WHEN CAN TEST-TIME SEARCH BE SKIPPED? SUCCESSOR-CHANNEL CRITICALITY FOR ACTION-CHUNKED POLICIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Test-time search can improve action-chunked robot policies, but running it at every decision is expensive and often does not change the outcome. We ask when search can be skipped. Our answer uses the policy's own candidate actions: if every candidate leads to a state with a similar capacity for future control, search has little to choose between. We measure *successor-channel criticality* as the variance across candidate actions of a future-controllability score. The score summarizes the spread of object positions reached by subsequent policy rollouts at a fixed resolution, and is computed in a resettable simulator without task reward or search labels. We prove that low criticality bounds the value of search, provided that value differences the score does not capture are also small. We evaluate gates with paired per-state interventions on fourteen datasets from three benchmark suites and two policy classes; criticality's ranking lies above the matched random band on all fourteen. Skipping search where estimated criticality is zero avoids about half of decisions while giving up 19.9% of the net benefit, less than the mean of the same score or a cheap rule based on episode time and sampled gripper disagreement. Distilled into a small network and added to that rule, a criticality head raises success on 600 held-out episodes by 4.2 points with 20.5% fewer search calls on a LIBERO image policy, and by 7.0 points at similar call counts on Stack. A network trained on the mean score also improves on that rule; on the image policy, averaged over three head seeds, the criticality head reaches 2.2 points higher success at similar calls when search is scarce and similar success with fewer calls at $p = 0.25$.

1 INTRODUCTION

Action-chunked policies predict several actions at each decision (Zhao et al., 2023; Chi et al., 2023; Black et al., 2025). Test-time search can improve these choices by sampling alternative chunks and evaluating them before acting. Its value, however, depends on the state. While the robot moves through free space, many chunks lead to equally good outcomes; near contact, small differences can decide whether the task succeeds. Searching at every decision therefore spends much of its computation on decisions where it has little expected benefit. A gate that decides when to search can remove this waste, provided it can tell the two kinds of state apart.

Existing gates use signals such as policy confidence, disagreement among samples, task difficulty or episode phase, and they are usually judged by final task success. Final success, however, mixes the quality of the signal with the policy, the search procedure and the budget. We instead measure the effect of search at each state directly: we run each probe state twice, once with search and once without, under shared random numbers, and record the difference in outcome. Measured this way, the usefulness of common signals is not stable. Ten of fourteen signals point in the right direction on some of our datasets and in the wrong direction on others, and one sampling-based signal changes direction when only the policy is retrained (Appendix C.1).

We therefore look at what the policy's candidate actions lead to. If every candidate chunk leads to a state from which the policy can reach a similar set of object configurations, the choice among them matters little and search can be skipped. If the candidates lead to states that differ in this respect, the

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

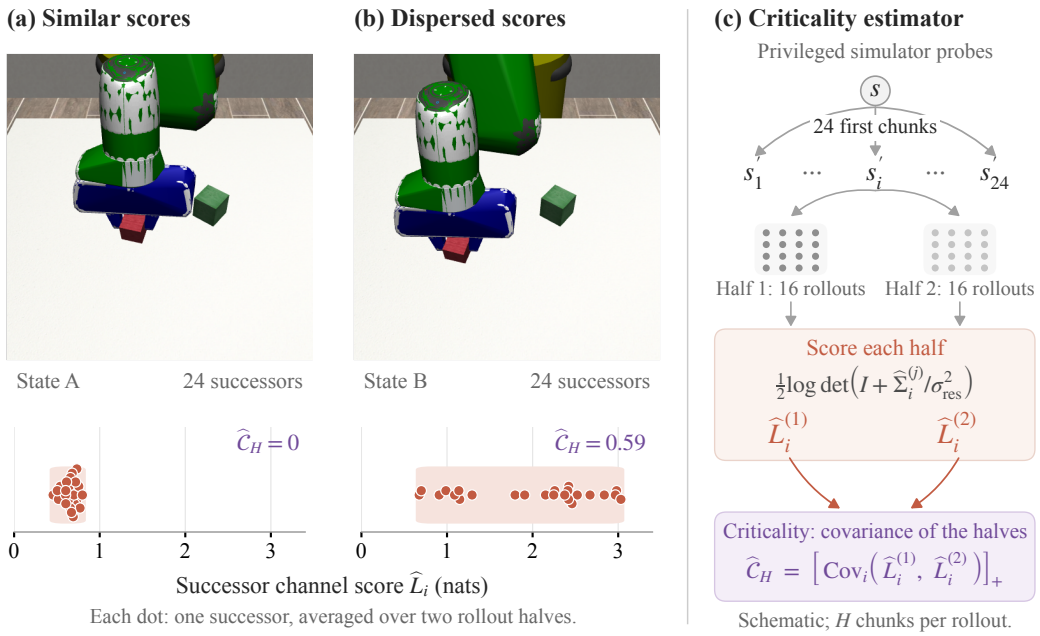

Figure 1: **The idea and the estimator.** (a,b) Two recorded Stack states with all 24 successor scores, each averaged over two rollout halves. State A has a narrow score spread and a clamped cross-half estimate of zero; State B has dispersed scores and positive estimated criticality. (c) Independent half estimates of the channel score (Eq. 4) are combined by cross-covariance (Eq. 6), correcting additive sampling-noise inflation (Proposition B.1).

choice matters. We call the variance of this reachability score across candidate actions *successor-channel criticality* (Figure 1, Eq. 5). It is computed from policy rollouts in a resettable simulator, without task reward or labels of search benefit. We prove that low criticality bounds the value of search, provided that value differences the score does not capture are also small (Section 5).

Three findings support the approach. First, skipping search where estimated criticality is zero avoids about half of all decisions while giving up 19.9% of the net benefit, compared with 25.8% for the mean of the same score and 23.5% for a cheap time-and-gripper rule. Second, criticality's ranking lies above the matched random band on all fourteen datasets, and a network distilled from it ranks states far better than one distilled from the mean (0.30 against 0.07 of the oracle gap on Stack). Third, in closed loop, adding the distilled network to a cheap gate raises success by 4.2 points on a LIBERO image policy and by 7.0 points on Stack. A network distilled from the mean also helps; on the image policy the criticality head adds success at the lowest budget and saves calls at $p = 0.25$ (Section 7). We use the estimated zero set to evaluate offline when search can be skipped, and distill the continuous score to allocate search online.

Contributions. (i) A reward-free signal for selective search, successor-channel criticality, with a split-half estimator that corrects additive sampling-noise inflation. (ii) A bound linking low criticality to low search value, with the conditions under which it holds stated explicitly. (iii) An evaluation protocol that measures the effect of search at each state, and experiments across three benchmark suites, two policy classes and two held-out deployments.

2 Related Work

Adaptive computation and metareasoning. Adaptive computation lets a model decide how much to compute for each input (Graves, 2016). For VLA policies, DeeR-VLA exits a multimodal model early (Yue et al., 2024); AC^2-VLA and ElegantVLA reuse computation based on action context and on temporal and motion cues (Yu et al., 2026; Li et al., 2026b); Act-Think-Abstain routes on task complexity (Izzo et al., 2026); and KV-cache gating recomputes when action confidence drops

(Wu et al., 2026). Closer to our setting, τ_0-VLA searches over subtask alternatives when needed (Cai et al., 2026), VLA-ATTC gates candidate generation using uncertainty (Li et al., 2026a), DIRECT allocates reasoning resources from scene context (Dao et al., 2026), VLA-Corrector triggers replanning from visual dynamics deviation (Pan et al., 2026), and in language models SEAG gates tree search on answer confidence (Lee et al., 2025). We evaluate allocation signals against a fixed operator's per-state causal benefit, separately from system-level success. This is metareasoning: computation has value through improved decisions, net of cost (Russell & Wefald, 1991; Horvitz, 1987; Dean & Boddy, 1988; Hay et al., 2012). DIAL learns sparse utility gates from counterfactual exploration after observing signal-direction reversals (Li et al., 2026c). Our teacher instead uses successor-channel structure without search-benefit labels.

Test-time selection. RoboMonkey and V-GPS improve VLA policies by sampling candidate actions and selecting one with a learned verifier or value function (Kwok et al., 2025; Nakamoto et al., 2025), and BID selects chunks through guided test-time sampling (Liu et al., 2025). Our search operator has the same sample-and-select form; we study when to invoke it. RGSC selects high-regret states for AlphaZero training using outcome supervision (Tsai et al., 2026), which our teacher does not need.

Action–future information and selective prediction. Empowerment is the capacity of the action–future information channel, maximized over the action distribution (Klyubin et al., 2005; 2007; Jung et al., 2011; Salge et al., 2014; Mohamed & Rezende, 2015). We fix the source to the current policy and distinguish the score at the current state from its variation across candidate successors. Causal action influence instead measures action effects on future observations (Seitzer et al., 2021). The veto objective resembles the coverage–risk tradeoff of selective prediction (Chow, 1970; Geifman & El-Yaniv, 2017).

3 Problem Setting and Causal Evaluation

Action-chunked policy and search. Let π be a frozen policy that emits an action chunk $A_t^k = (u_t, \ldots, u_{t+k-1}) \sim \pi(\cdot \mid s_t)$ at each decision step. A fixed search operator $\mathcal{S}$ may be invoked at a chunk boundary: it samples candidate chunks from π and selects one using an evaluator. A gate $g(s) \in \{0, 1\}$ decides whether search is invoked, with each call incurring cost c. The search operator is part of the problem definition; changing its candidate budget, evaluator, or continuation protocol changes the allocation problem itself.

Operator-specific causal search benefit. Let J denote the realized return from state s to episode termination. We define

$$\Delta_{\mathcal{S}}(s) = \mathbb{E}[J \mid \mathrm{do}(g = 1), s, \mathcal{S}] - \mathbb{E}[J \mid \mathrm{do}(g = 0), s, \mathcal{S}], \tag{1}$$

and, for an exchange rate $\lambda \geq 0$ between task return and computation,

$$\mathrm{VOC}_{\mathcal{S}}(s) = \Delta_{\mathcal{S}}(s) - \lambda c. \tag{2}$$

The intervention notation emphasizes that the gate is actively changed rather than merely observed. Thus $\Delta_{\mathcal{S}}$ measures the realized effect of invoking this particular search operator at this state; it is not a universal value of search.

Paired counterfactual labels. At each probe state, we estimate $\Delta_{\mathcal{S}}(s)$ using paired *single-switch* interventions. One branch invokes search at the probe state and the other does not, after which both are rolled to episode completion under the same continuation protocol and coupled with common random numbers. The resulting label measures realized return change, rather than merely how much search changes the proposed action or a critic's local value estimate.

These labels deliberately answer a local counterfactual question. They are not equivalent to executing a gate throughout an episode: an earlier decision to search can change the states encountered later. We therefore evaluate transfer from the single-switch surrogate to closed-loop gating separately in Section 7.

Matched-coverage ranking. A scalar score $r(s)$ ranks the probe states. At search coverage p, let $\mathcal{B}_r(p)$ denote the normalized weighted search benefit from selecting the top-p region (Appendix D.2). We compare it with matched random allocation and an oracle reference cross-fitted over paired rollout repeats at each state:

$$\mathrm{GapClosed}_r(p) = \frac{\mathcal{B}_r(p) - \mathcal{B}_{\mathrm{rand}}(p)}{\mathcal{B}_{\mathrm{oracle}}(p) - \mathcal{B}_{\mathrm{rand}}(p)}. \tag{3}$$

Random allocation therefore scores 0 and the held-out oracle reference scores 1. Values below 0 or above 1 are possible because the reference is cross-fitted rather than an in-sample upper bound. With $K = 12$ repeats, selection uses eleven and scoring uses the held-out repeat, averaged over all twelve splits. Uncertainty is estimated separately, by resampling episodes.

Veto evaluation. For state i, let w_i be its inverse inclusion probability and $\widehat{\Delta}_i$ its paired single-switch estimate. For veto set V, we report the weighted share of decisions skipped, $q(V)$, and the share of net benefit given up, $\ell(V)$:

$$q(V) = \frac{\sum_i w_i \mathbf{1}\{i \in V\}}{\sum_i w_i}, \qquad \ell(V) = \frac{\sum_i w_i \widehat{\Delta}_i \mathbf{1}\{i \in V\}}{\sum_i w_i \widehat{\Delta}_i}.$$

The signed ratio ℓ permits cancellation of beneficial and harmful search effects; it is not the fraction of positive opportunities missed.

4 Successor-Channel Criticality

Section 3 defines the operator-specific quantity that a search gate should predict. Rather than regressing directly on search benefit, we ask a narrower structural question: do the action chunks proposed by the current policy preserve different amounts of future controllability?

