m}}, \qquad \widehat{A} = \overline{J}_r - \overline{J}^\circ, \qquad \widehat{W} = \overline{J}^\circ - \overline{J}.$$

With probability at least $1 - \delta$,

$$|\widehat{A} - A| \le 2t + \omega, \qquad |\widehat{W} - W| \le 2t.$$

Proof. A bounded return has range HR_b. Hoeffding's inequality makes each sample mean accurate to t except with probability $\delta/3$. A union bound yields simultaneous accuracy. Each difference has sampling error at most $2t$; only the agent-regret estimate has the additional reference-policy error ω. Cross-policy independence is unnecessary, so paired random numbers are allowed if each policy's episodes remain independent.

This result requires the oracle-model counterfactual policy, not merely more episodes of deployment. It quantifies evaluation after the identifying intervention; it does not remove the intervention's cost.

C Decision Reliability: Full Derivations

Theorem 2 also controls planning regret. If $\widehat{J}(\widehat{\pi}) \ge \sup_{\pi\in\Pi} \widehat{J}(\pi) - \delta_p$, then the planning-regret inequality in Corollary 1 follows. The proof below takes a supremum and does not require an optimal policy to exist.

C.1 Robust policy contrasts and the order of evaluation

Fix a use request x, including the two output policies, and a nonempty class $\mathcal{C}$ of worlds sharing the evaluation interface. Bounded rewards imply $-HR_b \le \Gamma_{\mathcal{N}}(x) \le HR_b$. Let $\underline{\ell}(x)$ be any computable lower bound on $\ell_{\mathcal{C}}(x)$. By the definition of infimum,

$$\mathcal{M} \in \mathcal{C} \quad \Longrightarrow \quad \Gamma_{\mathcal{M}}(x) \ge \ell_{\mathcal{C}}(x) \ge \underline{\ell}(x).$$

On an event where $\mathcal{M} \in \mathcal{C}$, this deterministic implication holds for all requests at once, including any selected as a function of $\mathcal{C}$. No union bound over policies or use requests is required. For random

sets and selectors we assume the displayed infima and events are measurable; the explicit finite-state bounds used in the paper are measurable functions of finite arrays.

A uniform improvement guarantee with positive constant a exists if and only if $\ell_{\mathcal{C}}(x) > 0$: one direction takes infima, and the other chooses $a = \ell_{\mathcal{C}}(x)$. This equivalence is for the information represented by $\mathcal{C}$, not for the unknown true world alone. A computable lower bound may fail even when the exact infimum is positive.

If $\mathcal{C}$ is compact and $\Gamma_{\mathcal{N}}(x)$ is continuous in $\mathcal{N}$, its image is a nonempty compact subset of $\mathbb{R}$ and contains its infimum. Thus $\ell_{\mathcal{C}}(x) \leq 0$ exhibits a plausible world where the comparison is non-improving. Without attainment, $\ell_{\mathcal{C}} = 0$ may coexist with strictly positive contrast in every world, as for contrasts $1/j$, $j \geq 1$. It still precludes a uniform positive margin.

For finite spaces and finite horizon, the expectation of a fixed history-based policy is a finite sum of products of transition probabilities and rewards. It is continuous in these arrays. Closed row-confidence sets in the probability simplices are compact and nonempty; the empirical kernel is feasible. These observations justify attainment for the concrete verifier. Exact contrast minimization can remain computationally difficult, which is why the main text supplies tractable conservative bounds instead of assuming access to an exact robust optimizer.

A request compares a proposed behavior with an explicit reference behavior. The certificate does not establish global optimality; the planning bound controls suboptimality separately when an optimization residual is available.

C.2 Bellman operators and the telescoping identity

Proof of Theorem 2 and Corollary 1. For a policy π, define $r^\pi(s) = \int r(s,a)\pi(\mathrm{d}a \mid s)$ and

$$(P^\pi f)(s) = \int_{\mathcal{A}} \int_{\mathcal{S}} f(s') P(\mathrm{d}s' \mid s, a) \pi(\mathrm{d}a \mid s).$$

Time dependence is understood through the augmented state. Both Bellman recursions hold for bounded measurable values: $V_t^\pi = r^\pi + P^\pi V_{t+1}^\pi$ and $\widehat{V}_t^\pi = \widehat{r}^\pi + \widehat{P}^\pi \widehat{V}_{t+1}^\pi$. Subtracting gives the exact operator identity

$$\begin{aligned} V_t^\pi - \widehat{V}_t^\pi &= r^\pi - \widehat{r}^\pi + P^\pi V_{t+1}^\pi - \widehat{P}^\pi \widehat{V}_{t+1}^\pi \\ &= (r - \widehat{r})^\pi + (P^\pi - \widehat{P}^\pi)\widehat{V}_{t+1}^\pi + P^\pi (V_{t+1}^\pi - \widehat{V}_{t+1}^\pi). \end{aligned}$$

Let ρ_t be the true state law, $d_t = V_t^\pi - \widehat{V}_t^\pi$, and $b_t = \int \Delta_t^\pi(\cdot, a)\pi(\mathrm{d}a \mid \cdot)$. These local symbols are used only in this proof. Since $\rho_{t+1} = \rho_t P^\pi$ and $d_H = 0$, the operator recursion above yields

$$\begin{aligned} \int d_t \, \mathrm{d}\rho_t &= \int b_t \, \mathrm{d}\rho_t + \int P^\pi d_{t+1} \, \mathrm{d}\rho_t \\ &= \mathbb{E}_{\mu_t^\pi} \Delta_t^\pi + \int d_{t+1} \, \mathrm{d}\rho_{t+1}, \\ \int d_0 \, \mathrm{d}\rho_0 &= \sum_{t=0}^{H-1} \mathbb{E}_{\mu_t^\pi} \Delta_t^\pi + \int d_H \, \mathrm{d}\rho_H \\ &= \sum_{t=0}^{H-1} \mathbb{E}_{\mu_t^\pi} \Delta_t^\pi. \end{aligned}$$

Because the initial law is shared, the left-hand side is exactly $J(\pi) - \widehat{J}(\pi)$.

For finite signed measure $P - Q$ of total mass zero and bounded f, let $a = (\sup f + \inf f)/2$. Then

$$|(P - Q)f| = |(P - Q)(f - a)| \leq 2\,\mathrm{TV}(P, Q)\|f - a\|_\infty = \mathrm{span}(f)\,\mathrm{TV}(P, Q).$$

Applying this inequality to the residual gives

$$\begin{aligned}|J(\pi) - \widehat{J}(\pi)| &= \left|\sum_{t=0}^{H-1} \int \Delta_t^\pi \, \mathrm{d}\mu_t^\pi\right| \\ &\le \sum_{t=0}^{H-1} \int \left(|r - \widehat{r}| + \left|\int \widehat{V}_{t+1}^\pi \, \mathrm{d}(P - \widehat{P})\right|\right) \mathrm{d}\mu_t^\pi \\ &\le \sum_{t=0}^{H-1} \int \left(|r - \widehat{r}| + \operatorname{span}(\widehat{V}_{t+1}^\pi) \operatorname{TV}(P, \widehat{P})\right) \mathrm{d}\mu_t^\pi \\ &= \varepsilon_{\mathrm{I}}(\pi).\end{aligned}$$

No common support of the transition kernels is required.

For the two-policy comparison, let $e_\pi = J(\pi) - \widehat{J}(\pi)$. Then

$$\begin{aligned}J(\pi') - J(\pi) &= \widehat{J}(\pi') - \widehat{J}(\pi) + e_{\pi'} - e_\pi \\ &\ge \widehat{J}(\pi') - \widehat{J}(\pi) - |e_{\pi'}| - |e_\pi| \\ &\ge \widehat{J}(\pi') - \widehat{J}(\pi) - \varepsilon_{\mathrm{I}}(\pi') - \varepsilon_{\mathrm{I}}(\pi).\end{aligned}$$

The symmetric argument gives the corresponding upper bound. If an optimal π^* exists, the planning bound sharpens to

$$J(\pi^*) - J(\widehat{\pi}) \le \delta_p + \varepsilon_{\mathrm{I}}(\pi^*) + \varepsilon_{\mathrm{I}}(\widehat{\pi}).$$

If not, apply this inequality to a sequence approaching J^*, retaining the uniform supremum used in the main theorem.

C.3 Uniform bounds, equality examples, and horizon scaling

If $|r - \widehat{r}| \le \epsilon_r$ and $\operatorname{TV}(P, \widehat{P}) \le \epsilon_p$ uniformly, then $0 \le \widehat{V}_{t+1}^\pi \le (H - t - 1)R_b$ implies

$$|J(\pi) - \widehat{J}(\pi)| \le H\epsilon_r + \frac{H(H-1)}{2} R_b \epsilon_p.$$

Both model rewards and true rewards must obey the stated reward bound.

Proposition 3 (Two endpoints are necessary)**.** *For every $\beta \in (0, R_b/2]$, the right-hand side of the endpoint inequality above is attained with $\delta_p = 0$.*

Proof. Take a one-step two-action task with true rewards $2\beta, 0$ and model rewards β, β. Let the world-model optimizer break the tie toward the second action. Each endpoint value error is β, and the true regret is 2β. Thus one endpoint error cannot simply be removed from a general planning comparison.

The transition coefficient in the uniform value bound above is first-order sharp. Consider a deterministic true chain with zero initial reward and reward R_b at each later nonabsorbing state. The model enters a zero-reward absorbing state with independent probability ϵ_p at each transition. The actual discrepancy is

$$R_b \sum_{t=1}^{H-1} \left[1 - (1 - \epsilon_p)^t\right] = \frac{H(H-1)}{2} R_b \epsilon_p + O(H^3 \epsilon_p^2)$$

as $\epsilon_p \downarrow 0$ for fixed H. A uniform reward offset attains the linear reward coefficient. These examples establish value-bound scaling, not minimax sharpness of the gradient horizon exponent.

C.4 Transport to a new policy

Define the policy-independent envelope $\ell_t = |r - \widehat{r}| + (H - t - 1)R_b \operatorname{TV}(P, \widehat{P})$. We prove the transport inequality the transport inequality above.

Proposition 4 (Verified-policy transport). *If* $\sup_s \mathrm{TV}(\pi(\cdot \mid s), \pi_0(\cdot \mid s)) \le \delta$, *then*

$$\varepsilon_{\mathrm{I}}(\pi) \le \sum_{t=0}^{H-1} \mathbb{E}_{\mu_t^{\pi_0}} \ell_t + R_b \sum_{t=0}^{H-1} (H-t) \min\{1, (t+1)\delta\}.$$

Proof. Couple identical initial states. Whenever states agree, use a maximal coupling of the action kernels; conditional action disagreement has probability at most δ. If actions also agree, sample the identical true transition synchronously. After disagreement, any coupling preserving both marginal dynamics suffices.

The event that state–action pairs disagree by time t requires at least one of $t+1$ action disagreements. A union bound, or a first-disagreement decomposition, gives

$$\mathrm{TV}(\mu_t^{\pi}, \mu_t^{\pi_0}) \le \min\{1, (t+1)\delta\}.$$

For $0 \le f \le B$, the signed-measure argument above gives $\mathbb{E}_\mu f - \mathbb{E}_\nu f \le B\,\mathrm{TV}(\mu, \nu)$. Apply it to ℓ_t, whose range lies in $[0, (H-t)R_b]$. Since the integrand defining $\varepsilon_{\mathrm{I}}(\pi)$ is bounded by ℓ_t, summing proves the transport inequality above. Finally,

$$\sum_{t=0}^{H-1} (H-t)(t+1) = \frac{H(H+1)(H+2)}{6}.$$

A coverage assumption provides another route. If $\mu_t^{\pi} \ll \mu_t^{\pi_0}$ with Radon–Nikodym derivative bounded by c_t, then nonnegativity gives

$$\varepsilon_{\mathrm{I}}(\pi) \le \sum_{t=0}^{H-1} c_t \mathbb{E}_{\mu_t^{\pi_0}} \ell_t.$$

This is a stated density-ratio assumption, not something guaranteed by low passive prediction loss. If a candidate policy visits an action absent from the validation support, neither this bound nor unrestricted reuse of an on-policy certificate is justified.

C.5 Wasserstein and discounted variants

Suppose $(\mathcal{S}, d)$ is Polish, both next-state kernels have finite first moments, and $\widehat{V}_{t+1}^{\pi}$ is L_{t+1}-Lipschitz. The Kantorovich–Rubinstein dual formula gives

$$|J(\pi) - \widehat{J}(\pi)| \le \sum_{t=0}^{H-1} \mathbb{E}_{\mu_t^{\pi}} \left[|r - \widehat{r}| + L_{t+1} W_1(P, \widehat{P}) \right].$$

Here W_1 is the first Wasserstein distance for metric d. The proof substitutes $|(P - \widehat{P})\widehat{V}_{t+1}^{\pi}| \le L_{t+1} W_1(P, \widehat{P})$ into the exact identity. The additional Lipschitz assumption is indispensable; a small state-space metric error does not control an arbitrary discontinuous continuation value.