4.1 A policy-conditioned action–future channel

Let the simulator state decompose as $z(s) = [z^{\mathrm{robot}}(s), z^{\mathrm{world}}(s)]$. We construct the channel over the task-relevant world component z^{world}, excluding robot configuration so that robot self-motion does not dominate the score.

At state x, draw H consecutive policy chunks $A^k_{1:H} \sim \pi$. Under deterministic simulator transitions, their world endpoint is $W_H(x) = z^{\mathrm{world}}(F_H(x, A^k_{1:H}))$. Write $\Sigma_{W_H}(x) = \mathrm{Cov}[W_H(x) \mid x]$. With independent resolution noise, $Y_H(x) = W_H(x) + \xi$, $\xi \sim \mathcal{N}(0, \sigma^2_{\mathrm{res}} I)$, and the channel information $\mathcal{I}^{k,H}_\pi(x) = I(A^k_{1:H}; Y_H(x) \mid x)$.

The reachable endpoints need not be Gaussian. By the Gaussian maximum-entropy bound (derivation in Appendix A.3),

$$\mathcal{I}^{k,H}_\pi(x) \le \bar{\mathcal{I}}^{k,H}_\pi(x) = \frac{1}{2} \log\det\left(I + \frac{\Sigma_{W_H}(x)}{\sigma^2_{\mathrm{res}}}\right). \tag{4}$$

We call $\bar{\mathcal{I}}^{k,H}_\pi$ in Eq. 4 the *policy-conditioned channel score*. It is not classical empowerment capacity: the source distribution is fixed to the current policy rather than optimized, and the two quantities are not ordered. Logarithms are natural. Intuitively, $\bar{\mathcal{I}}^{k,H}_\pi$ is a resolution-regularized log-volume of the object positions the policy can reach in H chunks, in units of σ_{res}: it grows when the world endpoints vary across sampled action sequences and is zero when all sampled sequences reach the same world endpoint.

4.2 Successor-channel criticality

A large current-state level does not by itself make the current decision search-critical. Many futures may be reachable while every candidate action preserves similar downstream controllability. For a first chunk $A^k_1 \sim \pi(\cdot \mid s)$, let

$$s'_{A_1} = F_1(s, A^k_1).$$

We define successor-channel criticality as

$$\mathcal{C}_H(s) = \mathrm{Var}_{A_1^k \sim \pi(\cdot|s)} \left[\bar{\mathcal{I}}_\pi^{k,H}(s'_{A_1}) \right]. \tag{5}$$

The comparison signals are current-state level $L_H(s) = \bar{\mathcal{I}}_\pi^{k,H}(s)$ and successor mean $M_H(s) = \mathbb{E}_{A_1^k \sim \pi}[\bar{\mathcal{I}}_\pi^{k,H}(s'_{A_1})]$.

The successor mean uses the same candidate successors and channel lookahead as $\mathcal{C}_H$, making it the direct mean-versus-variance control.

The channel horizon starts at its input state: L_H reaches H chunks from s, whereas $\mathcal{C}_H$ first executes a candidate chunk and then probes H more. Thus its effective depth is $H+1$; L_{H+1} is the depth-matched current-state comparator, and M_H is matched by construction.

Monte-Carlo estimator. A naive variance of estimated successor-channel scores conflates genuine successor variation with Monte-Carlo estimation noise. We therefore split the channel rollouts at each successor into two independent halves with estimates $\widehat{L}_i^{(j)}$ and use their cross-covariance:

$$\widehat{\mathcal{C}}_{H,\mathrm{raw}}(s) = \frac{1}{n_{\mathrm{outer}} - 1} \sum_i \left(\widehat{L}_i^{(1)} - \overline{L}^{(1)} \right) \left(\widehat{L}_i^{(2)} - \overline{L}^{(2)} \right). \tag{6}$$

With $\overline{L}^{(j)}$ the mean over successors, we report

$$\widehat{\mathcal{C}}_H(s) = \max \left\{ \widehat{\mathcal{C}}_{H,\mathrm{raw}}(s), 0 \right\}. \tag{7}$$

We call the states with $\widehat{\mathcal{C}}_H(s) = 0$, where the raw cross-covariance is not positive, the *veto set*; skipping search there is the *veto*. Its membership depends on Monte-Carlo noise and finite-sample channel bias; Proposition B.1 separates these effects, and Section 6 evaluates the veto empirically.

From estimate to deployed gate. The estimate $\widehat{\mathcal{C}}_H$ serves two roles. Its veto set identifies candidate states for skipping, which we evaluate with paired interventions, and its continuous values serve as the targets of a privileged *teacher*. We distill these targets into a lightweight score h_θ using low-dimensional simulator observations and log-transformed targets (Appendix D.5). The teacher requires no counterfactual search-benefit or task-return labels. Online, calibrated combinations of the distilled score and cheap signals define a budget-dependent allocation rule. Section 7 evaluates this rule in closed loop, with training and cost details in Appendices D.5 and D.9.

5 One-Sided Bounds for Search Value

Search can improve a decision only if the candidate chunks differ in value. We show that this difference is bounded in terms of criticality and a residual term, the part of the value difference that the channel score does not see (Theorem 5.1). When both are small, search cannot pay for its cost (Corollary 5.2).

Search envelope. Let $Q(s,a)$ be expected return after chunk a and the fixed continuation. For $n \geq 2$ i.i.d. candidates $A^{(i)} \sim \pi(\cdot \mid s)$, define

$$G_n(s) = \mathbb{E}[\max_{i \leq n} Q(s, A^{(i)})] - \mathbb{E}_A Q(s,A).$$

An evaluator selecting from these candidates satisfies $\Delta_{\mathcal{S}}(s) \leq G_n(s)$. A current-state level bound additionally requires value to factor through the noisy world endpoint: Theorem A.1 gives $G_n \leq B \min\{c_n \sqrt{L_H}, \gamma_n\}$ under that condition, where $c_n = (n-1)/\sqrt{2(2n-1)} = c'_n/\sqrt{2}$ and $\gamma_n = \max_{u \in [0,1]}(u - u^n)$. The successor result below retains unexplained variation explicitly.

A bound through criticality. For a first chunk a, let $E_H(a) = \bar{\mathcal{I}}_\pi^{k,H}(F_1(s,a))$. It reaches depth $H+1$ from s, and $\mathcal{C}_H(s) = \mathrm{Var}[E_H(A)]$. At fixed s, define

$$g_H(e) = \mathbb{E}[Q(s,A) \mid E_H(A) = e], \qquad \rho_H^2(s) = \mathbb{E}[\mathrm{Var}(Q(s,A) \mid E_H(A))].$$

The *optionality defect* ρ_H measures value variation among candidates with the same channel score. It retains, for example, return differences between equally broad reachable sets in different regions.

Theorem 5.1 (Successor-channel decomposition)**.** *At each state s,*

$$\mathrm{Var}_{A\sim\pi}[Q(s,A)] = \mathrm{Var}_{A\sim\pi}[g_H(E_H(A))] + \rho_H^2(s). \tag{8}$$

If g_H is κ-Lipschitz on the score support, then

$$\mathrm{VOC}_{\mathcal{S}}(s) \le c'_n\sqrt{\kappa^2\mathcal{C}_H(s) + \rho_H^2(s)} - \lambda c, \qquad c'_n = \frac{n-1}{\sqrt{2n-1}}. \tag{9}$$

If $\mathrm{span}_a\, Q(s,a) \le B$, the right-hand side can be replaced by $\min\{c'_n\sqrt{\kappa^2\mathcal{C}_H + \rho_H^2}, B\gamma_n\} - \lambda c$.

The proof combines total variance, Lipschitz contraction and an expected-maximum inequality (Hartley & David, 1954) (Appendix A.5). Keeping the terms in quadrature is tighter than the additive relaxation $c'_n(\kappa\sqrt{\mathcal{C}_H} + \rho_H)$.

Corollary 5.2 (Low-criticality veto condition)**.** *Suppose $\kappa > 0$, $\rho_H(s) \le r$, and $\lambda c \ge c'_n r$. Then*

$$\mathcal{C}_H(s) \le \frac{(\lambda c/c'_n)^2 - r^2}{\kappa^2} \quad\Longrightarrow\quad \mathrm{VOC}_{\mathcal{S}}(s) \le 0.$$

At $\mathcal{C}_H(s) = 0$, $\mathrm{VOC}_{\mathcal{S}}(s) \le c'_n\rho_H(s) - \lambda c$.

This condition covers small positive criticality as well as exact zeros. It requires a residual bound; estimated scores additionally need error control (Appendix A.7). Section 6 measures what the veto gives up, while deployment uses a calibrated continuous score.

Why variation rather than mean? In a two-type successor model with scores $\mu \pm \eta$ and aligned values $q_0 \pm \beta\eta$, both equiprobable, the mean stays μ but $\mathcal{C}_H = \eta^2$ and $G_n = \beta\eta(1 - 2^{1-n})$. Thus search opportunity can vary at fixed channel mean, with the $\sqrt{\mathcal{C}_H}$ dependence of the bound (Appendix Figure 4). Empirical comparisons with M_H test whether this distinction is useful beyond the construction.

6 Empirical Evaluation: Ranking and Skipping

We ask two questions. Can a signal rank states by how much search helps there? And can it find states where search can be skipped at little cost? The first needs a good ordering of all states; the second needs only a reliable set of states to skip.

6.1 Experimental protocol

Ranking uses fourteen corpora from robosuite (Zhu et al., 2020), LIBERO (Liu et al., 2023) and Meta-World (Yu et al., 2020), all built on MuJoCo (Todorov et al., 2012), covering four task families (robosuite Stack and Lift, LIBERO object manipulation, Meta-World pick-place) and two policy classes: state-based behavior cloning and an image-conditioned flow-matching policy (Lipman et al., 2023). A *corpus* is one dataset of probe states with paired search labels, collected from one task and one policy. An *arm* scores a corpus with one estimator setting, mainly the channel horizon H; the veto analysis uses all thirty-one arms over seventeen corpora (Appendix D.1). Unless stated otherwise, $H = 2$ and search coverage $p = 0.25$. The search operator samples eight candidate chunks from the policy and picks one with a privileged rollout evaluator. Random-allocation bands are 90% ranges over 4,000 random draws, and paired differences use 95% episode-cluster bootstrap intervals (Appendix D; Agarwal et al., 2021).

6.2 Common signals change direction

Appendix Figure 5 scores fourteen gating signals against the same per-state benefit. Ten of them are better than random on some corpora and worse on others. Policy entropy and ensemble disagreement are not available for the image policy and are left out rather than set to zero. To isolate the policy, we evaluate a sample-disagreement cue adapted from VLA-ATTC on three policies that share task, architecture, chunk length and lookahead. It closes $+0.103$, -0.162 and -0.114 of the oracle gap, each outside its own random band: retraining the policy alone reverses the useful direction.

Table 1: **Summary of ranking and veto results.** Ranking rows count corpora where criticality is clearly better / worse / unresolved (95% intervals). Veto rows count where it gives up less at the same skipped share.

Comparison or measure	Result
Ranking vs. random	Above allocation band on 14/14 corpora
Ranking vs. cheap gate	5 / 4 / 5
Ranking vs. successor mean M_H	6 / 2 / 6
Ranking vs. depth-matched L_{H+1} ($H = 1$; $H = 2$)	3 / 0 / 7; 2 / 1 / 1
Veto: decisions skipped / benefit given up	49.4% / 19.9%
Veto vs. successor mean M_H	Gives up less on 21/31 arms
Veto vs. depth-matched L_{H+1} ($H = 1$; $H = 2$)	Gives up less on 7/10 arms; 3/4 corpora
Veto vs. cheap gate	Gives up less on 16/31 arms

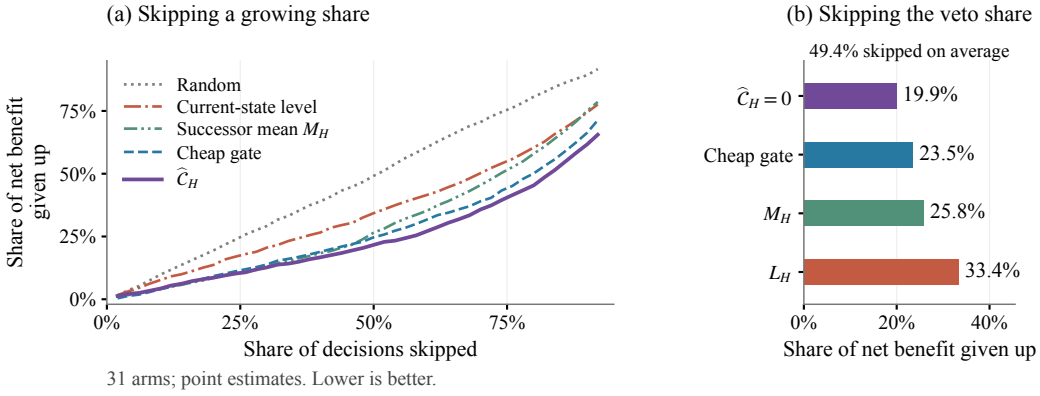

Figure 2: **Skipping search offline, 31 arms (point estimates).** (a) Benefit given up as each signal skips a growing share of decisions. (b) Each arm skips the share where its estimated criticality is zero, 49.4% on average; the other signals skip the same share in that arm. Lower is better. Arm-level ratios are averaged with equal weights.

6.3 Ranking

Criticality's point estimate lies above the matched random band on all fourteen corpora (Table 1). Against the cheap gate and the successor mean, the per-corpus results are mixed: criticality is clearly better on 5 and 6 corpora and clearly worse on 4 and 2. Against the current-state level at the same lookahead depth, criticality is never clearly worse on the ten $H = 1$ pairs (3 better, 7 unresolved). At the primary $H = 2$, it is clearly better on Stack ($+0.404$) and Meta-World ($+0.449$); on Lift the level is better (-0.147), and the image policy does not resolve (Appendix Table 5).

6.4 Skipping search where criticality is zero

Corollary 5.2 suggests skipping search on the veto set, where estimated criticality is zero (Section 4). Across thirty-one arms, the veto skips 49.4% of decisions (weighted) and gives up 19.9% of the net benefit of search. Skipping the same share with other signals gives up more: 23.5% for the cheap gate, 25.8% for the successor mean and 33.4% for the current-state level (Figure 2b); at matched depth, the veto also gives up less than L_3 on three of four corpora, none resolving alone. The gap to the successor mean, 5.9 points ($[0.8, 12.0]$ over 17 corpus clusters), is the most informative, because the two use the same rollouts and differ only in taking the variance instead of the mean. Against the cheap gate the difference is small and depends on how benefit is measured (Appendix D.6); per-arm counts are in Table 1.

Action-value variation is small on average in the veto group. The bound in Section 5 needs the value differences that criticality does not see to be small. On Stack, we re-evaluate 200 states per criticality group, with eight candidates and eight continuations each. The action-value variance is

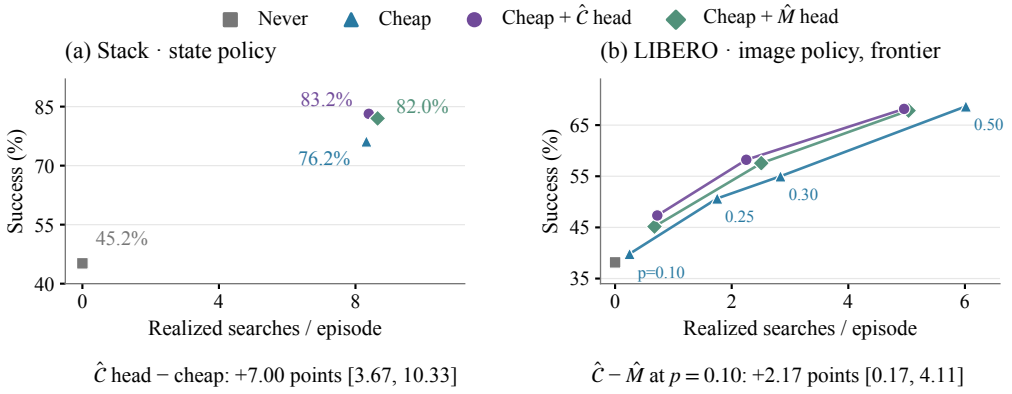

Figure 3: **Held-out success and search allocation in two settings.** Each panel uses 600 paired reset IDs; Cheap + $\widehat{C}$ head is the combined gate. (a) Stack at the pilot-selected operating point. (b) Image policy at nominal $p = 0.10, 0.25, 0.50$ for each gate, plotted at realized calls; head points at $p \leq 0.25$ average three head seeds (whiskers span them).

0.00033 ($[-0.00029, 0.00101]$) in the zero group, against 0.00143 and 0.00731 in the low and high groups (Appendix C.4). Because $\rho_H^2(s) \leq \mathrm{Var}_A[Q(s, A)]$ at every state, this supports a small group RMS residual in the veto group on this corpus; it is not a per-state certificate. The veto set is less stable than its size: on Stack, across eight estimator seeds its share barely moves, but 31.6% of probe states change membership at least once (Appendix B).

7 Closed-Loop Deployment

A gate that runs throughout an episode changes the states it later sees, so we test it separately from the single-switch labels. The deployed gate sums three calibrated scores: elapsed time, a gripper cue from twelve policy samples, and a small network distilled from criticality. It searches when the sum passes a threshold (Figure 6; Appendix D.5). We call this the *combined gate*, and the first two scores alone the *cheap gate*. The network reads privileged simulator state even for the image policy, where its teacher uses $H = 1$ and chunk length $k = 16$.

7.1 Held-out tests

Image policy. Before testing, a 60-episode pilot chose the cheap gate's coverage ($p = 0.30$) so that its search count matched the combined gate at $p = 0.25$ (2.82 against 2.70 calls per episode); success was not used for this choice. On 600 new paired episodes, the combined gate reaches 59.17% success, against 55.00% for the cheap gate and 38.17% without search. This is a gain of 4.17 points ($[1.67, 6.67]$; 41 episodes rescued, 16 broken). On test it also searches less than the cheap gate, 2.26 against 2.84 calls per episode (20.5% fewer). Because the call counts no longer match on test, this is a result at the chosen operating point rather than at equal spend. In a pre-registered frontier (Figure 3b; Appendix D.8), both head gates beat the cheap gate on success at $p = 0.10$ and 0.25 with more calls; at $p = 0.50$ success is similar with about one call fewer.

Stack. With a state-based policy, the combined gate reaches 83.17% success, against 76.17% for the cheap gate and 45.17% without search: a gain of 7.00 points ($[3.67, 10.33]$) at 8.39 against 8.32 calls per episode (Appendix D.7). On earlier 300-episode cohorts it also beats per-episode random allocation, by 21.0 points on Stack ($[15.3, 26.7]$) with 23% fewer calls and by 26.0 on Lift ($[19.9, 32.1]$); on the image policy it is 4.3 points ahead, which does not resolve ($[-0.3, 9.0]$; Appendix Table 13).

Does the gain come from criticality? Table 2 changes only the teacher of the network and keeps inputs, architecture and folds fixed. Offline, the criticality network ranks states far better than a

Table 2: **Effect of the teacher on Stack.** Offline GapClosed at $p = 0.25$, mean $\pm$ SD over three head seeds. The cheap gate alone scores 0.476.

Teacher	Head only	Cheap + head
Criticality	0.298 ± 0.051	0.639 ± 0.014
Successor mean	0.070 ± 0.030	0.627 ± 0.032
Random target	-0.106 ± 0.053	0.348 ± 0.025

network trained on the successor mean (0.30 against 0.07 of the oracle gap; Appendix D.6). Once the cheap signals are added, most of this difference disappears (0.64 against 0.63). With the undistilled scores, criticality still adds to a gate that already holds the cheap signals and the mean, raising the gap closed by 0.060 ($[0.015, 0.097]$; 23 of 31 arms, Appendix Figure 7), though not on the Stack corpus used for deployment. In closed loop the mean network also beats the cheap gate. On Stack criticality leads it by 1.17 points ($[-1.33, 3.67]$), unresolved. On the image policy, averaged over three head seeds, criticality has 2.17 points higher success at similar calls at $p = 0.10$ ($[0.17, 4.11]$) and similar success with 0.26 fewer calls at $p = 0.25$ ($[0.14, 0.37]$).

7.2 Cost

On a separate 100-episode image-policy cohort, the combined gate needs 46.0 seconds of online time per episode against 54.2 for the cheap gate. The network itself is cheap: the clock and network scores together take 0.075 ms per decision, and most of the gate's 26.9 ms goes to the twelve policy samples of the gripper cue. The planner accounts for 98.5% of the combined gate's time (Appendix D.9).

8 Limitations

Simulation and embodiment scope. The evaluation covers MuJoCo-based manipulation in robosuite, LIBERO and Meta-World, with setting-specific heads and calibration. Both teacher construction and the deployed head use simulator state; search uses a privileged rollout evaluator. Pixel-only gating, a learned evaluator and unseen tasks or embodiments have not yet been tested.

Operator-specific causal target. The causal target is defined for a fixed best-of-eight operator and continuation policy. Changing the operator changes Δ_S and can change the useful gating direction. This operator often cannot separate its candidates (argmax ties at an estimated 84.5% of Stack decisions), which bounds what any gate can predict.

Estimation and conditional theory. The bound requires control of the residual optionality defect, the veto set depends on estimation error, and several arms share a corpus.

Deployment. The deployment tests use held-out resets on settings studied earlier. The image-policy frontier covers three budgets and Stack one operating point. Criticality's closed-loop edge over a successor-mean head is small, rests on three head seeds on one task, and is unresolved on Stack.

9 Conclusion

We asked when test-time search can be skipped, and answered with the policy's own candidate actions: when they all lead to states with similar reach, search has little to gain. Successor-channel criticality measures this without reward or search labels, and a bound ties low criticality to low search value under stated conditions. Skipping search where estimated criticality is zero avoids about half of all decisions while giving up a fifth of the net benefit, less than the mean of the same score or a cheap time-and-gripper rule. The criticality head is the stronger standalone offline ranker on Stack, and both distilled heads improve deployment success, criticality's at lower search cost on the image policy. The main open steps are gating from pixels and a learned evaluator.

REPRODUCIBILITY STATEMENT

Theoretical results are proved in Appendices A and B. Appendix D lists every corpus and arm (D.1), the search operator, label protocol and seed keys (D.2), baseline formulas (D.3), estimator settings and world-state coordinates (D.4), distillation and calibration (D.5), teacher controls (D.6), pilot and test protocols with reset IDs and file hashes (D.7), and compute and timing (D.9). Table 21 states which reported quantities are regenerated automatically and which need extra care. Offline re-analyses read saved arrays and are deterministic given the stated bootstrap seeds. Anonymized code, the saved arrays behind every offline table and deployment interval, and the figure sources are in the supplementary material; we will release them publicly with the final version.

AI USE STATEMENT

AI tools assisted this manuscript revision through editing, code inspection, figure preparation, and scripts for reanalysis of existing experimental artifacts. This assistance did not generate new benchmark rollouts or train new policies. The authors remain responsible for the final manuscript and its claims.

REFERENCES

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

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Robert Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, Laura Smith, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0: A vision-language-action flow model for general robot control. In *Proceedings of Robotics: Science and Systems*, 2025. doi: 10.15607/RSS.2025.XXI.010. URL `https://www.roboticsproceedings.org/rss21/p010.html`.

Xiaowei Cai et al. τ_0-VLA: A hierarchical robot foundation model with world-model-guided test-time computation. arXiv preprint arXiv:2608.16885, 2026.

Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin C. M. Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Proceedings of Robotics: Science and Systems*, 2023. doi: 10.15607/RSS.2023.XIX.026.

C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970. doi: 10.1109/TIT.1970.1054406.

Jadelynn Dao, Milan Ganai, Yasmina Abukhadra, Ajay Sridhar, Mozhgan Nasr Azadani, Katie Luo, Clark Barrett, Jiajun Wu, Chelsea Finn, and Marco Pavone. DIRECT: When and where should you allocate test-time compute in embodied planners? arXiv preprint arXiv:2606.12402, 2026.

Thomas Dean and Mark Boddy. An analysis of time-dependent planning. In *Proceedings of the Seventh National Conference on Artificial Intelligence*, pp. 49–54, 1988.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, 2017.

Alex Graves. Adaptive computation time for recurrent neural networks. arXiv preprint arXiv:1603.08983, 2016.

H. O. Hartley and H. A. David. Universal bounds for mean range and extreme observation. *The Annals of Mathematical Statistics*, 25(1):85–99, 1954. URL `https://www.jstor.org/stable/2236514`.

Nicholas Hay, Stuart Russell, David Tolpin, and Solomon Eyal Shimony. Selecting computations: Theory and applications. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, 2012.

Eric J. Horvitz. Reasoning about beliefs and actions under computational resource constraints. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 1987.

Riccardo Andrea Izzo, Gianluca Bardaro, and Matteo Matteucci. Act, think or abstain: Complexity-aware adaptive inference for vision-language-action models. arXiv preprint arXiv:2603.05147, 2026.