For a stationary discounted problem with $\gamma \in (0,1)$, define $J^{\gamma}(\pi) = \mathbb{E} \sum_{t \ge 0} \gamma^t r_t$ and $\mu_{\gamma}^{\pi} = (1-\gamma) \sum_{t \ge 0} \gamma^t \mu_t^{\pi}$. The bounded Bellman resolvent yields

$$J^{\gamma}(\pi) - \widehat{J}^{\gamma}(\pi) = \frac{1}{1-\gamma} \mathbb{E}_{\mu_{\gamma}^{\pi}} [r - \widehat{r} + \gamma (P - \widehat{P}) \widehat{V}^{\pi}].$$

To see this, iterate $V - \widehat{V} = (r - \widehat{r})^{\pi} + \gamma(P^{\pi} - \widehat{P}^{\pi})\widehat{V} + \gamma P^{\pi}(V - \widehat{V})$. The remaining term after n iterations has sup norm at most $2\gamma^n R_b/(1-\gamma)$ and vanishes. Hence uniform reward and transition bounds imply

$$|J^{\gamma}(\pi) - \widehat{J}^{\gamma}(\pi)| \le \frac{\epsilon_r}{1-\gamma} + \frac{\gamma R_b \epsilon_p}{(1-\gamma)^2}.$$

This is a value extension only; the main finite-horizon statistical constants are not reused unchanged in infinite horizon.

C.6 CONDITIONAL CERTIFICATES AND CLOSED-LOOP COMPOSITION

Proof of Theorem 3. Fix a reference continuation π_0. For every admissible history state s at time t and first action a, let π^a choose a first and follow π_0 thereafter. This is a policy on the residual horizon $H-t$ with initial law concentrated at s. Applying the residual identity to this conditional problem gives

$$Q_t^0(s,a) - \widehat{Q}_t^0(s,a) = \sum_{j=t}^{H-1} \mathbb{E}_{P,\pi^a}\left[\Delta_j^{\pi^a}(s_j,a_j) \mid s_t = s\right].$$

Here $\Delta_j^{\pi^a}$ uses the learned continuation value of that same policy. Thus a valid conditional envelope is

$$b_t(s,a) = \sum_{j=t}^{H-1} \mathbb{E}_{P,\pi^a}\left[|r-\widehat{r}| + \operatorname{span}(\widehat{V}_{j+1}^{\pi_0})\,\mathrm{TV}(P,\widehat{P}) \mid s_t = s\right].$$

It is an analytical quantity unless its components are certified. Uniform row errors give

$$b_t(s,a) \le \sum_{j=t}^{H-1}[\epsilon_r + (H-j-1)R_b\epsilon_p] = (H-t)\epsilon_r + \frac{(H-t)(H-t-1)}{2}R_b\epsilon_p.$$

This proves (16). Finite-state audits bound all these conditional problems at once, including histories that a deployed agent has not yet visited.

Contrast-sensitive refinement. Let $\lambda_t = \kappa_t(\cdot \mid s) - \pi_{0,t}(\cdot \mid s)$, a signed measure of mass zero, and let $|\lambda_t|$ denote its total-variation measure. Then

$$|d_t - \widehat{d}_t| = \left|\int (Q_t^0 - \widehat{Q}_t^0)\,\mathrm{d}\lambda_t\right| \le \int b_t\,\mathrm{d}|\lambda_t| \le \int b_t\,\mathrm{d}(\kappa_t + \pi_{0,t}).$$

The first upper bound is the exact support function of the pointwise uncertainty class:

$$\sup_{|f(a)|\le b_t(s,a)} \left|\int f\,\mathrm{d}\lambda_t\right| = \int b_t\,\mathrm{d}|\lambda_t|.$$

To prove equality, take $f = b_t\,\mathrm{d}\lambda_t/\mathrm{d}|\lambda_t|$ on the support of $|\lambda_t|$. The Radon–Nikodym derivative is $+1$ or -1 almost everywhere, so this choice is admissible and realizes the integral. Sharpness is for the stated value-error class, not a claim that every extremizer is a realizable MDP. For uniform $b_t = b$ the refined radius is $2b\ \mathrm{TV}(\kappa_t, \pi_{0,t})$, and it vanishes when the proposed action law is unchanged. Either radius may be used in Theorem 3.

The performance-difference identity. Let $r_t = r(s_t,a_t)$ and write $\mathbb{E}_\nu$ for expectation under the true trajectory law of the gated policy. By conditional expectation and the reference Bellman identity,

$$\mathbb{E}_\nu[r_t + V_{t+1}^{\pi_0}(s_{t+1}) \mid s_t] = \int Q_t^0(s_t,a)\nu_t(\mathrm{d}a \mid s_t),$$

$$V_t^{\pi_0}(s_t) = \int Q_t^0(s_t,a)\pi_{0,t}(\mathrm{d}a \mid s_t),$$

$$\mathbb{E}_\nu[r_t + V_{t+1}^{\pi_0}(s_{t+1}) - V_t^{\pi_0}(s_t)] = \mathbb{E}_{\rho_t^\nu}[g_t d_t].$$

Summing the left side cancels every intermediate value:

$$\begin{aligned} J(\nu) - J(\pi_0) &= \sum_{t=0}^{H-1} \mathbb{E}_\nu[r_t + V_{t+1}^{\pi_0}(s_{t+1}) - V_t^{\pi_0}(s_t)] \\ &= \sum_{t=0}^{H-1} \int_{\mathcal{S}} g_t(s) d_t(s)\,\rho_t^\nu(\mathrm{d}s) \\ &\ge \sum_{t=0}^{H-1} \int_{\mathcal{S}} g_t(s)[\widehat{d}_t(s) - B_t(s)]\,\rho_t^\nu(\mathrm{d}s) \\ &\ge \sum_{t=0}^{H-1} \int_{\mathcal{S}} g_t(s) T_t(s)\,\rho_t^\nu(\mathrm{d}s) \ge 0. \end{aligned}$$

The first equality uses $V_H^{\pi_0} = 0$ and $\mathbb{E}_\rho V_0^{\pi_0} = J(\pi_0)$; the inequalities use $|d_t - \widehat{d}_t| \le B_t$, $\widehat{d}_t \ge S_t - \xi_t$, and $g_t = \mathbb{I}\{T_t > 0\}$. A finite sum of nonnegative integrable terms is strictly positive if one term is positive on an event of positive probability. The result guarantees expected improvement, not samplewise dominance of realized rewards.

Short rollouts and terminal estimates. Suppose a conditional model rollout stops after $h \in \{1, \dots, H-t\}$ steps and bootstraps with a measurable function $\widetilde{V}$ satisfying $\|\widetilde{V} - \widehat{V}_{t+h}^{\pi_0}\|_\infty \le \omega$. Let $\widetilde{Q}_t^0$ be the resulting expected truncated return. The tower property in the learned world gives

$$|\widetilde{Q}_t^0 - \widehat{Q}_t^0| \le \omega, \qquad |\widetilde{Q}_t^0 - Q_t^0| \le b_t + \omega.$$

Use $b_t + \omega$ in (13) and add the separate Monte Carlo estimation radius. Without a terminal-value error bound, an arbitrary score for one predicted observation is not a continuation-value certificate. The same applies to truncated language-model reasoning scores and learned critics.

D Differential Reliability of Predictive Policy Search

The parameter θ indexes behaviors evaluated through one fixed learned world model. The following results are optional local sufficient conditions for the same policy contrast used in the main text. They require differentiable policies and are not assumed for discrete action selection or frozen language-agent inference.

D.1 Differential certificates for predictive policy search

Assume shared rewards and freeze both worlds while differentiating. Policy densities on an open $\Theta \subseteq \mathbb{R}^d$ have common support, are continuously differentiable, and have score $\psi_\theta = \nabla_\theta \log \pi_\theta$ with $\|\psi_\theta\|_2 \le G$. Let $d_t = \mathbb{E}_{\mu_t^{\pi_\theta}} \mathrm{TV}(P, \widehat{P})$, $D_t = \min\{1, \sum_{j<t} d_j\}$, $e = \sum_{t=0}^{H-2} d_t$, and $C_H = GR_bH(H+1)/2$. Empty sums are zero.

Theorem 5 (Gradient reliability)**.** *For $g = \nabla_\theta J(\pi_\theta)$ and $\widehat{g} = \nabla_\theta \widehat{J}(\pi_\theta)$,*

$$\|g - \widehat{g}\|_2 \le 2GR_b \sum_{t=0}^{H-1} (t+1) D_t =: b_\theta \le 2C_H \min\{1, e\}.$$

If $b_\theta < \|\widehat{g}\|_2$, then $\langle g, \widehat{g}\rangle \ge \|\widehat{g}\|_2(\|\widehat{g}\|_2 - b_\theta) > 0$.

Proof of Theorem 5. Let $Q_t, \widehat{Q}_t$ denote the true and learned laws of the prefix through (s_t, a_t), and set $S_t = \sum_{j=0}^t \psi_\theta(s_j, a_j)$. Differentiation under the trajectory integral and score centering give

$$g = \sum_{t=0}^{H-1} \mathbb{E}_P[S_t r(s_t, a_t)], \qquad \widehat{g} = \sum_{t=0}^{H-1} \mathbb{E}_{\widehat{P}}[S_t r(s_t, a_t)].$$

Sequential coupling gives $\mathrm{TV}(Q_t, \widehat{Q}_t) \le D_t$. Thus

$$\|g - \widehat{g}\|_2 \le \sum_{t=0}^{H-1} 2\|S_t r(s_t, a_t)\|_\infty \mathrm{TV}(Q_t, \widehat{Q}_t) \le b_\theta,$$

where $\|F\|_\infty = \sup \|F\|_2$ for a vector-valued function. Cauchy–Schwarz and the gradient-ball geometry yield, for $b_\theta < \|\widehat{g}\|_2$,

$$\langle g, \widehat{g}\rangle \ge \|\widehat{g}\|_2(\|\widehat{g}\|_2 - b_\theta) > 0, \qquad \cos(g, \widehat{g}) \ge \sqrt{1 - \frac{b_\theta^2}{\|\widehat{g}\|_2^2}}.$$

Here $\cos(g, \widehat{g}) = \langle g, \widehat{g}\rangle / (\|g\|_2 \|\widehat{g}\|_2)$. The angular bound is attained in dimension at least two within the gradient-error ball. The following subsections supply the domination, coupling, and extremal calculations; this is geometric sharpness, not MDP minimax optimality.

Theorem 6 (Local improvement certificate)**.** *Suppose $\|\widetilde{g} - \widehat{g}\|_2 \le \xi$ and J is L-smooth along a feasible step $\theta^+ = \theta + \eta\widetilde{g}$, with $L, \eta > 0$. Set $b = b_\theta + \xi$ and $h = \|\widetilde{g}\|_2$. Then*

$$J(\pi_{\theta^+}) - J(\pi_\theta) \ge \eta h[(1 - L\eta/2)h - b].$$

If $h > b$, the feasible choice $\eta = (h-b)/(Lh)$ guarantees $(h-b)^2/(2L)$ improvement. If $h \le b$, no direction has positive worst-case first-order gain over the gradient ball.

Proof of Theorem 6. By the triangle inequality, $\|g - \widetilde{g}\|_2 \le b$. Smoothness therefore gives

$$J(\pi_{\theta^+}) - J(\pi_\theta) \ge \eta\langle g, \widetilde{g}\rangle - \tfrac{L}{2}\eta^2 h^2 \ge \eta h(h-b) - \tfrac{L}{2}\eta^2 h^2.$$

For a feasible displacement v, the exact supporting lower model is

$$\inf_{\|q-\widetilde{g}\|_2 \le b} \left\{ \langle q, v\rangle - \frac{L}{2}\|v\|_2^2 \right\} = \langle \widetilde{g}, v\rangle - b\|v\|_2 - \frac{L}{2}\|v\|_2^2.$$

For $v \neq 0$, equality holds at $q = \widetilde{g} - bv/\|v\|_2$. At length $t = \|v\|_2$, the maximum is $(h-b)t - Lt^2/2$, attained by alignment with $\widetilde{g}$ when $h > 0$. Its maximizing length is $[h-b]_+/L$, where $[x]_+ = \max\{x, 0\}$. If $h \le b$, zero belongs to the gradient ball. Quadratic objectives attain the lower model; sharpness is relative to this local uncertainty class. Appendix D.8 proves sharpness and the general-displacement certificate; Appendix D.9 states the requirements for practical optimizers.

We record the regularity and prefix constants used in this section:

$$\psi_\theta(s,a) = \nabla_\theta \log \pi_\theta(a \mid s), \qquad \|\psi_\theta(s,a)\|_2 \le G.$$

$$d_t = \mathbb{E}_{\mu_t^{\pi_\theta}} \mathrm{TV}(P, \widehat{P}), \qquad e = \sum_{t=0}^{H-2} d_t,$$

$$D_t = \min\left\{1, \sum_{j<t} d_j\right\}, \quad C_H = \frac{GR_bH(H+1)}{2},$$

With $S_t = \sum_{j=0}^{t} \psi_\theta(s_j, a_j)$, the common trajectory representation is given by the common trajectory representation above.

D.2 Differentiation under the trajectory integral

For each state, use a common sigma-finite action reference measure and a strictly positive policy density on parameter-independent support. Fix θ and a compact ball around it contained in Θ, with the bound the uniform score condition above throughout that ball. The mean-value theorem gives

$$\frac{\pi_{\theta+v}(a \mid s)}{\pi_\theta(a \mid s)} \le \exp(G\|v\|_2).$$

Let Q_θ be the trajectory law in one fixed world and $Z_v = \mathrm{d}Q_{\theta+v}/\mathrm{d}Q_\theta$. For $\|v\|_2 \le a$ within the chosen ball,

$$Z_v = \prod_{j=0}^{H-1} \frac{\pi_{\theta+v}(a_j \mid s_j)}{\pi_\theta(a_j \mid s_j)} \le e^{HGa},$$

$$\|\nabla_v Z_v\|_2 = \left\| Z_v \sum_{j=0}^{H-1} \psi_{\theta+v}(s_j, a_j) \right\|_2 \le HGe^{HGa}.$$

The return is bounded by HR_b. Dominated convergence permits differentiation under Q_θ. This argument is applied separately in each world and requires no likelihood ratio between P and $\widehat{P}$.