Tobias Jung, Daniel Polani, and Peter Stone. Empowerment for continuous agent–environment systems. *Adaptive Behavior*, 19(1):16–39, 2011. doi: 10.1177/1059712310392389.

Alexander S. Klyubin, Daniel Polani, and Chrystopher L. Nehaniv. All else being equal be empowered. In *European Conference on Artificial Life*, pp. 744–753, 2005.

Alexander S. Klyubin, Daniel Polani, and Chrystopher L. Nehaniv. Representations of space and time in the maximization of information flow in the perception-action loop. *Neural Computation*, 19(9):2387–2432, 2007. doi: 10.1162/neco.2007.19.9.2387.

Jacky Kwok, Christopher Agia, Rohan Sinha, Matt Foutter, Shulu Li, Ion Stoica, Azalia Mirhoseini, and Marco Pavone. RoboMonkey: Scaling test-time sampling and verification for vision-language-action models. In *Proceedings of the 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pp. 3200–3217, 2025.

Sungjae Lee, Hyejin Park, Jaechang Kim, and Jungseul Ok. Semantic exploration with adaptive gating for efficient problem solving with language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 588–606. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.29. URL https://aclanthology.org/2025.acl-long.29/.

Wenhao Li, Xiu Su, Yichao Cao, Hongyan Xu, Xiaobo Xia, Shan You, Yi Chen, and Chang Xu. VLA-ATTC: Adaptive test-time compute for VLA models with relative action critic model. arXiv preprint arXiv:2605.01194, 2026a.

Ye Li, Huanan Liu, Kangye Ji, Yuan Meng, Jiajun Fan, Yuansong Wang, Shiyu Qin, Chenglei Wu, Shu-Tao Xia, and Zhi Wang. ElegantVLA: Learning when to think for efficient vision-language-action models. arXiv preprint arXiv:2605.29438, 2026b.

Ziming Li, Jiatan Huang, Xiaoguang Guo, Guilin Wang, and Chuxu Zhang. Same signal, opposite meaning: Direction-informed adaptive learning for LLM agents. arXiv preprint arXiv:2605.06908, 2026c.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. LIBERO: Benchmarking knowledge transfer for lifelong robot learning. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

Yuejiang Liu, Jubayer Ibn Hamid, Annie Xie, Yoonho Lee, Maximilian Du, and Chelsea Finn. Bidirectional decoding: Improving action chunking via guided test-time sampling. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/0d78dd998f7b9ac79604d47a2d79bb0d-Abstract-Conference.html.

Shakir Mohamed and Danilo Jimenez Rezende. Variational information maximisation for intrinsically motivated reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Mitsuhiko Nakamoto, Oier Mees, Aviral Kumar, and Sergey Levine. Steering your generalists: Improving robotic foundation models via value guidance. In *Proceedings of The 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 4996–5013, 2025. URL https://proceedings.mlr.press/v270/nakamoto25a.html.

Yi Pan, Miao Pan, Qi Lu, Jiaming Huang, Man Zhang, Siteng Huang, Xin Li, Jie Zhang, Yongliang Shen, Xuhong Zhang, and Wenqi Zhang. VLA-Corrector: Lightweight detect-and-correct inference for adaptive action horizon. arXiv preprint arXiv:2607.01804, 2026.

Stuart Russell and Eric Wefald. Principles of metareasoning. *Artificial Intelligence*, 49(1–3):361–395, 1991. doi: 10.1016/0004-3702(91)90015-C.

Christoph Salge, Cornelius Glackin, and Daniel Polani. Empowerment—an introduction. In Mikhail Prokopenko (ed.), *Guided Self-Organization: Inception*, volume 9 of *Emergence, Complexity and Computation*, pp. 67–114. Springer, 2014. doi: 10.1007/978-3-642-53734-9_4.

Maximilian Seitzer, Bernhard Schölkopf, and Georg Martius. Causal influence detection for improving efficiency in reinforcement learning. In *Advances in Neural Information Processing Systems*, 2021.

Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A physics engine for model-based control. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033, 2012.

Yun-Jui Tsai, Wei-Yu Chen, Yan-Ru Ju, Yu-Hung Chang, and Ti-Rong Wu. Regret-guided search control for efficient learning in AlphaZero. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/9e720fce64f91114c49cfd640d821da3-Abstract-Conference.html.

Zhijie Wu, Kento Kawaharazuka, and Kei Okada. Neural introspection gating for adaptive KV-cache reuse in vision-language-action models. arXiv preprint arXiv:2608.10824, 2026.

Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-World: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Proceedings of the Conference on Robot Learning*, volume 100 of *Proceedings of Machine Learning Research*, pp. 1094–1100, 2020. URL https://proceedings.mlr.press/v100/yu20a.html.

Wenda Yu, Tianshi Wang, Fengling Li, Jingjing Li, and Lei Zhu. AC^2-VLA: Action-context-aware adaptive computation in vision-language-action models for efficient robotic manipulation. arXiv preprint arXiv:2601.19634, 2026.

Yang Yue, Yulin Wang, Bingyi Kang, Yizeng Han, Shenzhi Wang, Shiji Song, Jiashi Feng, and Gao Huang. DeeR-VLA: Dynamic inference of multimodal large language models for efficient robot execution. In *Advances in Neural Information Processing Systems*, 2024.

Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. In *Proceedings of Robotics: Science and Systems*, 2023.

Yuke Zhu, Josiah Wong, Ajay Mandlekar, Roberto Martín-Martín, Abhishek Joshi, Kevin Lin, Abhiram Maddukuri, Soroush Nasiriany, and Yifeng Zhu. robosuite: A modular simulation framework and benchmark for robot learning. arXiv preprint arXiv:2009.12293, 2020.

APPENDIX CONTENTS

A Proofs and Theoretical Extensions **14**

A.1 Notation and horizon convention 14

A.2 Channel-level theorem 15

A.3 Proof of Theorem A.1 15

A.4 Relaxing the world-state factorization 17

A.5 Proof of Theorem 5.1 17

A.6 Mean-matched successor construction 18

A.7 A finite-estimation veto condition 18

A.8 Why we do not use a reward-free bound on ρ_H 19

B Estimator Details **19**

C Additional Controls **21**

C.1 Direction instability across corpora and policy redraws 21

C.2 Figure 1: recorded score profiles and estimator schematic 21

C.3 Effective-depth-matched level comparison 22

C.4 Empirical audit of the optionality defect 23

D Experimental Setup and Reproducibility **23**

D.1 Corpus and arm registry 23

D.2 Empirical search operator and paired labels 24

D.3 Gating baselines 26

D.4 Criticality estimator and world-state representation 26

D.5 Distilled deployment predictor 27

D.6 Teacher controls and veto sensitivity 28

D.7 Pilot selection and held-out deployment tests 31

D.8 Image-policy success–calls frontier 31

D.9 Compute accounting 32

D.10 Reproducibility status 33

A Proofs and Theoretical Extensions

This appendix proves the results in Section 5 and records the horizon convention used throughout the paper. The notation matches Sections 3 and 4; in particular, the channel horizon H is never suppressed.

Table 3: Notation. Population quantities, Monte-Carlo estimates and the deployed head are distinct.

Symbol	Meaning
k, H	Primitive actions per chunk; channel lookahead in chunks.
$A_h^k,\ A^{(i)}$	Temporal chunk h; independent candidate chunk i at a fixed state.
n, K	Search candidates (8); paired label repeats (12).
$n_{\text{outer}}, n_{\text{inner}}$	Teacher successors (24); channel rollouts per successor (32, two halves of 16).
σ_{res}, d_w	World resolution (0.002 metres); world-projection dimension.
$\mathcal{I}_\pi^{k,H},\ \bar{\mathcal{I}}_\pi^{k,H} = L_H$	Policy-induced mutual information; its Gaussian log-det upper bound (the channel score).
Q, J, λ	Action value after a chunk; realized return; exchange rate between return and computation.
$E_H(a), M_H, \mathcal{C}_H$	Successor level, its candidate mean, and its candidate variance.
$\widehat{L}_i^{(j)}, \widehat{M}_H$	Half-j successor channel estimate; mean of half-averaged scores.
$\widehat{\mathcal{C}}_{H,\text{raw}}, \widehat{\mathcal{C}}_H$	Cross-half sample covariance; its nonnegative clamp.
h_θ, ς	Regressor of log-transformed criticality; positive target scale.
g_H, κ, ρ_H	Conditional value mean, its Lipschitz constant, residual standard deviation.
$B,\ \mathcal{B}_r(p)$	Span bound on action value; normalized weighted benefit of score r at coverage p.
$g, p, q(V), \ell(V)$	Binary gate; nominal search coverage; share of decisions skipped; signed share of benefit given up.
c, c_n, c_n', γ_n	Search cost; bound constants ($c_n = c_n'/\sqrt{2}$), not compute costs; range constant $\max_u(u - u^n)$.
η, β	Offset and value slope in the two-type construction (Appendix A.6).
$\Delta_{\mathcal{S}}, G_n$	Fixed-operator causal benefit; ideal best-of-n gain envelope.

A.1 Notation and horizon convention

Let

$$A_{1:H}^k = (A_1^k, \dots, A_H^k)$$

denote H consecutive action chunks sampled from the frozen policy, starting at state s. Let

$$W_H(s) = z^{\text{world}}\big(F_H(s, A_{1:H}^k)\big)$$

be the task-relevant world-state endpoint after exactly H chunks from s, and let

$$Y_H(s) = W_H(s) + \xi, \qquad \xi \sim \mathcal{N}(0, \sigma_{\text{res}}^2 I).$$

The current-state level is

$$L_H(s) = \bar{\mathcal{I}}_\pi^{k,H}(s) = \frac{1}{2} \log\det\left(I + \frac{\text{Cov}[W_H(s)]}{\sigma_{\text{res}}^2}\right).$$

For a fixed first chunk a, write

$$W_H^{\text{lev}}(s, a) \overset{d}{=} W_H(s) \mid A_1^k = a.$$

Equivalently, $W_H^{\text{lev}}(s, a)$ is obtained by executing a followed by $H - 1$ continuation chunks from π. Hence it lies exactly H chunks from the original state.

The first chunk also determines a successor

$$s_a' = F_1(s, a).$$

The successor-channel quantity used by criticality is

$$E_H(a) = \bar{\mathcal{I}}_\pi^{k,H}(s_a'),$$

which evaluates a fresh H-chunk channel starting from s'_a. Therefore $E_H(a)$ reaches $H+1$ chunks from the original state, and

$$\mathcal{C}_H(s) = \mathrm{Var}_{A\sim\pi(\cdot|s)}\left[E_H(A)\right].$$

This distinction is the reason the effective-depth-matched current-state comparison is $\mathcal{C}_H$ versus L_{H+1} rather than L_H.

Search operator. The idealized best-of-n search operator draws $A^{(1)},\dots,A^{(n)} \overset{\text{i.i.d.}}{\sim} \pi(\cdot \mid s)$ and returns the candidate with largest action value. Its gain relative to one policy draw is

$$G_n(s) := \mathbb{E}\left[\max_{1\le i\le n} Q(s,A^{(i)})\right] - \mathbb{E}_{A\sim\pi(\cdot|s)}\left[Q(s,A)\right]. \tag{10}$$

For the empirical search operator $\mathcal{S}$ used in the paper,

$$\Delta_{\mathcal{S}}(s) \le G_n(s), \qquad \mathrm{VOC}_{\mathcal{S}}(s) \le G_n(s) - \lambda c,$$

because an imperfect evaluator cannot outperform ideal selection from the same candidate set.

Finite-resolution value class. Fix $\sigma_{\mathrm{res}} > 0$ and horizon H. We say that the continuation value is σ_{res}-*resolvable at horizon H with span B* if there exists a function V_0 with

$$\mathrm{span}(V_0) \le B$$

such that, for every first chunk a,

$$Q(s,a) = \mathbb{E}\left[V_0\left(W_H^{\mathrm{lev}}(s,a) + \xi\right)\right], \qquad \xi \sim \mathcal{N}(0, \sigma_{\mathrm{res}}^2 I). \tag{11}$$

This class excludes reward differences that are only observable below the state resolution used by the channel.

A.2 Channel-level theorem

Use the best-of-n envelope in Eq. 10 and the σ_{res}-resolvable value class in Eq. 11. The conditional endpoint $W_H^{\mathrm{lev}}(s,a)$ already fixes the first chunk to a; its remaining randomness comes from continuation chunks. Let $L_H(s) = \bar{\mathcal{I}}_\pi^{k,H}(s)$, $c_n = (n-1)/\sqrt{2(2n-1)}$, and $\gamma_n = \max_{u\in[0,1]}(u - u^n)$.

Theorem A.1 (Finite-resolution channel bound)**.** *If the continuation value is σ_{res}-resolvable at horizon H with span B, then*

$$G_n(s) \le B\min\left\{c_n\sqrt{L_H(s)},\ \gamma_n\right\}. \tag{12}$$

Therefore,

$$\mathrm{VOC}_{\mathcal{S}}(s) \le B\min\left\{c_n\sqrt{L_H(s)},\ \gamma_n\right\} - \lambda c.$$

The proof is given in Appendix A.3. The bound is one-sided: a sufficiently small current-state level constrains the possible gain from search within the stated finite-resolution value class, whereas a large level does not imply that search is useful.

A.3 Proof of Theorem A.1

For a fixed first chunk a, let P_a^σ denote the law of $W_H^{\mathrm{lev}}(s,a) + \xi$, and let

$$\bar{P}^\sigma = \mathbb{E}_{a\sim\pi(\cdot|s)}\left[P_a^\sigma\right].$$

By Eq. 11,

$$Q(s,a) = \mathbb{E}_{y\sim P_a^\sigma}[V_0(y)].$$

Since $\mathrm{span}(V_0) \le B$,

$$|Q(s,a) - \mathbb{E}_\pi[Q(s,A)]| \le B\ \mathrm{TV}\left(P_a^\sigma, \bar{P}^\sigma\right). \tag{13}$$

Squaring and averaging over a gives

$$\mathrm{Var}_a[Q(s,a)] \le B^2\mathbb{E}_a\left[\mathrm{TV}^2\left(P_a^\sigma, \bar{P}^\sigma\right)\right]. \tag{14}$$

Pinsker's inequality therefore yields

$$\begin{aligned}\mathrm{Var}_a[Q(s,a)] &\le \frac{B^2}{2}\mathbb{E}_a\left[\mathrm{KL}\left(P_a^\sigma \| \bar{P}^\sigma\right)\right] \\ &= \frac{B^2}{2} I\left(A_1^k; Y_H(s) \mid s\right). \end{aligned} \tag{15}$$

For $H > 1$, the channel score is defined from the complete sampled chunk sequence. Because

$$A_1^k \to A_{1:H}^k \to Y_H(s)$$

forms a Markov chain conditional on s, data processing gives

$$I(A_1^k; Y_H(s) \mid s) \le I(A_{1:H}^k; Y_H(s) \mid s).$$

Conditioned on $A_{1:H}^k$, the endpoint is deterministic before the addition of ξ, so

$$\begin{aligned} I(A_{1:H}^k; Y_H(s) \mid s) &= h(Y_H(s) \mid s) - h(\xi) \\ &\le \frac{1}{2}\log\det\left(I + \frac{\mathrm{Cov}[W_H(s)]}{\sigma_{\mathrm{res}}^2}\right) \\ &= L_H(s), \end{aligned} \tag{16}$$

where the inequality is the Gaussian maximum-entropy bound. No Gaussian assumption is made on the reachable endpoints themselves. The bound follows in three steps. Given $A_{1:H}^k$ the endpoint $W_H(s)$ is fixed, so $h(Y_H(s) \mid A_{1:H}^k, s) = h(\xi) = \frac{1}{2}\log\det(2\pi e\,\sigma_{\mathrm{res}}^2 I)$. Because ξ is independent of $W_H(s)$, $\mathrm{Cov}[Y_H(s) \mid s] = \mathrm{Cov}[W_H(s)] + \sigma_{\mathrm{res}}^2 I$, and among densities with a given covariance the Gaussian has the largest differential entropy, so $h(Y_H(s) \mid s) \le \frac{1}{2}\log\det\big(2\pi e(\mathrm{Cov}[W_H(s)] + \sigma_{\mathrm{res}}^2 I)\big)$. Subtracting the two entropies gives Eq. 16, in nats. The noise is what makes the quantity finite: without it the endpoint is deterministic given the actions, $h(Y_H(s) \mid A_{1:H}^k, s) = -\infty$, and the mutual information diverges for continuous endpoints. σ_{res} sets the resolution at which endpoints count as distinct, and it can reorder states, not only rescale them: for large σ_{res} the score approaches $\mathrm{tr}\,\mathrm{Cov}[W_H]/(2\sigma_{\mathrm{res}}^2)$, while for small σ_{res} and full-rank covariance it approaches $\frac{1}{2}\log\det\mathrm{Cov}[W_H] - d\log\sigma_{\mathrm{res}}$, with d the world dimension.

Combining the previous inequalities,

$$\mathrm{sd}_a[Q(s,a)] \le B\sqrt{\frac{L_H(s)}{2}}. \tag{17}$$

For n i.i.d. candidate values, the Hartley–David–Gumbel expected-maximum inequality (Hartley & David, 1954) gives

$$G_n(s) \le \frac{n-1}{\sqrt{2n-1}}\,\mathrm{sd}_a[Q(s,a)].$$

Thus

$$G_n(s) \le B c_n \sqrt{L_H(s)}, \qquad c_n = \frac{n-1}{\sqrt{2(2n-1)}}. \tag{18}$$

A second bound follows only from the range of Q. After rescaling its support to $[0, B]$, with cumulative distribution function F,

$$\begin{aligned} G_n(s) &= \int_0^B \left(F(x) - F(x)^n\right) dx \\ &\le B \max_{u\in[0,1]} (u - u^n) = B\gamma_n. \end{aligned} \tag{19}$$

Taking the smaller of the two bounds proves Theorem A.1. □

Relation to empowerment capacity. The score $L_H(s) = \bar{\mathcal{I}}_\pi^{k,H}(s)$ upper-bounds the mutual information induced by the current policy,

$$\mathcal{I}_\pi^{k,H}(s) \le \bar{\mathcal{I}}_\pi^{k,H}(s).$$

Classical empowerment instead maximizes mutual information over the source action distribution, so

$$\mathcal{I}_\pi^{k,H}(s) \le \mathcal{E}_{\sigma_{\mathrm{res}}}^{k,H}(s),$$

where $\mathcal{E}_{\sigma_{\mathrm{res}}}^{k,H}(s)$, the empowerment capacity, is the supremum of $I(A_{1:H}^k; Y_H(s) \mid s)$ over source distributions of $A_{1:H}^k$. These two inequalities do not order $\bar{\mathcal{I}}_\pi^{k,H}$ and the empowerment capacity. The proof uses only the first inequality.

A.4 Relaxing the world-state factorization

Theorem A.1 assumes that continuation value is determined by the task-relevant world state at the resolution of the channel. Two actions can instead reach the same world configuration with different robot configurations and therefore different downstream value.

At the H-chunk endpoint, write (U_H, W_H) for robot and world components. Assume $Q(s,a) = \mathbb{E}[V(U_H, W_H) \mid A_1^k = a, s]$. Under the fixed policy-induced joint law at s, define

$$\bar{V}_s(w) = \mathbb{E}[V(U_H, W_H) \mid W_H = w, s], \qquad \delta_H(s) = \sup_w \operatorname{span}_u V(u,w),$$

over reachable endpoint states, and set $\bar{Q}(s,a) = \mathbb{E}[\bar{V}_s(W_H) \mid A_1^k = a, s]$. The same $\bar{V}_s$ is used for every candidate a.

Proposition A.2 (Factorization-defect extension)**.** *Suppose $\bar{Q}$ is σ_{res}-resolvable as in Eq. 11, with span B'. Then*

$$G_n(s) \le c_n B' \sqrt{L_H(s)} + c_n' \delta_H(s), \qquad c_n' = \frac{n-1}{\sqrt{2n-1}}.$$

To see this, write $Q = \bar{Q} + R$. The conditional average $\bar{V}_s(w)$ lies between the smallest and largest reachable values at w, hence $|R(s,a)| \le \delta_H(s)$ and $\mathrm{sd}_A R(s,A) \le \delta_H(s)$. The information term of Theorem A.1, applied to $\bar{Q}$, gives $\mathrm{sd}_A \bar{Q} \le B'\sqrt{L_H/2}$. Minkowski's inequality and the expected-maximum bound prove the result. The extra factorization and resolvability assumptions are explicit; this extension is not needed for the main criticality theorem.

Expected-maximum inequality. For completeness, let X have finite variance, mean μ_X and generalized quantile $q_X(u)$. The maximum of n independent copies has expectation $\int_0^1 q_X(u) n u^{n-1}\, du$. Hence

$$\mathbb{E}[\max_{i \le n} X_i] - \mu_X = \int_0^1 (q_X(u) - \mu_X)(n u^{n-1} - 1)\, du \le \mathrm{sd}(X) \frac{n-1}{\sqrt{2n-1}},$$

by Cauchy–Schwarz and $\int_0^1 (n u^{n-1} - 1)^2 du = (n-1)^2/(2n-1)$. Generalized quantiles make the argument valid for atoms as well as continuous distributions. Applying it to $X = Q(s,A)$ gives the search envelope used in both theorems (Hartley & David, 1954).

A.5 Proof of Theorem 5.1

Recall

$$E_H(A) = \bar{\mathcal{I}}_\pi^{k,H}(s'_A), \qquad \mathcal{C}_H(s) = \mathrm{Var}[E_H(A)].$$

Define

$$g_H(e) = \mathbb{E}[Q(s,A) \mid E_H(A) = e]$$

and

$$\rho_H^2(s) = \mathbb{E}\left[\mathrm{Var}\left(Q(s,A) \mid E_H(A)\right)\right].$$

The law of total variance gives

$$\begin{aligned} \mathrm{Var}[Q(s,A)] &= \mathrm{Var}\left(\mathbb{E}[Q(s,A) \mid E_H(A)]\right) + \mathbb{E}\left[\mathrm{Var}\left(Q(s,A) \mid E_H(A)\right)\right] \\ &= \mathrm{Var}[g_H(E_H(A))] + \rho_H^2(s). \end{aligned} \tag{20}$$

Let $E_H(A')$ be an independent copy of $E_H(A)$. If g_H is κ-Lipschitz,

$$\begin{aligned}\mathrm{Var}[g_H(E_H(A))] &= \frac{1}{2}\mathbb{E}\left[\left(g_H(E_H(A)) - g_H(E_H(A'))\right)^2\right] \\ &\le \frac{\kappa^2}{2}\mathbb{E}\left[\left(E_H(A) - E_H(A')\right)^2\right] \\ &= \kappa^2 \mathcal{C}_H(s). \end{aligned} \tag{21}$$

Therefore

$$\mathrm{sd}[Q(s,A)] \le \sqrt{\kappa^2 \mathcal{C}_H(s) + \rho_H^2(s)}.$$

Applying the same expected-maximum inequality as above,

$$\begin{aligned} G_n(s) &\le c_n' \sqrt{\kappa^2 \mathcal{C}_H(s) + \rho_H^2(s)} \\ &\le c_n' \left(\kappa \sqrt{\mathcal{C}_H(s)} + \rho_H(s)\right). \end{aligned} \tag{22}$$

Combining this with $\mathrm{VOC}_{\mathcal{S}}(s) \le G_n(s) - \lambda c$ proves Eq. 9 using the first line. The second line recovers the looser additive form. If Q has span at most B, the range argument in Appendix A.3 also gives $G_n \le B\gamma_n$; taking the minimum proves the capped form. □

Proof of Corollary 5.2. Under the stated condition, $\kappa^2 \mathcal{C}_H(s) \le (\lambda c/c_n')^2 - r^2$ and $\rho_H^2(s) \le r^2$, so $\kappa^2 \mathcal{C}_H(s) + \rho_H^2(s) \le (\lambda c/c_n')^2$. Taking square roots in Eq. 9 proves $\mathrm{VOC}_{\mathcal{S}} \le 0$. At $\mathcal{C}_H = 0$ the bound reduces to $c_n' \rho_H - \lambda c$, including when $\kappa = 0$. For this bound to certify the veto at every nonnegative price, including zero price, ρ_H must vanish. □

A.6 Mean-matched successor construction

Proposition A.3 (Variation at fixed successor mean)**.** *Fix $\mu > 0$, $\beta > 0$, and $q_0 \in (0,1)$. For $0 \le \eta \le \min\{\mu, q_0/\beta, (1-q_0)/\beta\}$, consider two successor types with equal policy probability and*

$$E_H(A) = \mu \pm \eta, \qquad Q(s,A) = q_0 \pm \beta\eta,$$

where signs agree. Then $M_H(s) = \mu$, $\rho_H(s) = 0$, $\mathcal{C}_H(s) = \eta^2$, and

$$G_n(s) = \beta\sqrt{\mathcal{C}_H(s)}(1 - 2^{1-n}).$$

Proof. The mean and variance follow directly. The conditional value is affine with slope β and zero residual. All n candidates have the lower value with probability 2^{-n}; otherwise their maximum is the higher value. Subtracting $\mathbb{E}Q = q_0$ gives the expression. □

The score values can be realized by one-dimensional Gaussian channel endpoints with variances $\sigma_{\mathrm{res}}^2[\exp(2(\mu \pm \eta)) - 1]$. A policy draws the endpoint via a stochastic action under deterministic dynamics, and continuation rewards can have the prescribed conditional expectations in $[0,1]$. Thus the construction is compatible with the channel score, rather than requiring an arbitrary signed feature. It establishes that successor mean alone cannot determine the best-of-n gain and that the square-root order is attained in this family. It does not assert value-score alignment in all tasks, optimality of the bound constant, or attainability by a noisy evaluator.