Differentiating policy normalization gives $\int \psi_\theta(s,a)\pi_\theta(\mathrm{d}a \mid s) = 0$. Let $r_t = r(s_t, a_t)$. The likelihood-ratio identity initially reads

$$\nabla J = \mathbb{E}_P \left[\left(\sum_{j=0}^{H-1} \psi_\theta(s_j, a_j) \right) \left(\sum_{t=0}^{H-1} r_t \right) \right].$$

Let $\mathcal{H}_j = \sigma(s_0, a_0, \ldots, a_{j-1}, s_j)$ be the history before action a_j. For $j > t$, r_t is $\mathcal{H}_j$-measurable, and

$$\mathbb{E}[\psi_\theta(s_j, a_j) r_t] = \mathbb{E}[r_t \mathbb{E}[\psi_\theta(s_j, a_j) \mid \mathcal{H}_j]] = 0.$$

Removing these terms gives

$$\nabla J = \mathbb{E}_P F_\theta, \qquad F_\theta = \sum_{t=0}^{H-1} S_t r_t, \qquad S_t = \sum_{j=0}^{t} \psi_\theta(s_j, a_j).$$

Since $\|S_t\|_2 \le (t+1)G$, $\|F_\theta\|_2 \le C_H$. The same functional represents the imagined gradient when rewards are shared.

D.3 Sequential coupling without a support assumption

For standard Borel kernels a measurable maximal coupling can be constructed from their common part. At a state–action pair let $\nu = P + \widehat{P}$, write their densities relative to ν as $p, \widehat{p}$, and use $\min(p, \widehat{p})\nu$ as the common subprobability. Its mass is $1 - \mathrm{TV}(P, \widehat{P})$. Sample identically from this part and couple the residual parts arbitrarily. This gives both required marginals and disagreement probability equal to the local total variation.

Run that coupling only while the histories agree, using identical policy draws on common histories. Let E_j be first disagreement at transition j. If A_j is agreement through (s_j, a_j), then

$$\mathbb{P}(E_j) = \mathbb{E}\Big[\mathbb{I}\{A_j\}\,\mathrm{TV}\{P(\cdot \mid s_j, a_j), \widehat{P}(\cdot \mid s_j, a_j)\}\Big] \le d_j.$$

Let $Q_t, \widehat{Q}_t$ be the prefix laws through (s_t, a_t). The first-disagreement events are disjoint, so

$$\mathrm{TV}(Q_t, \widehat{Q}_t) \le \mathbb{P}\left(\bigcup_{j<t} E_j\right) = \sum_{j<t} \mathbb{P}(E_j) \le \sum_{j<t} d_j.$$

Combining with $\mathrm{TV}(Q_t, \widehat{Q}_t) \le 1$ gives D_t.

For any vector-valued F with $\|F\|_2 \le C$,

$$\|\mathbb{E}_P F - \mathbb{E}_Q F\|_2 = \sup_{\|v\|_2=1} |\mathbb{E}_P\langle v, F\rangle - \mathbb{E}_Q\langle v, F\rangle| \le 2C\,\mathrm{TV}(P, Q).$$

Using $\|S_t r_t\|_\infty \le (t+1)GR_b$,

$$\begin{aligned}\|g - \widehat{g}\|_2 &\le \sum_{t=0}^{H-1} \left\| \int S_t r_t \,\mathrm{d}(Q_t - \widehat{Q}_t) \right\|_2, \\ &\le 2GR_b \sum_{t=0}^{H-1} (t+1) D_t \le 2C_H \min\{1, e\}.\end{aligned}$$

With uniform local error ϵ,

$$b_\theta \le 2GR_b \sum_{t=0}^{H-1} (t+1)\min\{1, t\epsilon\}.$$

For small $H\epsilon$, this is $2GR_b\epsilon H(H-1)(H+1)/3$. We do not claim this cubic horizon dependence is minimax optimal over policy-gradient MDPs.

D.4 Prediction error and comparison-specific sensitivity

Proof of Corollary 2. Fix $\epsilon \in (0, 1/2]$ and a two-step task with initial actions a_0, a_1, terminal states g, b, zero initial reward, and terminal reward $c \in (0, R_b]$ at g. In the true world let

$$P(g \mid a_0) = \frac{1}{2}, \qquad P(g \mid a_1) = \frac{1}{2} - \frac{\epsilon}{2},$$

whereas the learned world model satisfies

$$\widehat{P}(g \mid a_0) = \frac{1}{2}, \qquad \widehat{P}(g \mid a_1) = \frac{1}{2} + \frac{\epsilon}{2}.$$

The remaining probability is assigned to b, and the two worlds agree elsewhere. Hence

$$\sup_a \mathrm{TV}\Big(P(\cdot \mid a), \widehat{P}(\cdot \mid a)\Big) = \epsilon,$$

but the learned world model assigns comparison $c\epsilon/2$ to replacing a_0 by a_1, while the true comparison is $-c\epsilon/2$. Thus arbitrarily small uniform transition error reverses the action ranking.

Parameterize the two-action construction of Corollary 2 by $\pi_\theta(a_1) = \sigma(\theta) = (1 + \exp(-\theta))^{-1}$. Then

$$J(\theta) = \frac{c}{2} - \frac{c\epsilon}{2}\sigma(\theta), \qquad \widehat{J}(\theta) = \frac{c}{2} + \frac{c\epsilon}{2}\sigma(\theta).$$

The same construction has a differential consequence. The learned world model ranks a_1 above a_0 while the true world ranks them in the opposite order; hence world-model-greedy planning is wrong without any parameter update. Increasing the logistic probability of a_1 strictly decreases the true return, so optimizing the predictive surrogate follows the wrong direction.

For the converse, let the next state be $(y, v) \in \{0, 1\}^2$ and the terminal reward be cy, with $c \in (0, R_b]$. At the initial state, actions a_0, a_1 produce $y = 1$ with probabilities $1/4, 3/4$ in both worlds. The true world sets $v = 0$ and the learned world sets $v = 1$. Terminal dynamics and rewards agree. The two initial next-state laws have disjoint supports, so their total variation is one. Nevertheless every policy has the same value in both worlds: only its initial action affects reward, and the terminal action is immaterial. With logistic initial action probability,

$$J(\theta) = \widehat{J}(\theta) = c[1/4 + \sigma(\theta)/2].$$

Values, action ordering, and policy gradients are all exact despite maximal transition discrepancy. This is why a total-variation-based sufficient gate need not characterize every valid use.

Qualification is comparison-specific. Consider three actions with terminal success probabilities

$$p_0 = 1/4, \qquad p_1 \in [1/4 - \epsilon/2, 1/4 + \epsilon/2], \qquad p_2 = 3/4,$$

where $0 < \epsilon \leq 1/4$. Supply the upper-endpoint model and take the lower endpoint as the true world. Replacing a_0 by a_2 has contrast $c/2$ in every plausible world; replacing a_0 by a_1 has contrast $-c\epsilon/2$ in truth and $c\epsilon/2$ in the model. Thus a certificate for the first comparison does not certify the second, even within the same task and with the same predictive model.

D.5 Why value equality and joint parameter updates are insufficient

In a two-step task, initial actions a_1, a_0, in this order, lead to terminal reward c with true success probabilities $1, 0$ and learned success probabilities $0, 1$. For the logistic policy $\pi_\theta(a_1) = \sigma(\theta)$,

$$J(\theta) = c\sigma(\theta), \qquad \widehat{J}(\theta) = c[1 - \sigma(\theta)].$$

At $\theta = 0$ the values both equal $c/2$, but the gradients are $c/4$ and $-c/4$. Policy scores are bounded by one. Equality of scalar values at a point therefore provides no gradient certificate, even with bounded rewards and regular policies.

Freezing the world is also essential. Consider a one-action bandit with true reward $1/2$ and model reward $\widehat{r}_\theta = 1/2 + \theta$ for $|\theta| < 1/4$. At $\theta = 0$ model error is zero and the policy cannot change the true return. Nevertheless the total derivative of the model return is one. This derivative changes the model, not the policy. The main gradient theorem controls derivatives through a fixed predictive world; it does not certify unrestricted shared-parameter co-training. Shared-parameter methods require separating these derivative paths or proving an additional bound.

D.6 Reward-model error

Let $\widehat{r}$ also be fixed with respect to θ, and bounded in $[0, R_b]$. The learned gradient now uses $\widehat{F}_\theta = \sum_t S_t \widehat{r}(s_t, a_t)$. Adding and subtracting $\mathbb{E}_{\widehat{P}} F_\theta$ gives

$$\|g - \widehat{g}\|_2 \leq 2GR_b \sum_{t=0}^{H-1} (t+1) D_t + G \sum_{t=0}^{H-1} (t+1) \mathbb{E}_{\widehat{\mu}_t^{\pi_\theta}} |r - \widehat{r}|.$$

Indeed, the first difference uses the common true-reward functional and prefix coupling; the second is bounded pointwise by $\sum_t (t+1) G |r - \widehat{r}|$ and integrated under the learned occupancy. A uniform reward error ϵ_r yields the additional radius $G\epsilon_r H(H+1)/2$. Every gate must include this additional radius when rewards are learned. A sampled reward discrepancy is not a uniform bound unless separately certified.

D.7 Sharp angular geometry

We prove the sharp angular consequence the angular inequality above for Theorem 5. Here $\cos(g, \widehat{g}) = \langle g, \widehat{g} \rangle / (\|g\|_2 \|\widehat{g}\|_2)$.

Let $v \neq 0$, $h = \|v\|_2$, and suppose $\|q - v\|_2 \leq b < h$. Put $u = v/h$ and decompose $q = xu + y$ with $y \perp u$. The constraint is $(x-h)^2 + \|y\|_2^2 \leq b^2$, so $x > 0$. The largest possible squared tangent of the angle is

$$\max_{x \in [h-b, h+b]} \frac{b^2 - (x-h)^2}{x^2}.$$

Writing $F(x) = [b^2 - (x-h)^2]/x^2$, we have

$$F'(x) = \frac{2(h^2 - b^2 - hx)}{x^3}, \qquad x_* = (h^2 - b^2)/h, \qquad F(x_*) = \frac{b^2}{h^2 - b^2}.$$

Since x_* lies in $[h-b, h+b]$ and the derivative changes from positive to negative there,

$$\frac{\langle q, v \rangle}{\|q\|_2 \|v\|_2} \geq \frac{\sqrt{h^2 - b^2}}{h}.$$

For dimension at least two and $b > 0$, choose a perpendicular y attaining the boundary to obtain equality. If $b = 0$ or the dimension is one, the cosine is one; the displayed lower bound still holds. If $b \geq h$, the uncertainty ball includes zero, and for $b > h$ it includes vectors pointing against v. This proves the exact robust threshold for an acute-angle guarantee.

D.8 Optimal robust displacement and sharpness

For every displacement v we prove the exact robust-support formula the robust-support identity above.

Suppose a differentiable objective has unknown gradient in the ball $\{q : \|q - \widetilde{g}\|_2 \leq b\}$ and satisfies the local smooth lower model $J(\theta + v) - J(\theta) \geq \langle q, v \rangle - L\|v\|_2^2/2$. For $v \neq 0$, the adverse gradient is $q = \widetilde{g} - bv/\|v\|_2$. Hence the robust-support identity above holds exactly; for $v = 0$ both sides are zero.

Put $h = \|\widetilde{g}\|_2$ and let $\Phi(v)$ denote the right-hand side of the robust-support identity above. Cauchy–Schwarz and one-variable maximization give

$$\sup_{\|v\|_2 = t} \Phi(v) = (h-b)t - \tfrac{L}{2}t^2,$$

$$\sup_{v \in \mathbb{R}^d} \Phi(v) = \sup_{t \geq 0}\{(h-b)t - \tfrac{L}{2}t^2\} = \frac{[h-b]_+^2}{2L}.$$

For $h > b$, the maximizer is $v_* = (h-b)\widetilde{g}/(Lh)$; for $h \leq b$, it is $v_* = 0$. The unconstrained maximum is attainable as a certified step only if the segment from θ to $\theta + v_*$ is feasible and lies in the smoothness region.

For a specified v, the quadratic

$$f(\theta + w) = f(\theta) + \langle q, w \rangle - \frac{L}{2}\|w\|_2^2$$

attains the lower model. These local quadratics establish sharpness given gradient-ball and smoothness information. They are not asserted to be globally bounded MDP returns; a tighter guarantee exploiting further MDP structure is not ruled out.

A sufficient smoothness bound is available when policies are twice continuously differentiable and $\|\nabla_\theta^2 \log \pi_\theta(a \mid s)\|_2 \leq K$ uniformly on a convex region. Differentiating the prefix expression for expected reward yields

$$\nabla_\theta^2 J = \mathbb{E}_P \sum_{t=0}^{H-1} r_t \left[S_t S_t^\top + \sum_{j=0}^{t} \nabla_\theta^2 \log \pi_\theta(a_j \mid s_j) \right].$$

The domination argument used for the first derivative extends using these bounds. Therefore

$$L = R_b \left[G^2 \frac{H(H+1)(2H+1)}{6} + K \frac{H(H+1)}{2} \right]$$

is sufficient. A smaller valid local L improves the gate, but a numerical curvature estimate alone is not a guaranteed upper bound.