A.7 A finite-estimation veto condition

Proposition A.4 (Conditional finite-estimation certificate)**.** *At a fixed state, let $\varepsilon \ge 0$ and suppose an event of probability at least $1 - \alpha$ guarantees $\mathcal{C}_H(s) \le \widehat{\mathcal{C}}_H(s) + \varepsilon$ and $\rho_H(s) \le r$. For a valid Lipschitz bound $\kappa > 0$, if $\lambda c \ge c_n' r$ and*

$$\widehat{\mathcal{C}}_H(s) + \varepsilon \le \frac{(\lambda c/c_n')^2 - r^2}{\kappa^2},$$

then $\mathrm{VOC}_{\mathcal{S}}(s) \le 0$ on that event. The guarantee is for the implication: $\Pr($*the condition holds and* $\mathrm{VOC}_{\mathcal{S}}(s) > 0) \le \alpha$. *It is not a bound on* $\Pr(\mathrm{VOC}_{\mathcal{S}}(s) \le 0 \mid$ *the condition holds*$)$.

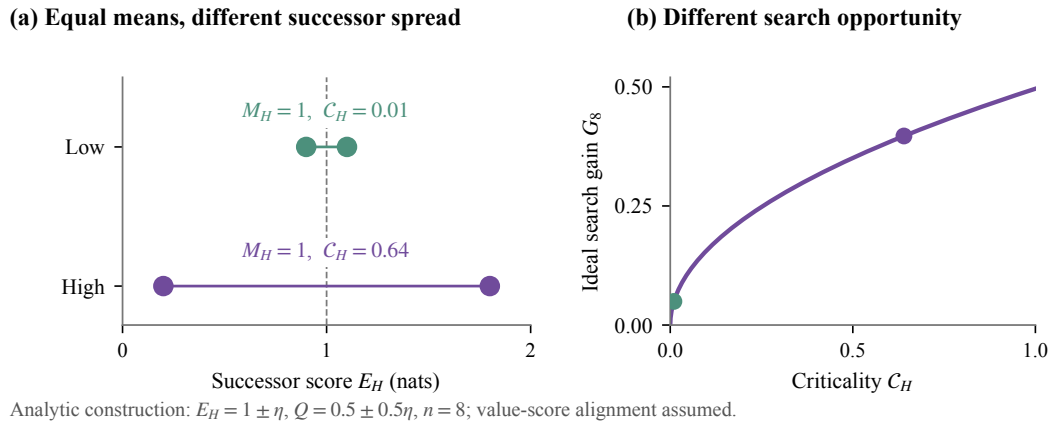

Figure 4: **Equal successor means can hide different search opportunities.** Two equiprobable successor types have scores $1 \pm \eta$ and values $0.5 \pm 0.5\eta$. Their mean score stays fixed, while the exact ideal gain is $G_8 = \frac{1}{2}\sqrt{C_H}(1-2^{-7})$. This analytic construction assumes value-score alignment; it is not benchmark data.

Proof. On the stated event, substitute the upper bounds into Eq. 9 and apply Corollary 5.2. □

This is a pointwise sufficient condition, not a confidence bound supplied by the split-half estimator. A data-dependent choice across many states would require simultaneous coverage. In particular, ε must account for finite-sample channel bias as well as sampling variation; a clamped zero does not provide it. The deployed CDF gate is evaluated empirically rather than asserted to satisfy this certificate.

Why the condition remains conditional. When $C_H(s) = 0$, $E_H(A)$ is almost surely constant, so

$$\rho_H^2(s) = \mathrm{Var}[Q(s, A)].$$

Criticality therefore does not control value differences that are invisible after successor controllability has been compressed to the scalar $E_H(A)$. This is why the paper retains the optionality defect rather than claiming an unconditional reward-free certificate.

A.8 Why we do not use a reward-free bound on ρ_H

We also attempted to upper-bound ρ_H using the same information-theoretic chain as Theorem A.1. The resulting bound is numerically vacuous on the states for which it would matter most. In a separate diagnostic it was approximately $14\times$ the realized quantity overall, and on the additional criticality-veto states it was nearly identical to the channel-level bound despite near-zero realized search benefit. The failure is structural: when $C_H = 0$, the conditioning variable is constant, and equality of successor-channel volumes does not imply equality of the corresponding reachable sets. We therefore make no reward-free theoretical claim about ρ_H and audit an action-value-variance upper bound below.

Where the channel-level bound is informative. On the $H = 2$ arms (61,877 states; $n = 8$, $B = 1$, $\gamma_n = 0.650$), $Bc_n\sqrt{L_H}$ is below the range bound $B\gamma_n$ on 45.2% of states. On 35.4% the plug-in expression is zero because the estimated level is zero; on the remaining 9.75% it is positive and below the range bound. These are estimated bound values, conditional on the theorem's value-factorization assumption; a zero estimate does not by itself certify zero population search gain.

B Estimator Details

At each probe state we draw $n_{\text{outer}} = 24$ candidate first chunks. For each successor we draw $n_{\text{inner}} = 32$ channel rollouts and split them into two disjoint halves of 16, producing channel estimates $\widehat{L}_i^{(1)}$ and $\widehat{L}_i^{(2)}$, where $\overline{L}^{(j)} = n_{\text{outer}}^{-1} \sum_i \widehat{L}_i^{(j)}$.

A naive variance across estimated successor scores contains both true successor variation and Monte-Carlo estimation noise. We instead use the split-half cross-covariance

$$\widehat{\mathcal{C}}_{H,\mathrm{raw}}(s) = \frac{1}{n_{\mathrm{outer}} - 1} \sum_{i=1}^{n_{\mathrm{outer}}} \left(\widehat{L}_i^{(1)} - \overline{L}^{(1)}\right)\left(\widehat{L}_i^{(2)} - \overline{L}^{(2)}\right).$$

What split-half correction estimates. The following result separates inner-rollout noise from successor-dependent channel bias. All expectations in this paragraph condition on the probe state s.

Proposition B.1 (Split-half target and bias transfer)**.** *Let $X_1, \ldots, X_N$ be i.i.d. candidate successors, $N \geq 2$. Conditional on X_i, let $\widehat{L}_i^{(1)}$ and $\widehat{L}_i^{(2)}$ be independent, identically distributed estimates with conditional mean $m(X_i)$ and variance $v(X_i)$; the pairs are independent across i. Assume the score variables below have finite second moments. Write $L(X) = \overline{\mathcal{I}}_\pi^{k,H}(X)$, $b(X) = m(X) - L(X)$, $\mathcal{C}_m = \mathrm{Var}[m(X)]$, and $\delta_b = \mathrm{sd}[b(X)]$. Let $\widehat{V}_{\mathrm{avg}}$ denote the unbiased sample variance of $(\widehat{L}_i^{(1)} + \widehat{L}_i^{(2)})/2$. Then*

$$\mathbb{E}[\widehat{\mathcal{C}}_{H,\mathrm{raw}}] = \mathcal{C}_m, \qquad \mathbb{E}[\widehat{V}_{\mathrm{avg}}] = \mathcal{C}_m + \tfrac{1}{2}\mathbb{E}[v(X)], \tag{23}$$

$$\left|\sqrt{\mathcal{C}_m} - \sqrt{\mathcal{C}_H}\right| \leq \delta_b. \tag{24}$$

For the channel score in Eq. 4, suppose each half uses an unbiased sample covariance $\widehat{\Sigma}$ from i.i.d. endpoints. If its conditional squared Frobenius error is finite, define

$$a(X) = \frac{\mathbb{E}[\|\widehat{\Sigma} - \Sigma_{W_H}(X)\|_F^2 \mid X]}{4\sigma_{\mathrm{res}}^4}.$$

Then $-a(X) \leq b(X) \leq 0$ and, whenever $\mathbb{E}[a(X)^2] < \infty$, $\delta_b \leq \sqrt{\mathbb{E}[a(X)^2]}$.

Proof. Total covariance and conditional independence give $\mathrm{Cov}(\widehat{L}^{(1)}, \widehat{L}^{(2)}) = \mathrm{Var}[m(X)]$. The unbiased sample covariance has this expectation. Total variance of the half average gives the second identity in Eq. 23. Since $m = L + b$, the reverse triangle inequality for centered L^2 variables gives Eq. 24. For the remaining statement set $f(S) = \frac{1}{2}\log\det(I + S/\sigma_{\mathrm{res}}^2)$ and $D = \widehat{\Sigma} - \Sigma_{W_H}(X)$. The entire line segment $S_t = \Sigma_{W_H}(X) + tD$, $0 \leq t \leq 1$, is positive semidefinite, and

$$\frac{d^2}{dt^2} f(S_t) = -\tfrac{1}{2}\,\mathrm{tr}\left[(\sigma_{\mathrm{res}}^2 I + S_t)^{-1} D (\sigma_{\mathrm{res}}^2 I + S_t)^{-1} D\right] \in \left[-\frac{\|D\|_F^2}{2\sigma_{\mathrm{res}}^4},\, 0\right].$$

Taylor's formula with integral remainder and $\mathbb{E}[D \mid X] = 0$ therefore give $-a(X) \leq b(X) \leq 0$. Finally, $\mathrm{Var}(b) \leq \mathbb{E}[b^2] \leq \mathbb{E}[a^2]$. □

The split removes the nonnegative inflation $\frac{1}{2}\mathbb{E}[v(X)]$ from the naive half-average variance, even when noise varies across successors. A constant channel bias does not affect criticality; its variation is what matters. The covariance-error bound is conservative, particularly at small σ_{res}, and is not estimated in the deployment experiments. If an outer-sampling event guarantees $|\widehat{\mathcal{C}}_{H,\mathrm{raw}} - \mathcal{C}_m| \leq t$ and a valid bound $\delta_b \leq \bar{\delta}_b$ is available, then on that event

$$\mathcal{C}_H \leq \left(\sqrt{\widehat{\mathcal{C}}_H + t} + \bar{\delta}_b\right)^2.$$

This separates the two estimation terms needed by Appendix A.7; it does not certify the empirical zero set.

Because the target is a variance, we report

$$\widehat{\mathcal{C}}_H(s) = \max\left\{\widehat{\mathcal{C}}_{H,\mathrm{raw}}(s), 0\right\}.$$

Clamping retains positive raw estimates and replaces negative ones by zero. It does not imply that the veto set is contained in the population zero set: negative error can create false vetoes at positive criticality, while positive error can prevent vetoes at zero criticality. The observed 14.1% disagreement between zero estimated current-state level and positive estimated criticality is a descriptive cross-estimator statistic, not an estimate of either population error rate.

For positive-semidefinite $\widehat{\Sigma}$ and $\sigma_{\mathrm{res}} > 0$, $I + \widehat{\Sigma}/\sigma_{\mathrm{res}}^2$ is positive definite even if $\widehat{\Sigma}$ is rank deficient. The implementation's requirement $n_{\mathrm{inner}} > 2d_w$ is a sampling heuristic, not a mathematical condition for existence of the log determinant or a guarantee of estimator accuracy.

Sensitivity to estimator seed. Reanalysis of eight saved Stack BC draw-1 estimates at $H = 2$ finds sample veto shares between 26.05% and 27.49%. Mean pairwise Jaccard similarity is 0.606 and mean pairwise membership disagreement is 13.20%; 31.55% of states change membership at least once. Thus stable share does not imply a stable veto set. At seed zero, sample share is 27.00% and inverse-inclusion-weighted share is 30.63%. These are different estimands, not competing estimates of the same sample quantity.

The primary experiments use

$$H = 2, \qquad \sigma_{\text{res}} = 0.002.$$

The noise scale changes which directions of future variation are treated as resolvable and can therefore change the ranking rather than merely rescale the score.

C Additional Controls

C.1 Direction instability across corpora and policy redraws

Figure 5 gives the full comparison summarized in Section 6.2. The cross-corpus signs are point estimates; the three matched policy redraws are compared with their own random-allocation bands.