D.9 Actual optimizers and latent interfaces

The comparison certificate applies to a realized output of any predictive search procedure. The differential condition is useful when that procedure optimizes a parameterized behavior. It validates the proposed displacement; it does not require a synthetic-data training pipeline.

For differentiable policies, a proposed displacement need not be parallel to a policy-gradient estimate. If $\|g - \widetilde{g}\|_2 \le b$ and the proposed segment is L-smooth, then for every feasible v,

$$J(\theta + v) - J(\theta) \ge \langle \widetilde{g}, v \rangle - b\|v\|_2 - \frac{L}{2}\|v\|_2^2.$$

This follows by inserting the gradient-ball support function into the smoothness inequality. It can certify a preconditioned, clipped, or otherwise proposed optimizer displacement without identifying it with the exact gradient. The estimate must target the score-function gradient of the frozen imagined world. If a practical estimate has an additional bias bounded by ζ, replace b by $b_\theta + \xi + \zeta$. Merely naming a critic or using automatic differentiation does not establish such a bound.

A common history interface need not require reconstructing every observation. The following bridge states precisely what is sufficient for a fixed latent policy interface.

Proposition 5 (History-to-latent reliability bridge)**.** *Let h be a true observable history, $s = f(h)$ a measurable, parameter-independent representation, and let the policy use only s. Start the latent model at the pushforward of the true initial history law. Write $K(\cdot \mid h, a)$ for the true conditional law of the next representation and $\widehat{P}(\cdot \mid f(h), a)$ for its learned prediction. Suppose, uniformly in admissible histories and actions,*

$$|r(h,a) - \widehat{r}(f(h),a)| \le \epsilon_r, \qquad \mathrm{TV}\{K(\cdot \mid h,a), \widehat{P}(\cdot \mid f(h),a)\} \le \epsilon_p.$$

Then every such policy satisfies

$$|J - \widehat{J}| \le H\epsilon_r + \frac{H(H-1)}{2} R_b \epsilon_p.$$

Under the same fixed-interface score regularity as Theorem 5, its gradient discrepancy is at most

$$\|g - \widehat{g}\|_2 \le 2GR_b \sum_{t=0}^{H-1} (t+1)\min\{1, t\epsilon_p\} + \frac{GH(H+1)}{2}\epsilon_r.$$

The true representation process need not be Markov.

Proof. Lift the learned value to histories as $\widehat{V}_t(f(h))$. Subtract its learned Bellman recursion from the true history recursion and add the true conditional expectation of $\widehat{V}_{t+1}(f(h'))$. The residual is the reward difference plus $(K - \widehat{P})\widehat{V}_{t+1}$, bounded by $\epsilon_r + (H - t - 1)R_b\epsilon_p$. The remaining true conditional difference telescopes exactly as in Theorem 2. Summing proves the latent value bound above.

For gradients, keep the complete true history in the construction while coupling the latent trajectories. On agreement of latent prefixes, the two action laws coincide. Couple the true next representation, conditional on its full history, with the learned latent transition by maximal coupling; disagreement probability is at most ϵ_p. A regular conditional distribution of the next true history given its representation preserves the full true marginal. Standard Borel assumptions ensure these conditional kernels exist. Once prefixes separate, continue with any marginal-preserving coupling. The latent prefix discrepancy through time t is at most $\min\{1, t\epsilon_p\}$.

Use the learned reward functional in both worlds. The true gradient's reward-replacement error is at most $G\epsilon_r \sum_t (t+1)$. The remaining common functional depends only on latent prefixes; the score is $\nabla \log \pi_\theta(a \mid f(h))$ in the true process and the same function of (s, a) in the learned process. Applying the prefix argument proves the latent gradient bound above. Parameter independence of f and the supplied world is required for this common functional.

The premise includes representation aliasing: histories with the same representation must all have predictions close enough to the supplied kernel. Passive reconstruction loss does not establish it.

If the representation admits an exact Markov kernel, finite-state row verification can be applied to that kernel; otherwise a latent-row average needs an additional uniform aliasing bound. Encoder updates or shared-parameter model updates require rechecking the interface and derivative premises. These distinctions allow the theory to be used with learned representations without assuming their sufficiency by definition.

D.10 Dimension-free imagined-gradient sampling

A sampling radius for the differential certificate is

$$\xi = \frac{C_H}{\sqrt{N}}\left(1 + \sqrt{2\log(1/\delta)}\right)$$

where N is the number of independent imagined trajectories and $\delta \in (0,1)$.

Proposition 6 (Bounded-vector sampling)**.** *Condition on a fixed policy and learned world. Let $Y_1, \dots, Y_N$ be independent samples of $\widehat{F}_\theta$ with $\|Y_i\|_2 \le C_H$, and let $\widetilde{g} = N^{-1}\sum_i Y_i$. Then $\|\widetilde{g} - \widehat{g}\|_2 \le \xi$ with probability at least $1-\delta$, for ξ in the stated sampling radius. The radius may be truncated at $2C_H$.*

Proof. Independence and centering cancel cross terms:

$$\mathbb{E}\|\widetilde{g} - \widehat{g}\|_2^2 = \frac{1}{N^2}\sum_i \mathbb{E}\|Y_i - \widehat{g}\|_2^2 \le \frac{C_H^2}{N}.$$

Jensen gives $\mathbb{E}\|\widetilde{g} - \widehat{g}\|_2 \le C_H/\sqrt{N}$. Replacing one sample changes this norm by at most $2C_H/N$. The bounded-differences inequality therefore bounds the probability of exceeding its expectation by t by $\exp[-Nt^2/(2C_H^2)]$. Take $t = C_H\sqrt{2\log(1/\delta)/N}$. Finally $\|\widetilde{g}\|_2, \|\widehat{g}\|_2 \le C_H$, giving the deterministic truncation.

The bound has no parameter-dimension factor because the assumed norm bound already controls the entire vector. It applies to independent trajectory estimators of the stated gradient, not automatically to replay-correlated minibatches, bootstrapped actor losses, or biased value-gradient estimators. An additional estimator-bias radius must be added for those alternatives.

E Statistical Certification and Adaptive Verification

E.1 Calibrating decision-relevant prediction error

Let $x = (z, \widehat{\mathcal{M}}, \pi_0, \pi_1)$ be a complete decision request, with $e(x) = |\widehat{\Gamma}(x) - \Gamma_{\mathcal{M}}(x)|$. An independently fitted score u predicts this nonnegative error. Conditional on that fit, assume exchangeability of the calibration pairs (x_j, e_j) and one future pair. For $\alpha \in (0,1)$ put

$$q_j = e_j - u(x_j), \quad k = \lceil (n+1)(1-\alpha) \rceil, \quad U_\alpha(x) = [u(x) + q_{(k)}]_+.$$

Set $q_{(k)} = +\infty$ if $k > n$. The request includes the actual model, reference, candidate selection procedure and continuation; changing their distribution is a change of the calibration population.

Proposition 7 (Calibrated contrast and false-promotion risk)**.** *Under the preceding exchangeability assumption, $\mathbb{P}\{e(x) > U_\alpha(x)\} \le \alpha$. Suppose $\mathbb{P}\{\widehat{\Gamma}(x) < S(x) - \xi(x)\} \le \delta$. Define*

$$T(x) = S(x) - U_\alpha(x) - \xi(x).$$

Then

$$\mathbb{P}\{T(x) > 0,\ \Gamma_{\mathcal{M}}(x) \le 0\} \le \alpha + \delta.$$

Proof. After independent randomized tie-breaking, exchangeability makes the future residual's rank K uniform on $\{1, \dots, n+1\}$. Therefore

$$\begin{aligned}\mathbb{P}\{q > q_{(k)}\} &\le \mathbb{P}\{K > k\} \\ &= \sum_{j=k+1}^{n+1} \mathbb{P}\{K = j\} = \frac{n+1-k}{n+1} \le \alpha.\end{aligned}$$

Since $q = e - u(x)$ and $U_\alpha = [u(x) + q_{(k)}]_+$, the event $e > U_\alpha$ is contained in $q > q_{(k)}$. On its complement and on $\widehat{\Gamma} \geq S - \xi$,

$$\begin{aligned}\Gamma_{\mathcal{M}} &= \widehat{\Gamma} + (\Gamma_{\mathcal{M}} - \widehat{\Gamma}) \\ &\geq \widehat{\Gamma} - |\Gamma_{\mathcal{M}} - \widehat{\Gamma}| \\ &\geq S - \xi - U_\alpha = T.\end{aligned}$$

Hence

$$\begin{aligned}\{T > 0, \Gamma_{\mathcal{M}} \leq 0\} &\subseteq \{e > U_\alpha\} \cup \{\widehat{\Gamma} < S - \xi\}, \\ \mathbb{P}\{T > 0, \Gamma_{\mathcal{M}} \leq 0\} &\leq \mathbb{P}\{e > U_\alpha\} + \mathbb{P}\{\widehat{\Gamma} < S - \xi\} \\ &\leq \alpha + \delta.\end{aligned}$$

This is split conformal calibration applied to a decision-relevant error target (Angelopoulos & Bates, 2022). It does not identify the true error from a single transition. A valid upper label such as $\varepsilon_{\mathrm{I}}(\pi_0) + \varepsilon_{\mathrm{I}}(\pi_1)$ may replace e when exact contrast errors are unavailable, subject to the labeling conditions below.

Calibration versus raw uncertainty scores. The certificate concerns a calibrated upper envelope of the decision-relevant error $e(x) = |\widehat{\Gamma}(x) - \Gamma_{\mathcal{M}}(x)|$. A confidence, disagreement, entropy, or critic score is therefore not itself an error certificate: it becomes usable in the gate only after an argument establishes the required upper-error guarantee for the request population under consideration.

E.2 Split conformal coverage and the exact promotion event

Condition on the training data used to fit u. Let $q_1, \ldots, q_n, q$ be the exchangeable calibration and future residuals, and let K denote the rank of q after independent randomized tie-breaking. For $k \leq n$,

$$\begin{aligned}\mathbb{P}(K = j) &= (n+1)^{-1}, \qquad j = 1, \ldots, n+1, \\ \mathbb{P}\{q > q_{(k)}\} &\leq \mathbb{P}(K > k) = \frac{n+1-k}{n+1} \leq \alpha.\end{aligned}$$

Since $q = e - u(x)$ and $U_\alpha(x) = \max\{0, u(x) + q_{(k)}\}$,

$$\{e > U_\alpha(x)\} \subseteq \{q > q_{(k)}\}, \qquad \mathbb{P}\{e > U_\alpha(x)\} \leq \alpha.$$

For $k = n + 1$, set $q_{(k)} = +\infty$; coverage is then immediate.

Define the events $E = \{e \leq U_\alpha\}$ and $F = \{\widehat{\Gamma} \geq S - \xi\}$. On $E \cap F$, positive T implies strictly positive true gain. Therefore

$$\{T > 0,\ \Gamma_{\mathcal{M}}(x) \leq 0\} \subseteq E^c \cup F^c.$$

Consequently, without independence of E, F,

$$\mathbb{P}\{T > 0,\ \Gamma_{\mathcal{M}}(x) \leq 0\} \leq \mathbb{P}(E^c) + \mathbb{P}(F^c) \leq \alpha + \delta.$$

In contrast, if $p = \mathbb{P}(T > 0) > 0$, the argument gives at most

$$\mathbb{P}\{\Gamma_{\mathcal{M}}(x) \leq 0 \mid T > 0\} \leq \min\{1, (\alpha + \delta)/p\},$$

not $\alpha + \delta$. A generic marginal certificate can fail exclusively on a selected minority: let X be Bernoulli with mean α, let $e(X) = X$, and set $U(X) = 0$. Marginal coverage is $1 - \alpha$, yet conditional miscoverage given $X = 1$ is one. This demonstrates why marginal validity alone cannot justify selection-conditional validity; it is not a claim that this particular U was produced by split conformal prediction.

For multiple future snapshots chosen before observing calibration outcomes, one may compute each marginal certificate at level α/m for a fixed pool of size m. A union bound gives simultaneous coverage at least $1 - \alpha$ over the pool, permitting arbitrary subsequent selection within it. Each pair must still have the required marginal exchangeability; the pool cannot be manufactured adaptively from the same calibration residuals without further analysis.

E.3 NOISY LABELS AND FINITE STRATA

If exact errors are unavailable, suppose an identically applied labeling procedure produces exchangeable upper labels $\overline{e}_i$, including a hypothetical future $\overline{e}$, and $\mathbb{P}(e > \overline{e}) \le \beta$. Calibrate the residuals $\overline{e}_i - u(x_i)$ instead. Conformal coverage of $\overline{e}$ and a union bound then yield

$$\mathbb{P}\{e > U_\alpha(x)\} \le \alpha + \beta.$$

This statement requires upper labels and their separate validity bound, not merely unbiased estimates. A pixel discrepancy or mean state error may instead define a different target; conformal coverage of that target does not bound e without a proven bridge.

For a prespecified finite partition $c(x) \in \{1, \dots, m\}$, calibrate separately within each stratum. Conditional on a future stratum h, the fitted score, and its calibration count n_h, assume the corresponding residuals are exchangeable. The same rank argument with $k_h = \lceil (n_h + 1)(1 - \alpha) \rceil$ gives

$$\mathbb{P}\{e \le U_h(x) \mid c(x) = h\} \ge 1 - \alpha.$$

The partition is fixed independently of calibration errors. Conditioning on arbitrary within-stratum promotion decisions is still not covered.

E.4 FRESH FINITE-STATE VERIFICATION

This construction supplies a fully observable certificate under stronger access assumptions. Fix a finite task with $D = |\mathcal{S}|$ and $m = |\mathcal{S}||\mathcal{A}|$ state–action rows. After choosing the task, obtain n independent next-state samples per row from a generative oracle. For learned rewards, also obtain bounded reward samples in $[0, R_b]$ with the correct conditional means. Samples within each row are conditionally independent given the pre-audit history. Across-row independence is unnecessary. Write $\overline{P}, \overline{r}$ for the empirical kernels and means and define

$$a_n = \sqrt{\frac{D \log 2 + \log(4m/\alpha)}{2n}}, \qquad c_n = R_b \sqrt{\frac{\log(4m/\alpha)}{2n}}.$$

For each row set

$$u(s,a) = \min\{1, \mathrm{TV}(\overline{P}, \widehat{P}) + a_n\}, \qquad v(s,a) = \min\{R_b, |\overline{r} - \widehat{r}| + c_n\}.$$

The symbols s, a in row arguments denote a state and an action.

Proposition 8 (Simultaneous fresh-audit certificate)**.** *Conditional on the pre-audit history, with probability at least $1 - \alpha$, every row obeys $\mathrm{TV}(P, \widehat{P}) \le u$ and $|r - \widehat{r}| \le v$. The statement holds simultaneously for all bounded world-model candidates, including candidates fitted using the audit. If $u_* = \max_{s,a} u(s,a)$ and $v_* = \max_{s,a} v(s,a)$, then every policy has*

$$\varepsilon_{\mathrm{I}}(\pi) \le H v_* + \frac{H(H-1)}{2} R_b u_*,$$

and every policy satisfying the score bound has

$$\|g - \widehat{g}\|_2 \le 2GR_b \sum_{t=0}^{H-1} (t+1) \min\{1, t u_*\} + \frac{GH(H+1)}{2} v_*.$$

Proof. For any fixed subset of the D next states, its empirical probability is the mean of Bernoulli variables. Hoeffding bounds absolute error above a_n by $2\exp(-2na_n^2)$. Union over at most 2^D subsets and m rows gives failure probability at most $\alpha/2$ for $\mathrm{TV}(P, \overline{P}) \le a_n$. A separate Hoeffding and row union bound give $|r - \overline{r}| \le c_n$ with failure probability at most $\alpha/2$.

On this common event, triangle inequalities give the row envelopes above. The event constrains only the true and empirical kernels; hence the inequalities hold for every supplied world model, including data-dependent ones. Clipping is valid because total variation is at most one and rewards differ by at most R_b. The uniform value inequality proves the audited value bound above. The reward-gradient bound and $d_t \le u_*$ prove the audited gradient bound, uniformly over the admitted policies.

This verifier costs mn real row queries and does not infer total variation from prediction loss. Its strength is simultaneous validity after policy selection; its weakness is dependence on finite spaces and generative coverage. In large or continuous spaces it must be replaced by a justified structured confidence set or a direct return audit, not presented as computationally free.

E.5 SHARED EVIDENCE UNDER ADAPTIVE PREDICTIVE DECISIONS

Proof of Theorem 4. Let E be the simultaneous event from Proposition 8. It constrains the true and empirical transition rows, and reward means when audited, independently of the candidate policies or learned world models. For every selected world model, let u_i, v_i be the maxima of the transition and reward envelopes defined above. Then

$$B_i = 2Hv_i + R_bH(H-1)u_i, \qquad |\Gamma_{\mathcal{M}}(x_i) - \widehat{\Gamma}_i| \le B_i \quad \text{on } E.$$

For known shared rewards put $v_i = 0$. The triangle inequalities hold simultaneously for models fitted on the audit as well as for independently fitted models.

Let $\mathcal{H}_i$ contain the audit, prior requests and observations, the current world model, and the selected policy pair, before estimation of the current contrast. For $F_i = \{\widehat{\Gamma}_i \ge S_i - \xi_i\}$ assume $\mathbb{P}(F_i^c \mid \mathcal{H}_i) \le \delta_i$. Write $\mathcal{I}$ for the random set of accepted indices. On $E \cap F_i$, any accepted request satisfies

$$\begin{aligned}\Gamma_{\mathcal{M}}(x_i) &\ge \widehat{\Gamma}_i - B_i \\ &\ge S_i - \xi_i - B_i = T_i > 0.\end{aligned}$$

Consequently, the bad-promotion event obeys

$$\mathcal{B} = \bigcup_{i\ge1} \big(\{i \in \mathcal{I}\} \cap \{\Gamma_{\mathcal{M}}(x_i) \le 0\}\big) \subseteq E^c \cup \bigcup_{i\ge1} F_i^c.$$

The conditional validity of the fresh estimate and the tower property give

$$\begin{aligned}\mathbb{P}(\mathcal{B}) &\le \mathbb{P}(E^c) + \mathbb{P}\left(\bigcup_{i\ge1} F_i^c\right) \\ &\le \mathbb{P}(E^c) + \sum_{i\ge1} \mathbb{P}(F_i^c) \\ &= \mathbb{P}(E^c) + \sum_{i\ge1} \mathbb{E}[\mathbb{P}(F_i^c \mid \mathcal{H}_i)] \\ &\le \alpha + \sum_{i\ge1} \delta_i.\end{aligned}$$

No conditional coverage of E after selection is used. The argument requires the unconditional common event and conditional validity of each fresh estimation step.

For a selected pair, draw N independent paired imagined episodes and let S_i be the sample mean difference. Each difference lies in $[-HR_b, HR_b]$, so

$$\widehat{\Gamma}_i \ge S_i - \xi_i, \qquad \xi_i = HR_b\sqrt{\frac{2\log(1/\delta_i)}{N}}$$

with conditional probability at least $1 - \delta_i$. Pairing within an index is allowed; pairs must be independent. If the candidate is selected from K prespecified candidates using these same samples, replace δ_i by δ_i/K and union the bounds. Unrestricted selection requires fresh evaluation or a justified uniform estimator. This sampling restriction is separate from reuse of the model audit.

For Theorem 3, the uniform row event bounds every conditional Q_t^0 through (16). Exact world-model evaluation then gives simultaneous local certificates at all states. With sampled local evaluation, either use a uniform finite-state correction or conditional fresh estimates along the actual trajectory. The latter controls bad local decisions along that trajectory; it must not be presented as a uniform bound over all unvisited histories.

For a fixed task archive, allocate audit risk α_z to task z. Total false-promotion risk is at most $\sum_z \alpha_z + \sum_i \delta_i$, with each stationary task's model risk charged once. If a task's transition kernel drifts by at most β in uniform TV, replace u_i by $\min\{1, u_i + \beta\}$. Reward drift analogously increases v_i. These statements follow from triangle inequalities. A new task or unbounded drift requires new coverage evidence.

E.6 AMORTIZED EVIDENCE COST AT POSITIVE DECISION MARGINS

Assume shared rewards, $H \geq 2$ and $R_b > 0$. Set

$$K = R_b H(H-1), \qquad q_i = \max_{s,a} \mathrm{TV}(P, \widehat{P}_i), \qquad a_i = \widehat{\Gamma}_i - Kq_i.$$

The quantity a_i is a sufficient margin involving the unknown true model error, not an observable gain estimate.

Proposition 9 (Amortized verification)**.** *On the row-confidence event, $u_i \leq q_i + 2a_n$ simultaneously for every world model. Exact world-model evaluation therefore gives* (28)*. The sample size* (29) *certifies every request with $a_i/K \geq s$. For a sampled estimate satisfying $|S_i - \widehat{\Gamma}_i| \leq \xi_i$, the corresponding bound is*

$$T_i \geq a_i - 2Ka_n - 2\xi_i.$$

Proof. For every row, the triangle inequality gives

$$\mathrm{TV}(\overline{P}, \widehat{P}_i) + a_n \leq \mathrm{TV}(P, \widehat{P}_i) + 2a_n.$$

Maximization and clipping preserve this upper bound. With $S_i = \widehat{\Gamma}_i$ and $\xi_i = 0$, $T_i = \widehat{\Gamma}_i - Ku_i \geq a_i - 2Ka_n$. Condition (29) is exactly $2a_n < s$. If $a_i/K \geq s$, it follows that $T_i \geq K(s - 2a_n) > 0$. For sampled evaluation, $S_i \geq \widehat{\Gamma}_i - \xi_i$ gives the sampled-margin inequality above.

The real evidence cost is mn, instead of Nmn for repeating that row audit before each of N comparisons. Imagined evaluation and candidate search remain additional costs. Positive margins are a premise; their existence is not implied by confidence coverage. The statement is deterministic on the common event, so it remains valid for adaptive requests meeting that premise. A post-selection random margin cannot be treated as a prespecified unconditional power guarantee. When $H = 1$ and rewards are shared, transitions do not affect returns and no transition audit is required.

E.7 STOPPING AN AUDIT AT A POSITIVE MARGIN

For adaptively chosen audit size, use

$$\alpha_n = \frac{\alpha}{n(n+1)}, \qquad a_n = \sqrt{\frac{D\log 2 + \log(4m/\alpha_n)}{2n}}.$$

For each row take successive prefixes of an independent sample stream. Since $\sum_{n\geq 1} \alpha_n = \alpha$, one event covers all integer sample sizes and permits stopping at the first positive margin. The log factor now grows with n; the fixed-size bound is not an anytime bound.

For a fixed request with exact imagined contrast and positive sufficient margin $a = \widehat{\Gamma} - Kq$, fixed-size verification needs at most any integer satisfying

$$n > \frac{2K^2}{a^2}[D\log 2 + \log(4m/\alpha)].$$

The right side is a sufficient audit budget, not a bound on how quickly learning reduces the model error q.

E.8 A LOWER BOUND ON THE EVIDENCE NEEDED FOR PROMOTION

We prove the necessary query budget (30), with Bernoulli relative entropy kl defined below.

Proposition 10 (Evidence required by a useful gate)**.** *For every $\gamma \in (0, 1/4]$ there are two two-step worlds and one fixed supplied world model such that one fixed candidate decision is beneficial in one and harmful in the other. If the audit's only world-dependent observations are row-query answers, promotion probability at least $1 - \delta$ in the first world and at most δ in the second requires* (30)*, for $0 < \delta < 1/2$, even with adaptive query selection.*

We prove Proposition 10. All side information at the start of the audit is identical under the two hypotheses; only new oracle answers distinguish them. Use horizon two and states $\{s, g, b\}$, with zero reward at s, b and terminal reward $c \in (0, R_b]$ at g. There are two actions. In both worlds $P(g \mid s, a_0) = 1/2$; in the plus and minus worlds, respectively, $P(g \mid s, a_1) = 1/2 + \gamma$ and $1/2 - \gamma$. All other rows agree. The supplied world model is the plus world, and the same policy family is used in both: $\pi_\theta(a_1 \mid s) = \sigma(\theta)$. Then

$$J_+(\theta) = c[1/2 + \gamma\sigma(\theta)], \qquad J_-(\theta) = c[1/2 - \gamma\sigma(\theta)].$$

Replacing a_0 by a_1 has contrasts $c\gamma$ and $-c\gamma$, respectively. Thus a safe and useful promotion decision must distinguish the two worlds. The logistic parameterization merely interpolates between the same two actions.

For $p, q \in (0, 1)$ define

$$\mathrm{kl}(p, q) = p \log(p/q) + (1 - p) \log[(1 - p)/(1 - q)].$$

Use the usual limiting conventions at zero and one. One informative query, at (s, a_1), has relative entropy

$$\mathrm{kl}(1/2 + \gamma, 1/2 - \gamma) = 2\gamma \log \frac{1 + 2\gamma}{1 - 2\gamma} \le 16\gamma^2.$$

For the inequality use $\log[(1 + x)/(1 - x)] \le 2x/(1 - x)$ at $x = 2\gamma \le 1/2$. Queries at other rows have zero relative entropy.

Let an audit make at most n adaptively chosen row queries; include its independent randomization in the transcript $Y_{1:n}$. Conditional on $\mathcal{H}_{t-1} = \sigma(Y_{1:t-1})$, its selection kernel is identical under both hypotheses. Writing $\mathbb{P}^n_+, \mathbb{P}^n_-$ for the transcript laws, the relative-entropy chain rule yields

$$\begin{aligned} D_{\mathrm{KL}}(\mathbb{P}^n_+ \| \mathbb{P}^n_-) &= \sum_{t=1}^{n} \mathbb{E}_+[D_{\mathrm{KL}}(\mathbb{P}_+(Y_t \mid \mathcal{H}_{t-1}) \| \mathbb{P}_-(Y_t \mid \mathcal{H}_{t-1}))] \\ &\le \sum_{t=1}^{n} \mathrm{kl}(1/2 + \gamma, 1/2 - \gamma) \\ &\le 16n\gamma^2. \end{aligned}$$

If the audit stops early, append uninformative queries to reach n; this changes neither the decision nor the divergence bound.