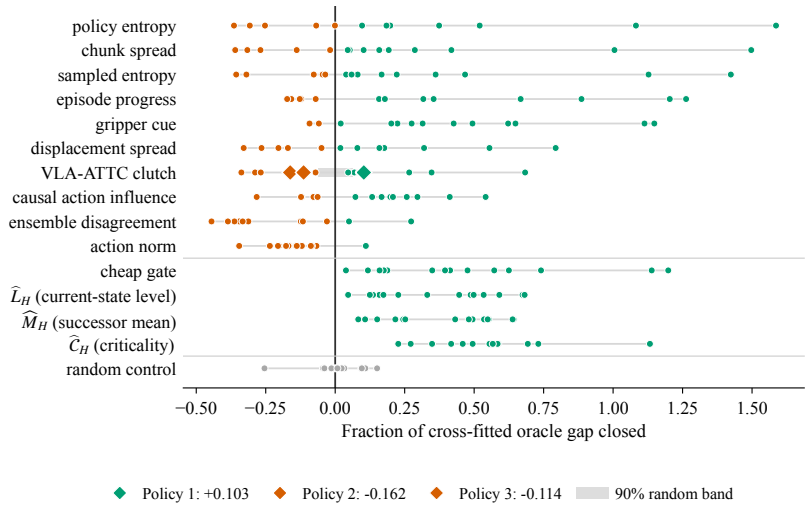

Figure 5: **The useful direction of common search-gating signals is setting-dependent.** Each point is the fraction of the cross-fitted oracle gap closed on one primary corpus at $p = 0.25$ and $H = 2$; zero denotes matched random allocation. Ten of fourteen non-random signals take both positive and negative point estimates; only the cheap gate and the three channel scores are above random on every corpus, and criticality has the highest minimum (+0.23). Diamonds show the adapted VLA-ATTC cue on three matched policy redraws ($k = 16$, $H = 1$) ($+0.103, -0.162, -0.114$), each outside its own random-allocation band.

C.2 Figure 1: recorded score profiles and estimator schematic

Figure 1 uses two unaltered 480×480 frames from one archived Stack base-policy rollout. States A and B correspond to zero-based chunk-decision indices 12 and 8 in the trace cache, respectively; these indices are not primitive simulator timesteps. Their frames match the endpoint-cache images pixel for pixel.

Each plot contains all 24 saved successor scores $\widehat{L}_i = (\widehat{L}_i^{(1)} + \widehat{L}_i^{(2)})/2$, computed from six-dimensional world positions with $H = 2$ and 16 rollouts per half. The horizontal scales are shared; vertical displacement only separates overlapping dots. State A's scores range from 0.455 to 0.798, while state B's range from 0.675 to 3.032. Their means are 0.654 and 1.975; the examples illustrate score dispersion and are not a mean-matched control.

The archive selected B by maximum estimated criticality and A by maximum current-state level among estimated-zero states in that episode. The sample variances of the displayed half-averaged scores are 0.00755 and 0.60483, whereas the archived clamped criticalities are 0 and 0.59110. These are different statistics: the latter use cross-half covariance, as drawn in panel (c). The cache does not retain the two complete half-score vectors, so no empirical half-pair scatter is plotted. The sample glyphs in panel (c) show the nested sampling design; their positions carry no spatial or numerical measurements. No new simulation was used for this redraw.

Figure 6 separately depicts the implemented deployment rule. The head takes privileged simulator state; the gripper cue uses twelve policy samples. Continuous channel targets supervise the head, while paired single-switch labels evaluate the offline veto (Section 3). Raw teacher computation is accounted for in Appendix D.9. Source arrays, hashes and drawing code accompany the figure package.

C.3 Effective-depth-matched level comparison

Because $\mathcal{C}_H$ first executes one candidate chunk and then evaluates an H-chunk channel from the successor, its effective depth from the original state is $H + 1$. The fair current-state comparison is therefore $\mathcal{C}_H$ versus L_{H+1}.

At $p = 0.25$, using the same episode-cluster bootstrap protocol as in the main text, ten corpus/horizon pairs have both quantities available at the required depths:

Table 4: Effective-depth-matched comparison $\mathcal{C}_1$ versus L_2 ($H = 1$ versus $H + 1 = 2$) on ten corpora. Positive values favor $\mathcal{C}_H$. Intervals are 95% episode-cluster bootstrap intervals.

Corpus	$\mathcal{C}_H - L_{H+1}$	95% CI
LIBERO t3 BC d1	-0.131	$[-0.273, +0.031]$
LIBERO t3 BC d2	$\mathbf{+0.109}$	$[\mathbf{+0.002}, \mathbf{+0.242}]$
LIBERO t3 BC d3	-0.070	$[-0.188, +0.061]$
LIBERO t3 VLA d1	$+0.148$	$[-0.016, +0.370]$
LIBERO t3 VLA d2	$+0.031$	$[-0.182, +0.243]$
LIBERO t4 BC d1	-0.058	$[-0.214, +0.106]$
Lift BC d1	-0.098	$[-0.210, +0.007]$
MetaWorld pp d1	$+0.079$	$[-0.108, +0.355]$
MetaWorld pp d2	$\mathbf{+0.383}$	$[\mathbf{+0.093}, \mathbf{+0.659}]$
Stack BC d1	$\mathbf{+0.188}$	$[\mathbf{+0.079}, \mathbf{+0.329}]$

Three of the ten comparisons are clearly positive, none is clearly negative, and seven are unresolved. This control is not a perfect estimator twin: L_{H+1} is estimated from the $n_{\text{outer}} = 24$ outer rollouts at depth $H + 1$, whereas the successor channel estimates inside $\mathcal{C}_H$ are formed from two halves of 16 samples. The exactly matched mean-versus-variance control is the successor mean M_H, which uses the same successors, depth, and channel rollouts as $\mathcal{C}_H$.

The primary horizon. All ten pairs above compare $\mathcal{C}_1$ with L_2. To test the primary $H = 2$, we scored four corpora at $H = 3$, each on the machine that labelled it and with its probe set verified identical to the labelled one. Table 5 reports $\mathcal{C}_2$ against L_3 at $p = 0.25$, for ranking and for the veto at criticality's own skipped share.

Table 5: Depth-matched comparison at the primary horizon, $\mathcal{C}_2$ against L_3, $p = 0.25$. Ranking is $\mathrm{GapClosed}(\mathcal{C}_2) - \mathrm{GapClosed}(L_3)$; veto is the benefit L_3 gives up minus the benefit the veto gives up at the same skipped share. Positive values favor criticality; 95% episode-cluster intervals.

Corpus	Ranking		Skipped	Veto	
Stack BC d1	$\mathbf{+0.404}$	$[\mathbf{+0.224}, \mathbf{+0.541}]$	0.306	$+0.024$	$[-0.017, +0.072]$
Meta-World pp d1	$\mathbf{+0.449}$	$[\mathbf{+0.073}, \mathbf{+0.979}]$	0.653	$+0.222$	$[-0.001, +0.403]$
Lift BC d1	$\mathbf{-0.147}$	$[\mathbf{-0.249}, \mathbf{-0.047}]$	0.422	$+0.023$	$[-0.021, +0.060]$
LIBERO t3 VLA d1	$+0.011$	$[-0.120, +0.177]$	0.572	-0.020	$[-0.109, +0.070]$

Ranking against the matched level is task-dependent: two corpora clearly favor criticality, Lift clearly favors the level, and the image policy does not resolve. On Lift the level gains most from its extra chunk (L_2 to L_3: 0.446 to 0.606), while on Stack it gains little (0.159 to 0.183). As a veto, criticality gives up less than L_3 on three of four corpora, and no single corpus resolves. One step deeper, $\mathcal{C}_3 - L_4$ is $+0.284$ $[0.123, 0.441]$ on Stack, $+0.389$ $[-0.046, 0.928]$ on Meta-World and -0.104 $[-0.216, -0.008]$ on Lift. These arms are not added to the 31-arm aggregates.

C.4 Empirical audit of the optionality defect

We audit action-value variance on Stack BC draw 1 at $H = 2$. By total variance, $\rho_H^2 \leq \mathrm{Var}_A\, Q$ in every stratum. Equality is guaranteed at population $\mathcal{C}_H = 0$, whereas our strata are selected using the clamped estimate $\widehat{\mathcal{C}}_H$. We therefore interpret the audit as estimating an upper bound on the residual variance in all groups. We sample 200 states from each of three groups and estimate action-value variance using eight candidate first chunks and eight continuation rollouts per candidate, with common random numbers and split-half debiasing.

Table 6: Audit of action-value variance across criticality strata. All groups use estimated criticality; population action-value variance upper-bounds residual variance in every group.

Group	n	$\widehat{\mathrm{Var}}_A[Q]$	95% CI
$\widehat{\mathcal{C}}_H = 0$	200	$+0.00033$	$[-0.00029, +0.00101]$
$0 < \widehat{\mathcal{C}}_H \leq$ median	200	$+0.00143$	$[+0.00025, +0.00284]$
$\widehat{\mathcal{C}}_H >$ median	200	$+0.00731$	$[+0.00408, +0.01103]$

The zero-criticality interval includes zero, whereas both positive- criticality groups are clearly above zero. The high-criticality point estimate is approximately $22\times$ the veto-group estimate. These estimates concern group-mean action-value variance. A square root of that mean bounds group RMS residual, not pointwise residual, and an interval including zero does not establish an exactly zero population value.

To separate the criticality veto from the current-state level veto, the 200 zero-criticality states can be partitioned further:

Table 7: Decomposition of the audited zero-criticality set. The "extra veto" consists of states rejected by criticality but not by the current-state level.

Subset	n	$\widehat{\mathrm{Var}}_A[Q]$	95% CI	mean $\widehat{\Delta}$
$\widehat{\mathcal{C}}_H = 0,\ \widehat{L}_H = 0$	131	$+0.00056$	$[-0.00028, +0.00149]$	$+0.00094$
$\widehat{\mathcal{C}}_H = 0,\ \widehat{L}_H > 0$	69	-0.00014	$[-0.00096, +0.00064]$	-0.00505

The extra-veto subset again has action-value variance statistically compatible with zero in the underlying audit, while its mean estimated search benefit is slightly negative. These measurements are specific to one corpus and policy draw and are not used as a universal guarantee.

D Experimental Setup and Reproducibility

This appendix gives the implementation and registry details omitted from the main text. A *corpus* is one labelled probe-state dataset; an evaluation *arm* re-scores a corpus at one estimator setting. The registry contains **31 arms over 17 corpora**. The primary ranking benchmark uses the 14 corpora available at $H = 2$ and $p = 0.25$, while the veto analysis uses all 31 arms. The four $H = 3$ arms of Appendix C.3 are reported separately and are not part of this registry.

D.1 Corpus and arm registry

Table 8 reports the primary $H = 2$ corpora. Here, N is the number of labelled probe states, $\overline{\Delta}_{\mathcal{S}}$ is the mean paired search benefit. The last column is the unweighted *sample* veto share; the headline

skipped share uses inverse inclusion weights. The random column symmetrizes the 90% allocation band using the larger absolute endpoint; it is not a 95% confidence interval.

Table 8: Primary $H = 2$ corpora used for ranking at $p = 0.25$. The random column is the matched random-allocation band for GapClosed.

Corpus	N	$\overline{\Delta}_S$	GapClosed	Random	Sample veto
robosuite					
Stack BC d1	4460	0.0190	0.587	±0.097	0.270
Stack BC d2	5064	0.0255	0.584	±0.080	0.227
Stack BC d3	5027	0.0188	0.564	±0.090	0.199
Lift BC d1	3972	0.0203	0.459	±0.089	0.386
Lift BC d2	4205	0.0148	0.227	±0.087	0.323
Lift BC d3	3937	0.0100	0.417	±0.123	0.379
LIBERO					
LIBERO t3 BC d1	4664	0.0088	0.555	±0.121	0.522
LIBERO t3 BC d2	4688	0.0132	0.731	±0.113	0.515
LIBERO t3 BC d3	4810	0.0101	0.495	±0.119	0.471
LIBERO t4 BC d1	4945	0.0118	0.567	±0.144	0.515
LIBERO t3 VLA d1	4092	0.0100	0.349	±0.151	0.546
LIBERO t3 VLA d2	4013	0.0119	0.272	±0.138	0.567
Meta-World					
MetaWorld pp d1	4000	0.0030	1.132	±0.300	0.632
MetaWorld pp d2	4000	0.0042	0.693	±0.194	0.638

The 31-arm registry is more compactly shown by grouping arms that share the same corpus. Each listed horizon is one arm; the horizon counts therefore sum to 31.