Writing p and q for the plus and minus promotion probabilities, data processing for the binary decision gives transcript divergence at least $\mathrm{kl}(p, q)$. This follows by partitioning the likelihood-ratio integral over promotion and rejection and applying Jensen on each part. Since $p \ge 1 - \delta > \delta \ge q$, Bernoulli relative entropy is increasing in p and decreasing in q on this region; consequently

$$\begin{aligned} 16n\gamma^2 &\ge D_{\mathrm{KL}}(\mathbb{P}^n_+ \| \mathbb{P}^n_-) \\ &\ge \mathrm{kl}(p, q) \\ &\ge \mathrm{kl}(1 - \delta, \delta), \end{aligned}$$

which proves (30). The bound is for evidence needed by a safe, high-power promotion test, not a general lower bound on world-model training. It shows that an arbitrarily small unresolved intervention margin cannot be certified from a fixed amount of real evidence.

Together with the sufficient audit bound in Appendix E.6, this establishes the matching $\Theta(\gamma^{-2})$ dependence on the unresolved decision margin. We do not claim matching state-space or horizon dependence: those factors arise from the particular simultaneous finite-state verifier used for the upper bound.

E.9 Direct return verification for general interfaces

A finite-state confidence set is not necessary if true episodic evaluation is available. Choose policies π, π' before collecting fresh evaluations. Let Z_i be the difference of their returns in independent episode pairs. Pairing within an episode index is allowed, and $Z_i \in [-HR_b, HR_b]$. Then

$$J(\pi') - J(\pi) \ge \overline{Z} - HR_b \sqrt{\frac{2 \log(1/\alpha)}{n}}$$

with probability at least $1-\alpha$, by one-sided Hoeffding. A positive lower bound directly certifies the candidate improvement, including for black-box planners and history-based agents.

This audit is more general but can be expensive: it validates each candidate with real episodes instead of amortizing model reliability across many imagined candidates. Its false-positive event, conditional on pre-audit choice, fits Theorem 7. If candidates are selected using these same evaluations, use a simultaneous finite-pool correction or fresh evaluation again.

E.10 Elementary concentration tools

For completeness, both concentration inequalities used above follow from a bounded exponential moment. If $X \in [a,b]$, let $K(\lambda) = \log \mathbb{E}\exp[\lambda(X - \mathbb{E}X)]$. Under the exponentially tilted law, $K''(\lambda)$ is the variance of X, bounded by $(b-a)^2/4$: center X at $(a+b)/2$ and use that variance is at most the corresponding second moment. Since $K(0) = K'(0) = 0$, integrating twice gives

$$\mathbb{E}\exp[\lambda(X - \mathbb{E}X)] \le \exp\{\lambda^2(b-a)^2/8\}.$$

For independent variables, multiply these bounds. Markov's inequality for the exponential and optimization over $\lambda > 0$ yield $\mathbb{P}\{\overline{X} - \mathbb{E}\overline{X} \ge t\} \le \exp[-2nt^2/(b-a)^2]$. Apply the same argument to $-X$ and add probabilities for a two-sided bound.

If a function $f(X_1, \dots, X_n)$ of independent variables changes by at most c_i when only X_i changes, reveal the variables sequentially. The resulting Doob martingale difference has conditional mean zero and conditional range of length at most c_i: averaging over the remaining independent variables preserves the coordinate-wise difference bound. Applying the bounded exponential-moment inequality above conditionally and iterating gives

$$\mathbb{P}\{f - \mathbb{E}f \ge t\} \le \exp\left[-\frac{2t^2}{\sum_i c_i^2}\right].$$

Using $c_i = 2C_H/N$ proves the tail step in Proposition 6. These arguments also hold conditional on a pre-audit history when the required conditional independence is satisfied.

F Adaptive Co-learning and Resource Accounting

Theorem 7 (Adaptive fresh verification)**.** *Let $\mathcal{H}_{k-1}$ contain all information before round k's fresh audit, including its task choice. If the combined conditional failure probability is at most ε_k and only positive certified comparisons are executed, let $\mathcal{B}$ denote the event of at least one non-improving execution in this sequence. Then*

$$\mathbb{P}(\mathcal{B}) \le \sum_{k \ge 1} \varepsilon_k, \qquad \mathbb{E}N_K \le \sum_{k=1}^{K} \varepsilon_k,$$

where N_K counts non-improving executions through round K.

Proof of Theorem 7. Each bad execution implies audit or estimation failure. Conditional expectation and the tower property give its marginal bound; countable subadditivity and linearity give the two conclusions. Unlike Theorem 4, this permits arbitrary task changes, but pays for fresh validity.

F.1 Filtration and repeated promotion

At round k, $\mathcal{H}_{k-1}$ contains previous world-model fits, policies, generator choices, and audit outcomes, together with the present task choice made before its fresh random samples. Let E_k be the event on which all current world-model-error and sampling bounds used by the gate hold. Assume $\mathbb{P}(E_k^c \mid \mathcal{H}_{k-1}) \le \varepsilon_k$, where ε_k is deterministic or predictable with a deterministic summable upper envelope. For simplicity the main theorem uses deterministic ε_k.

Let I_k indicate that the algorithm accepts a candidate behavior and its true target return does not increase. The deterministic margin theorem gives $I_k \le \mathbb{I}\{E_k^c\}$. Consequently

$$\mathbb{E}[I_k \mid \mathcal{H}_{k-1}] \le \mathbb{E}[\mathbb{I}\{E_k^c\} \mid \mathcal{H}_{k-1}] = \mathbb{P}(E_k^c \mid \mathcal{H}_{k-1}) \le \varepsilon_k,$$
$$\mathbb{E}\sum_{k=1}^{K} I_k = \sum_{k=1}^{K} \mathbb{E}[\mathbb{E}(I_k \mid \mathcal{H}_{k-1})] \le \sum_{k=1}^{K} \varepsilon_k.$$

Moreover, the event of at least one bad accepted comparison is contained in $\bigcup_{k\ge1} E_k^c$, whence

$$\mathbb{P}(\mathcal{B}) \le \mathbb{P}\left(\bigcup_{k\ge1} E_k^c\right) \le \sum_{k\ge1} \mathbb{P}(E_k^c)$$
$$= \sum_{k\ge1} \mathbb{E}[\mathbb{P}(E_k^c \mid \mathcal{H}_{k-1})] \le \sum_{k\ge1} \varepsilon_k.$$

Choosing $\varepsilon_k = \varepsilon/[k(k+1)]$ protects every round with total risk at most ε.

Proposition 8 supplies conditional world-model validity after adaptive task choice; the conditional sampling inequality above supplies simulation accuracy. Their error budgets add. This argument never conditions a marginal guarantee on a data-dependent promotion event. Between rounds all world models and task choices may change; this particular argument restores validity by fresh verification. For repeated uses on one fixed task, Theorem 4 instead retains a simultaneous world-confidence event and charges its failure probability only once.

F.2 Fixed-objective progress and stationarity

The main-text contrast telescope (26) requires no derivatives. For differentiable predictive policy search, the local certificate also yields the following stationarity specialization. The contrast telescope controls expected return rather than samplewise episode rewards. Because $J(\pi) \le HR_b$, any fixed positive certified margin can occur only finitely many times along a sequence of accepted updates; this is the content of (27). The statement does not bound the waiting time, number of rejected proposals, or real samples required between accepted updates. On the event of simultaneous validity, enumerate accepted policy updates by $i = 0, \dots, n-1$. Model-only steps between them do not change the policy. Assume they optimize one fixed bounded objective J, have common smoothness L, and total gradient radius $b_i \le \kappa h_i$, where $h_i = \|\widetilde{g}_i\|_2$ and $0 \le \kappa < 1$. Use $\eta = (1-\kappa)/L$ and require each segment to remain in the admissible region. Theorem 6 gives

$$J(\theta_{i+1}) - J(\theta_i) \ge \frac{(1-\kappa)^2}{2L} h_i^2.$$

Since $\|\nabla J(\theta_i)\|_2 \le (1+\kappa)h_i$, summing yields

$$\frac{(1-\kappa)^2}{2L(1+\kappa)^2} \sum_{i=0}^{n-1} \|\nabla J(\theta_i)\|_2^2 \le J(\theta_n) - J(\theta_0) \le \sup_\theta J(\theta) - J(\theta_0).$$

Therefore

$$\min_{0\le i<n} \|\nabla J(\theta_i)\|_2^2 \le \frac{1}{n}\sum_{i=0}^{n-1} \|\nabla J(\theta_i)\|_2^2 \le \frac{2L(1+\kappa)^2[\sup_\theta J(\theta) - J(\theta_0)]}{(1-\kappa)^2 n}.$$

If infinitely many updates are accepted, boundedness of J also gives

$$\sum_{i=0}^{\infty} \|\nabla J(\theta_i)\|_2^2 < \infty \quad \Longrightarrow \quad \lim_{i\to\infty} \|\nabla J(\theta_i)\|_2 = 0.$$

Existence of infinitely many certified updates is not assumed to follow from validity alone.

A fixed distribution over tasks can be incorporated in the initial state, defining one expected-return objective. By contrast, certifying improvement only for a selected task does not certify improvement of the task average. If objectives change to J_i with $\sup_\theta |J_{i+1}(\theta) - J_i(\theta)| \le v_i$, then

$$\sum_{i=0}^{n-1} [J_i(\theta_{i+1}) - J_i(\theta_i)] \le HR_b + \sum_{i=1}^{n-1} v_{i-1}.$$

To verify this, telescope while adding $J_{i-1}(\theta_i) - J_i(\theta_i)$ at every internal point. The endpoint range is at most HR_b and each internal discrepancy is at most its stated drift bound. Substituting this right-hand side in the preceding calculation gives a drift-corrected average bound for $\|\nabla J_i(\theta_i)\|_2^2$. Without drift control it is not a convergence theorem for a fixed target.

F.3 What reliability does and does not allocate

The frontiers in (18)–(19) are disjoint: $\mathcal{F}_{\mathrm{W}}$ has nonpositive certified margin, while $\mathcal{F}_{\mathrm{A}}$ has positive margin. Membership in $\mathcal{F}_{\mathrm{W}}$ requests evidence or model refinement for a promising comparison. Membership in $\mathcal{F}_{\mathrm{A}}$ authorizes the candidate behavior. Neither identifies failure ownership or supplies an environment-wide guarantee.

A model can be reliable while its imagined return estimate is too noisy. Increasing the rollout count reduces ξ without changing the predictive kernel. A request with no statistically resolved apparent benefit is deferred. This third possibility prevents a two-way routing rule from treating lack of evidence as proof of model ignorance or agent incompetence.

To define the next task, a generator may prioritize uncertainty, measured progress, diversity, or intervention coverage. The theory constrains the use of a selected task after verification; it does not claim an optimal open-ended task generator.

F.4 Deterministic and stochastic precedence accounting

Let $\mathcal{Z}$ now be a finite task archive. For each task fix nonnegative integers w_z, a_z. A productive model update decrements w_z by one. A productive policy update decrements a_z by one only once $w_z = 0$. Updates have no transfer and no regression. An update outside these conditions is ineffective. These assumptions are an explicit learning-response abstraction, not consequences of a value-error bound.

Theorem 8 (Exact precedence accounting)**.** *Let $B^* = \sum_z (w_z + a_z)$. Every completed schedule has length*

$$B = B^* + m,$$

where m is its number of ineffective updates. An eligible work-conserving schedule attains B^*.

More generally, suppose an eligible model trial on task z succeeds independently with probability $p_z > 0$ at cost $c_z > 0$, and an eligible policy trial succeeds with probability $q_z > 0$ at cost $d_z > 0$. Each success decrements its corresponding integer by one. Every schedule that uses only eligible trials and continues until all tasks finish has expected cost

$$B^* = \sum_z \left(\frac{c_z w_z}{p_z} + \frac{d_z a_z}{q_z} \right).$$

Extra ineffective trials add their expected costs.

Proof. For the deterministic case, let $\Phi_k = \sum_z (w_z(k) + a_z(k))$ be the remaining work after k updates, and let I_k indicate that update k is ineffective. At completion time B,

$$\Phi_0 = B^*, \qquad \Phi_B = 0, \qquad \Phi_{k-1} - \Phi_k = 1 - I_k.$$

Writing $m = \sum_{k=1}^{B} I_k$, summation yields

$$B^* = \sum_{k=1}^{B} (\Phi_{k-1} - \Phi_k) = B - m, \qquad B = B^* + m.$$

Eligible schedules have $m = 0$ and attain B^*.

For the stochastic case, attach independent Bernoulli sequences to each task and stage. A nonanticipating schedule reveals the next unused entry only when that stage is attempted. Let U_z, V_z be the numbers of eligible model and policy trials required to exhaust the two quotas. Geometric waiting times give

$$\mathbb{E}U_z = w_z/p_z, \qquad \mathbb{E}V_z = a_z/q_z.$$

Every completing eligible schedule reveals exactly these prefixes, irrespective of interleaving. If C is its total cost, then

$$C = \sum_z (c_z U_z + d_z V_z), \qquad \mathbb{E}C = \sum_z \left(\frac{c_z w_z}{p_z} + \frac{d_z a_z}{q_z} \right) = B^*.$$

Ineffective trials neither reveal a productive entry nor change a quota; their costs add pathwise.

Thus order inside an eligible frontier need not be uniquely optimal: the theorem characterizes a class of schedules with identical completion cost in this abstraction. It does not prove dominance over every scalar score, because a scalar rule can itself respect precedence. Cross-task transfer, warm-start benefits of early policy training, uncertain learning curves, forgetting, and model updates that never improve certification all change the allocation problem.

If every harmful false promotion in this abstraction wastes at most c units of work, Theorem 7 bounds its expected wasted work through round K by $c \sum_{k=1}^{K} \varepsilon_k$. This is the cost of erroneous certification alone. Conservative rejection, audit cost, and ordinary optimization failure are additional costs and are not hidden inside that bound.

G Controlled Learned-World-Model Experiment

G.1 Experimental execution

We evaluate action-conditioned predictors learned from sampled experience in finite-horizon worlds. Sparse, chain, and grid families each use 16 states, four actions, known support, bounded shared rewards, and horizons $H \in \{4, 8, 12\}$. Forty development worlds fix the protocol, 80 independent worlds supply the finite-world audit quantities, and 240 held-out worlds are used once for evaluation. World identifiers and random streams are disjoint across these roles. The predictor is the Beta-smoothed empirical transition kernel fitted from nested row observations; a separately acquired 32-query-per-row model defines the fixed reference behavior. Every gate sees only the learned predictor and its permitted audit evidence. Exact simulator returns are revealed only after routing and are never used to choose a proposal, a gate, or an evidence location.

The protocol has three complementary components. First, 400 paired constructions share the learned model and passive record but differ on one unobserved action, isolating whether intervention evidence identifies the otherwise hidden world effect. Second, held-out planning and teaching requests compare unconditional positive-imagination acceptance, uncertainty-only rejection, and decision-relative qualification; the same requests measure value error, update-direction cosine, harmful-use rate, and realized gain. Third, online runs compare ungated, uniformly gated, and decision-directed evidence acquisition at an equal real-transition budget. Candidate proposals, reference behavior, and evaluation streams are matched within every comparison, so the only manipulated object is whether the world-model prediction is qualified and where new evidence is acquired.

Two theory-directed supplements reuse the same finite-world row-audit primitive. For evidence-complexity scaling, the audited Bernoulli row has means $1/2 - \gamma$ and $1/2 + \gamma$ in the two possible worlds, with 13 logarithmically spaced margins $\gamma \in [0.02, 0.20]$. At each margin we enumerate sample sizes and report the smallest $n_{0.9}$ for which the exact randomized Neyman–Pearson test reaches 90% power while controlling false promotion at $\alpha = 0.05$; the log–log slope is fitted once across all registered margins. For evidence reuse, we fix $\gamma = 0.08$, audit all eight transition rows, and evaluate $K \in \{10, 20, 50, 100, 200\}$ later comparisons. The shared rule pays for one simultaneous α-level audit; the re-audit baseline pays for K fresh tests, each assigned error α/K and separately sized to the same 90% power. Consequently both procedures control familywise false promotion, while their real-query costs are measured under identical accuracy requirements. Exact integer results and the complete plotting records are retained with the experiment archive.

Theory-to-measurement correspondence. The finite-world design gives each theoretical statement an observable counterpart. Paired worlds implement passive non-identifiability and interventional recovery (Theorem 1 and Proposition 2); attribution accuracy and model-effect error are evaluated before any decision gate is scored. Planning curves test decision reliability and the separation between predictive accuracy and decision safety (Theorem 2 and Corollary 2) through realized

gain, harmful-use rate, and risk–coverage. Online runs instantiate the closed-loop and adaptive-acquisition statements (Theorem 3 and Theorem 7), while gradient and teaching diagnostics test gradient reliability and local safe improvement (Theorem 5 and Corollary 6). The log–log experiment tests the evidence lower-bound order (Proposition 10), and the reuse experiment tests amortized verification and precedence accounting (Proposition 9 and Theorem 8). This one-to-one mapping makes the appendix a validation suite for the theory rather than a collection of unrelated plots.

G.2 Results and supplementary summaries

The results separate the three theoretical claims cleanly. The paired construction reaches 100% attribution accuracy at the largest intervention budget, while model-effect error decreases from 0.08999 to 0.01888 (Figure 7). On held-out planning requests, accepting every positive imagined advantage produces 5.10% harmful uses; decision-relative qualification reduces this to 0.30% while preserving positive aggregate gain (Figures 8–9). The update-direction supplement raises mean true/imagined gradient cosine from 0.761 to 0.9994 as evidence increases. In the online study, ungated, uniform-gated, and decision-directed acquisition incur 44, 10, and 4 harmful updates, respectively (Figures 10–12). The evidence-complexity fit has slope 2.014 on the log–log scale, matching the predicted γ^{-2} order. At $K = 200$ downstream comparisons, one shared audit uses 664 real queries versus 350,400 for fresh re-auditing, a $527.7\times$ reduction under the same registered familywise error and power targets.

These conclusions do not rely on post-hoc selection: all registered worlds and margins are retained, gates never observe exact evaluation returns, and paired methods reuse identical stochastic streams. The simultaneous theorem-derived certificate can abstain completely at small budgets, and adaptive acquisition is not uniformly superior in every cell; we report these limitations rather than replacing them with a favorable subset. The experiment therefore checks the finite-world assumptions and scaling laws directly, while the agent benchmark in Appendix H separately tests whether the same verify-then-promote structure remains useful without claiming a finite-state certificate.

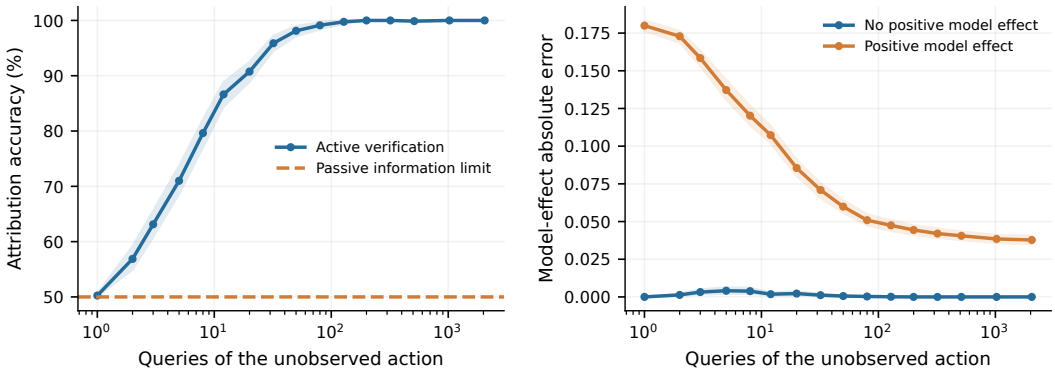

Figure 7: Attribution and model-effect error under increasing intervention evidence. Active action-conditioned queries resolve worlds that are indistinguishable under the shared passive record.

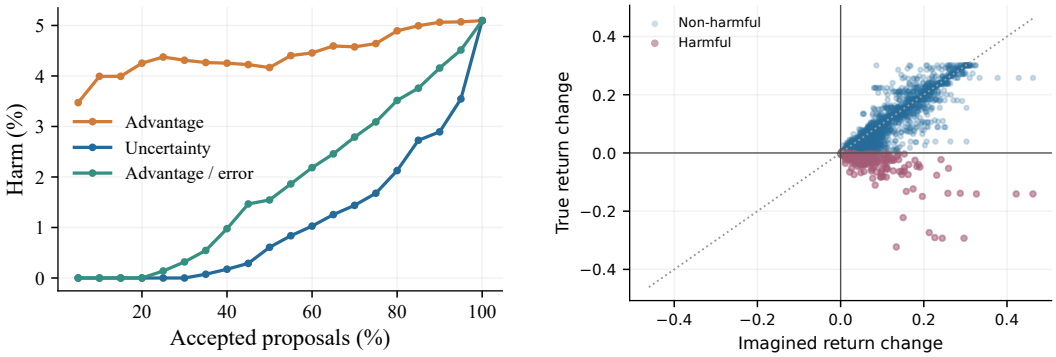

Figure 8: Decision-relative planning qualification. The risk–coverage curve (left) and imagined-versus-realized changes (right) show that screening the proposed use, rather than trusting positive imagination alone, removes most harmful promotions.

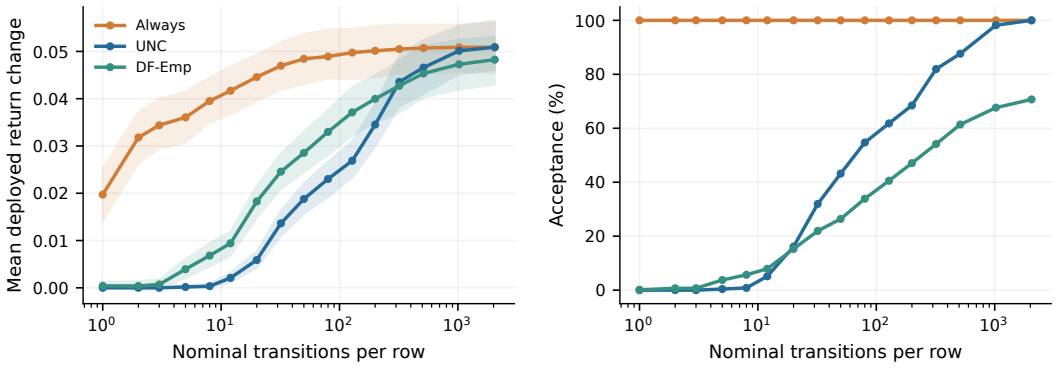

Figure 9: Planning gain and promotion rate as action-conditioned evidence increases. Reliability improves without changing the candidate requests or reference policy.

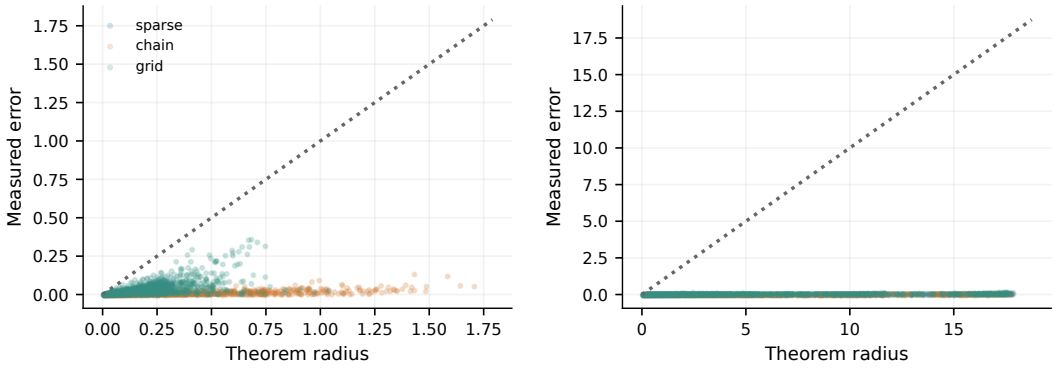

Figure 10: Certificate diagnostics. Measured decision errors remain below their theorem-derived radii over the registered finite-world requests.

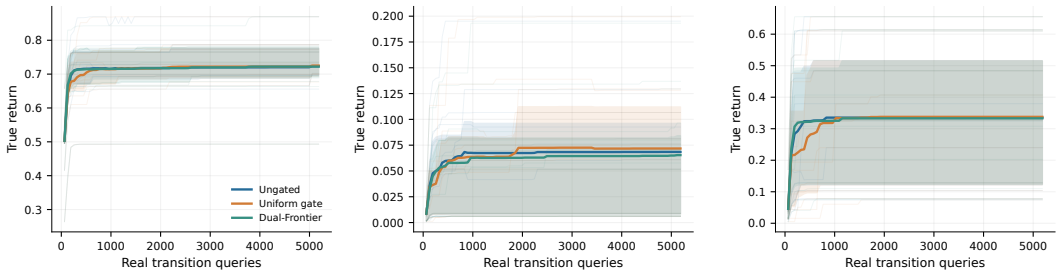

Figure 11: Online qualification and acquisition under equal real-transition budgets. Decision-directed evidence reduces harmful updates relative to ungated and uniform alternatives.

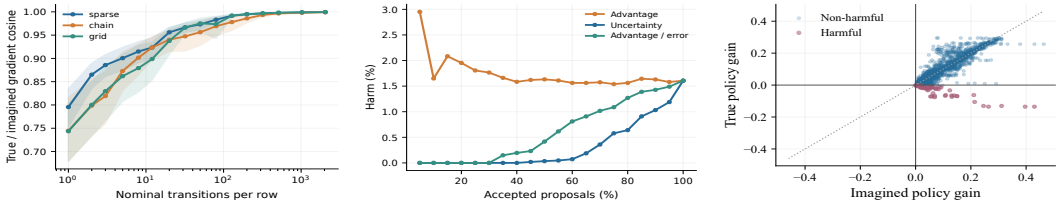

Figure 12: Transfer diagnostics for update-direction alignment (left), teaching risk–coverage (center), and imagined-versus-realized teaching gains (right).

H AGENT–WORLD-MODEL BENCHMARK EVALUATION

The benchmark study evaluates the Dual-Frontier admission rule of Section 5 under a language-world-model interface. We fix Qwen-AgentWorld as the learned world model, vary the agent backbone between Llama-3.1-8B-Instruct and Qwen3-8B, and evaluate BFCL (Patil et al., 2025), API-Bank (Li et al., 2023), and NexusRaven (Srinivasan et al., 2023). For each benchmark, the request manifest is partitioned once into a 20% calibration split and an 80% held-out test split. The partition is fixed before calibration and routing evaluation and is shared across all compared methods; the same request partition is used for both agent backbones.

Each request defines a reference–candidate comparison before admission: the base-agent output is the reference and a fixed Qwen-AgentWorld proposal is the candidate. Base actions and world-model generations are produced once and cached, and every non-Agent-only routing rule acts on the same candidate records. Calibration requests are used only to construct the decision-relevant error radius entering the Dual-Frontier margin. Their benchmark references are accessed only after all corresponding model-side predictions have been frozen. Test references are never used in candidate construction, calibration, or admission and are loaded only after the routed test outputs are fixed. Prompts, tool schemas, sampling budgets, parsers, generation order, and candidate-generation records are otherwise shared across routing rules.

H.1 DECISION RULES AND CALIBRATED DUAL-FRONTIER INSTANTIATION

Predicted advantage and estimation uncertainty. Once the shared proposal c_i^K is fixed, the world model evaluates its consequence relative to the reference action c_i^0 using a fixed repeated-evaluation budget. Let $\widehat{\gamma}_{i\ell} \in [-1, 1]$, $\ell = 1, \dots, L$, denote the resulting normalized candidate-minus-reference consequence scores. We define

$$S_i = \frac{1}{L} \sum_{\ell=1}^{L} \widehat{\gamma}_{i\ell}, \qquad \xi_i = \sqrt{\frac{2 \log(1/\delta)}{L}}, \tag{31}$$

where the same L and δ are used throughout evaluation. Thus S_i is the benchmark-side predicted advantage and ξ_i accounts for finite-generation estimation uncertainty.

Decision-relevant model-error calibration. The repeated world-model records additionally provide proposal-level reliability information. We collect these quantities in

$$z_i = \big(1 - \bar{\rho}_i,\, 1 - \kappa_i,\, h_i^{\max}\big), \qquad \psi_i = g(z_i), \tag{32}$$

where g is a fixed nonnegative aggregation rule specified before calibration and used unchanged at test time. The individual confidence, agreement, and consequence-risk signals therefore contribute to the model-error estimate rather than acting as separate Dual-Frontier admission thresholds.

For each calibration request j, all model-side quantities are frozen before the benchmark evaluator provides the normalized candidate-minus-reference contrast Γ_j^{eval}. We calibrate the residual decision discrepancy not already accounted for by ξ_j,

$$e_j^B = \big[\big|S_j - \Gamma_j^{\mathrm{eval}}\big| - \xi_j\big]_+, \qquad q_j = e_j^B - \psi_j. \tag{33}$$

For a calibration set of size n_{cal}, let

$$k = \lceil (n_{\mathrm{cal}} + 1)(1 - \alpha) \rceil, \qquad B_i = \big[\psi_i + q_{(k)}\big]_+, \tag{34}$$

with $q_{(k)}$ the corresponding calibration quantile. Calibration is performed only on the designated 20% split; B_i is then evaluated without test labels on the remaining requests. Under the exchangeability condition of Proposition 7, the same rank argument gives

$$\Pr\big\{e_i^B > B_i\big\} \le \alpha. \tag{35}$$

Dual-Frontier promotion rule. The benchmark implementation uses the same three-term margin as Eq. (17):

$$T_i^{\mathrm{DF}} = S_i - B_i - \xi_i. \tag{36}$$

A world-model proposal is promoted only when $T_i^{\mathrm{DF}} > 0$. On the calibration-coverage event in (35),

$$\Gamma_i^{\mathrm{eval}} \geq S_i - B_i - \xi_i = T_i^{\mathrm{DF}}, \tag{37}$$

so a positive benchmark margin has exactly the lower-bound semantics required by the Dual-Frontier decision rule. This construction separates predicted benefit, decision-relevant model error, and finite-generation estimation uncertainty while keeping the practical gate identical in form to the theoretical one.

Six routing rules. Let $a_i^r \in \{0, 1\}$ denote whether rule r admits the shared proposal c_i^K. If no valid proposal exists, all non-Agent-only rules fall back to c_i^0. For the three single-signal comparison rules, we retain fixed thresholds

$$\tau_\rho = 0.70, \qquad \tau_\kappa = 3/7, \qquad \tau_h^P = 0.22, \tag{38}$$

used only to define the corresponding baselines. The routing decisions are

$$\begin{aligned} &a_i^A = 0, && a_i^W = 1, && a_i^C = \mathbb{I}\{\bar{\rho}_i \geq \tau_\rho\}, \\ &a_i^K = \mathbb{I}\{\kappa_i \geq \tau_\kappa\}, && a_i^P = \mathbb{I}\{h_i^{\max} \leq \tau_h^P\}, && a_i^D = \mathbb{I}\{T_i^{\mathrm{DF}} > 0\}. \end{aligned} \tag{39}$$

Here A, W, C, K, P, D denote Agent-only, Always-WM, Confidence, Consistency, Pessimistic, and Dual-Frontier, respectively. The candidate is identical across W, C, K, P, D; only the admission rule changes. The routed output is

$$c_i^r = \begin{cases} c_i^K, & a_i^r = 1, \\ c_i^0, & a_i^r = 0. \end{cases} \tag{40}$$

The proposal-generation protocol, uncertainty construction, calibration level, and routing rules are fixed before evaluation of the held-out test split. Calibration references are confined to the designated calibration requests, and test references are inaccessible until (40) has been frozen. Every held-out request is retained in evaluation, including fallbacks and malformed outputs. Oracle is excluded because it observes reference outcomes and is not deployable.

H.2 Metrics

For request i and routing rule r, let c_i^0 denote the base-agent output and c_i^r the frozen routed output in (40). We use the fixed benchmark-aware evaluator of each benchmark throughout. Let $y_i(c) \in \{0, 1\}$ denote strict task success, $p_i(c) \in [0, 1]$ parameter-level accuracy, $u_i(c) \in [0, 1]$ the decision-quality score assigned to output c, and $q_i(c_i^0, c) \in [0, 1]$ the corresponding reliability-loss score relative to the base action. All evaluator definitions are fixed before routing-rule replay.

Evaluation follows the native structural conventions of each benchmark. BFCL v4 preserves ordered and parallel-call multiplicity and validates function names and arguments against the supplied schema. API-Bank matches the required API identity and its normalized parameter dictionary. NexusRaven canonicalizes its Python-style function expression before comparing function identity, argument names, values, and multiplicity. Thus the metrics below share a common form while retaining the native call semantics of each benchmark.

Task success. The strict task-level metric is

$$\mathrm{TS}_r = \frac{1}{N} \sum_{i=1}^{N} y_i(c_i^r). \tag{41}$$

It requires the complete function-call sequence, including function names, arguments, and call multiplicity, to satisfy the benchmark reference.

Parameter accuracy. We report the average parameter-level score

$$\mathrm{PA}_r = \frac{1}{N} \sum_{i=1}^{N} p_i(c_i^r), \tag{42}$$

where $p_i(\cdot)$ follows the corresponding benchmark's native parameter matching and normalization rules.

Net decision gain. To measure the signed effect of routing through the learned world model, we use

$$\mathrm{NDG}_r = \frac{1}{N}\sum_{i=1}^{N}\left[u_i(c_i^r) - u_i(c_i^0)\right]. \tag{43}$$

Positive values indicate that the admitted world-model revisions improve the average decision quality relative to the base agent, whereas negative values indicate net degradation. We report NDG in percentage points.

Harmful revisions. Let $h_i(c_i^0, c_i^r) \in [0,1]$ denote the evaluator's degradation score for the routed output relative to the base action. We report

$$\mathrm{HRR}_r = \frac{1}{N}\sum_{i=1}^{N} h_i(c_i^0, c_i^r). \tag{44}$$

This quantity measures the overall exposure to harmful interventions, including both their occurrence and decision-level severity.

Revision coverage. The fraction of requests on which rule r admits a world-model revision is

$$\mathrm{RC}_r = \frac{1}{N}\sum_{i=1}^{N} a_i^r. \tag{45}$$

Coverage is descriptive rather than an objective by itself and should be read jointly with the reliability metrics.

Selective risk. Among admitted revisions, we measure the residual reliability loss as

$$\mathrm{SR}_r = \frac{\sum_{i=1}^{N} a_i^r\, q_i(c_i^0, c_i^r)}{\sum_{i=1}^{N} a_i^r}, \qquad \sum_i a_i^r > 0. \tag{46}$$

Selective risk is undefined when a rule never revises. The pair $(\mathrm{RC}, \mathrm{SR})$ therefore provides the empirical risk–coverage view of the verify-then-promote decision rule.

TS, PA, and NDG are higher-is-better metrics, whereas HRR and SR are lower-is-better; RC is descriptive. Here N always denotes requests in the held-out 80% test split; calibration requests are excluded from every reported benchmark metric. The reported Avg. column is the equal-weight arithmetic mean of the three benchmark percentages rather than a pooled micro-average.

H.3 Additional results

The component ablation removes one term at a time from the same Dual-Frontier decision margin while keeping the shared proposal, calibrated quantities, and cached world-model records fixed. Advantage-only admits when $S_i > 0$; w/o world-model error uses $S_i - \xi_i > 0$; w/o estimation error uses $S_i - B_i > 0$; and full Dual-Frontier uses

$$S_i - B_i - \xi_i > 0.$$

The comparison therefore isolates the contribution of predicted advantage, decision-relevant model error, and finite-generation estimation uncertainty without changing proposal generation or test requests. Table 6 reports the Qwen3-8B results; the corresponding Llama-3.1-8B-Instruct results appear in Table 2.

The complete margin is consistently strongest. Relative to Advantage-only, Dual-Frontier improves the equal-weight Success/Acc. averages by $7.05/4.34$ points for Llama-3.1-8B-Instruct and $6.27/3.92$ points for Qwen3-8B. Removing either uncertainty term recovers only part of this gain, while their joint use yields the strongest performance for both backbones. This pattern is consistent with Section 5: predicted benefit identifies potentially useful interventions, whereas B_i and ξ_i determine whether that apparent advantage remains sufficiently supported for promotion.

Table 6: Additional Qwen3-8B ablation results. Success and Acc. denote task success and parameter accuracy; Avg. is the uniform mean over BFCL v4, API-Bank, and NexusRaven.

Models	Methods	BFCL v4		API-Bank		NexusRaven		Avg.	
		Success	Acc.	Success	Acc.	Success	Acc.	Success	Acc.
Qwen3-8B	Advantage-only	67.80	84.94	95.07	96.41	94.40	95.71	85.76	92.35
	w/o world-model error	72.32	88.01	96.88	97.50	96.02	96.68	88.41	94.06
	w/o estimation error	77.79	91.18	98.40	98.31	97.11	97.52	91.10	95.67
	Dual-Frontier	**79.38**	**92.20**	**98.92**	**98.64**	**97.80**	**97.96**	**92.03**	**96.27**

Coverage and conditional reliability. Tables 7 and 8 report how often each deployable rule admits a world-model revision and the residual risk among those admitted revisions. Agent-only is omitted because it never revises; consequently, its selective risk is undefined rather than zero. Revision coverage is descriptive, whereas lower selective risk is better. Their joint reading is essential: Dual-Frontier deliberately operates at moderate coverage while removing most unsafe promotions, exactly the selective qualification behavior predicted by the theory.

Across the three benchmarks, Dual-Frontier revises 57.59% of Llama and 57.51% of Qwen requests. At this nontrivial coverage, its average selective risk is 28.20% and 18.23%, respectively, versus 67.68% and 78.20% for the strongest competing selective baseline. The reductions of 39.48 and 59.97 percentage points explain why the primary-metric gains are not a consequence of indiscriminate revision: the method rejects precisely the candidate groups most likely to erase a correct base decision. Always-WM provides the opposite endpoint, with full coverage but selective risks of 78.89% and 86.29%.

Table 7: Revision coverage (%) across agent backbones and benchmarks. Avg. is the uniform mean over BFCL v4, API-Bank, and NexusRaven.

Models	Methods	BFCL v4	API-Bank	NexusRaven	Avg.
Llama-3.1-8B Instruct	Always-WM	100.00	100.00	100.00	100.00
	Confidence	99.50	99.61	98.74	99.28
	Consistency	80.88	84.65	83.33	82.95
	Pessimistic	82.25	85.43	79.87	82.52
	Dual-Frontier	55.62	61.81	55.35	57.59
Qwen3-8B	Always-WM	100.00	100.00	100.00	100.00
	Confidence	99.62	99.61	99.69	99.64
	Consistency	78.25	81.30	81.45	80.33
	Pessimistic	82.50	79.53	84.91	82.31
	Dual-Frontier	53.62	57.28	61.64	57.51

Table 8: Selective risk (%) among admitted world-model revisions. Lower is better; Avg. is the uniform mean over the three benchmarks.

Models	Methods	BFCL v4	API-Bank	NexusRaven	Avg.
Llama-3.1-8B Instruct	Always-WM	93.12	73.43	70.13	78.89
	Confidence	90.95	66.21	57.64	71.60
	Consistency	89.49	61.86	51.70	67.68
	Pessimistic	90.43	65.67	56.69	70.93
	Dual-Frontier	**9.12**	**33.66**	**41.82**	**28.20**
Qwen3-8B	Always-WM	97.88	79.53	81.45	86.29
	Confidence	96.99	72.53	76.03	81.85
	Consistency	96.49	67.07	71.04	78.20
	Pessimistic	96.97	72.52	75.56	81.68
	Dual-Frontier	**3.43**	**27.36**	**23.90**	**18.23**

AI USE STATEMENT

Generative AI tools were used solely for formatting checks and language polishing. The authors reviewed all resulting revisions and take full responsibility for the final manuscript.