Table 9: Full evaluation registry grouped by corpus. Each entry in the H column denotes one arm. In arm names, "VLA" denotes the image-conditioned flow-matching policy, not a pretrained foundation model.

Corpus	H	chunk	Artifact ID
robosuite			
Stack BC d1	$\{1, 2, 4, 8\}$	4	`f1_stack_big`
Stack BC d2	$\{2\}$	4	`f1_stack_s1`
Stack BC d3	$\{2\}$	4	`f1_stack_s2`
Lift BC d1	$\{1, 2, 4\}$	4	`f1_lift_v3`
Lift BC d2	$\{2\}$	4	`f1_lift_s1`
Lift BC d3	$\{2\}$	4	`f1_lift_s2`
LIBERO			
LIBERO t3 BC d1	$\{1, 2\}$	4	`f1_libero_full`
LIBERO t3 BC d2	$\{1, 2\}$	4	`f1_libero_s1`
LIBERO t3 BC d3	$\{1, 2\}$	4	`f1_libero_s2`
LIBERO t4 BC d1	$\{1, 2\}$	4	`f1_libero_t4`
LIBERO t3 VLA d1	$\{1, 2\}$	4	`f1_vla_libero_object_3`
LIBERO t3 VLA d2	$\{1, 2\}$	4	`f1_vla_libero_object_3_d1`
LIBERO t3 VLA c16 d1	$\{1\}$	16	`f1_vla_libero_object_3_c16`
LIBERO t3 VLA c16 d2	$\{1\}$	16	`f1_vla_libero_object_3_d1_c16`
LIBERO t3 VLA c16 d3	$\{1\}$	16	`f1_vla_libero_object_3_d2_c16`
Meta-World			
MetaWorld pp d1	$\{1, 2, 4\}$	4	`f1_pp_s1`
MetaWorld pp d2	$\{1, 2\}$	4	`f1_pp_s2`

The primary 14-corpus subset keeps stamped $H = 2$ arms that provide the cheap gate, then deduplicates by labels path. The `c16` corpora are excluded because their available arm has $H = 1$. Comparisons additionally require matching H, σ_{res}, n_{inner}, n_{outer}, planner, and chunk length.

D.2 Empirical search operator and paired labels

All causal labels are specific to `PrivilegedPlanner(n_candidates=8, n_eval=8)`. At a probe state, the operator draws eight candidate chunks from the frozen policy. Each candidate is

evaluated with eight simulator rollouts that execute the candidate and then follow the same frozen policy to episode termination. A rollout receives the binary ever-success value $\max_t \mathbf{1}\{\text{success at } t\}$, and the candidate with the highest mean value is selected. Thus the evaluator uses neither shaped reward nor a learned value bootstrap.

Ties. The planner uses standard first-occurrence `argmax` tie-breaking. Since candidate chunks are i.i.d. draws from the same prior, this is distributionally equivalent to uniform tie-breaking. The analysis-side top-p selector instead randomises ties. A separate 600-state stratified Stack audit redraws the evaluator and, using inverse-inclusion weights and stratum masses, estimates an argmax tie rate of 84.5% and an all-zero rate of 57.8% for the source episode population. These are independent evaluator draws, not the evaluations that generated each paired label. In the same audit, the weighted $|\widehat{\Delta}|$ ratio between above-median criticality and the veto set is 3.24 over all states and 1.79 once states on which no candidate scores are removed; 65% $[53\%, 76\%]$ of the excess ratio reflects the evaluator failing to rank. On the non-degenerate states the veto set still shows no detectable benefit ($\widehat{\Delta} = -0.0049$, $[-0.0285, 0.0192]$, $n = 75$), whereas the two positive-criticality strata carry $+0.058$ and $+0.060$ and are indistinguishable from each other.

Paired labels. Each probe state uses $K = 12$ paired single-switch rollouts,

$$\widehat{\Delta}_{\mathcal{S}}(s) = \frac{1}{K} \sum_{j=1}^{K} \left(J^{(j)}_{\text{search}} - J^{(j)}_{\text{no-search}} \right).$$

Both branches share the continuation protocol and common random numbers. Seeds are counter-based and keyed by (`base_seed`, `state_id`, `rep`, `counter`); the branch is intentionally omitted from the key. Candidate chunks are drawn from one shared stream, while evaluation rollouts are keyed separately by candidate and rollout index. Random-allocation bands use 4,000 matched draws and report the 5th–95th percentiles (90%). Paired intervals use 4,000 episode-cluster bootstrap resamples (95%). The oracle reference leaves out one paired repeat at each state, selects with the remaining eleven, scores with the held-out repeat and averages over twelve splits. Distillation, separately, uses held-out episode folds. Planner evaluation and outcome continuation use distinct random streams.

Weighted ranking formulas. Let $b_{r,i}(p)$ indicate the selected states and let $\widehat{\Delta}_i = K^{-1} \sum_j D_{ij}$, where D_{ij} is paired return difference. The implementation computes

$$\mathcal{B}_r(p) = \frac{\sum_i w_i \widehat{\Delta}_i b_{r,i}(p)}{\sum_i w_i}, \qquad \mathcal{B}_{\text{rand}}(p) = p \frac{\sum_i w_i \widehat{\Delta}_i}{\sum_i w_i}.$$

For oracle split j, let $b_{-j,i}(p)$ select using the other $K - 1$ repeats. Then $\mathcal{B}_{\text{oracle}}(p) = K^{-1} \sum_j [\sum_i w_i D_{ij} b_{-j,i}(p) / \sum_i w_i]$. The discrete top-$p$ selector never exceeds target mass p and uses a fixed pseudorandom tie permutation. Thus realized selected mass may fall below p, whereas the random reference above uses nominal p. Allocation bands use the same discrete selector; exact-mass sensitivity is a separate check. Recomputing both signal and oracle at realized mass shifts mean GapClosed by between -0.00003 and $+0.00021$ across four signals and three coverages (largest single-arm shift 0.00102); the floor removes at most one state, equally for every signal, and reorders nothing.

Discrete veto matching. Archived baseline veto summaries sort scores stably and include the first state whose cumulative inverse-inclusion weight reaches the target veto mass. Their achieved mass can exceed the target by at most one sampled state's weight, and boundary ties retain stored array order. This differs from the randomized tie rule of the ranking selector; the reported veto comparisons use the archived discrete-cutoff convention.

All ranking and veto conclusions are operator-specific: changing the candidate budget, evaluator, or continuation protocol changes $\Delta_{\mathcal{S}}$ and may change the ordering of gates.

D.3 GATING BASELINES

All signals use one orientation: larger score means "search here." No signal is re-oriented per corpus; the only explicit negation is episode progress, $-t$. Sample-based proxies share the same twelve policy draws per state. Table 10 gives the exact scores.

Table 10: Gating signals. For sampled chunk $A^{(i)}$, $u^{(i)}_{\tau,d}$ is action coordinate d at within-chunk step τ; P indexes positional coordinates and d_g the gripper. Larger scores request search.

Signal	Score	Availability / source
Cheap and sample-based		
chunk spread	$\mathrm{mean}_d \, \mathrm{sd}_i \, u^{(i)}_{0,d}$	12 samples
sampled entropy	$\frac{1}{2} \sum_d \log(2\pi e \, \widehat{v}_d)$	12 samples
episode progress	$-t$	clock
gripper cue	$\mathrm{sd}_i \, u^{(i)}_{0,d_g}$	12 samples
displacement spread	$\sqrt{\mathrm{tr} \, \mathrm{Cov}_i[\sum_{\tau=0}^{k-1} u^{(i)}_{\tau,P}]}$	12 samples
ATTC-style clutch (adapted)	$\mathrm{DTW}(A^{(1)}_{:,P}, A^{(2)}_{:,P})$	2 samples
action norm	$\mathrm{mean}_i \lVert u^{(i)}_0 \rVert_2$	12 samples
cheap gate	$\mathrm{rank}(-t) + \mathrm{rank}(u_g)$	clock + samples
Policy-specific		
policy entropy	$\frac{1}{2} \sum_{\tau,d} \log(2\pi e \sigma^2_{\tau,d})$	analytic BC only
ensemble disagreement	$\mathrm{mean}_{\tau,d} \, \mathrm{sd}_m \, \mu^{(m)}_{\tau,d}$	BC ensemble only
Simulator / structural		
one-step influence proxy	one-action channel score (below)	simulator + samples
current-state level	$\widehat{L}_H(s)$	simulator + prior
successor mean	$\widehat{M}_H(s) = n^{-1}_{\mathrm{outer}} \sum_i \widehat{L}_i$	simulator + prior
criticality	$\widehat{\mathcal{C}}_H(s)$, Eq. 7	simulator + prior

Here $\widehat{L}_i = (\widehat{L}^{(1)}_i + \widehat{L}^{(2)}_i)/2$; u_g denotes the gripper cue, not an action coordinate. For sampled entropy, $\widehat{v}_d = \max\{\widehat{\mathrm{Var}}_i[u^{(i)}_{0,d}], 10^{-12}\}$ uses sample variance with divisor 11. The spread and gripper standard deviations use divisor 12; displacement covariance uses 11, and ensemble dispersion uses the number of members. Analytic entropy and ensemble disagreement span the full chunk. Here $\sigma_{\tau,d}$ is the pre-clipping Gaussian scale, so analytic entropy is a Gaussian proxy; $\mu^{(m)}$ is member m's clipped mean chunk. Policy entropy is unavailable for the flow-matching policy because it has no closed-form $\sigma_{\tau,d}$; ensemble disagreement is unavailable for a single network. These entries are missing, not zero. The offline cheap gate uses average ranks for ties. The ATTC-style clutch adapts the published cue: DTW between the positional coordinates of two chunks sampled from our policy. It does not include VLA-ATTC's relative action critic, so it is a signal comparison, not a comparison with the full method.

The one-step influence proxy, inspired by causal action influence (Seitzer et al., 2021), draws first actions $u^{(i)}_0$, executes one primitive simulator step, and applies the same Gaussian log-determinant estimator to the resulting world endpoints. It is the one-action analogue of Eq. 4, not a full policy chunk with $H = 1$ and not an optimized empowerment capacity.

D.4 CRITICALITY ESTIMATOR AND WORLD-STATE REPRESENTATION

The raw teacher uses $n_{\mathrm{outer}} = 24$ candidate chunks, $n_{\mathrm{inner}} = 32$ channel rollouts per successor, $H = 2$, and $\sigma_{\mathrm{res}} = 0.002$ in the primary setting. The 32 inner samples are split into two disjoint halves; criticality is estimated by cross-half covariance and clamped at zero. A reported zero means only that the raw covariance is non-positive. No significance test or confidence bound is used to define this veto set; Appendix B details its limitations.

The channel is evaluated on z^{world} only, excluding robot configuration. Table 11 lists the exact observation slices.

Table 11: World-state coordinates used by the channel estimator. All listed components are absolute world-frame positions in metres.

Suite / task	obs. dim.	d_w	indices	components
robosuite Stack	50	6	$(7, 8, 9, 17, 18, 19)$	cubeA xyz, cubeB xyz
robosuite Lift	40	3	$(7, 8, 9)$	cube xyz
LIBERO object:3	50	6	$(7, 8, 9, 17, 18, 19)$	bbq sauce xyz, basket xyz
LIBERO object:4	50	6	$(7, 8, 9, 17, 18, 19)$	ketchup xyz, basket xyz
Meta-World pick-place-v3	39	3	$(4, 5, 6)$	object COM xyz

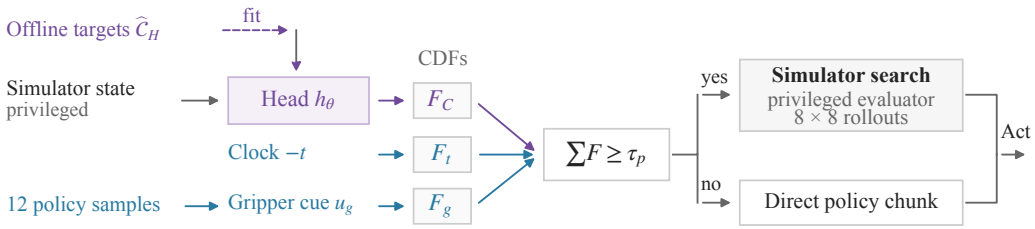

Figure 6: **Distilled online allocation.** The head, clock and sampled gripper cue enter separate empirical CDFs; their sum selects simulator search or a direct policy chunk. Dashed supervision fits the head offline; online inputs and the privileged evaluator are shown explicitly.

No per-dimension standardisation or whitening is applied on this scoring path. Therefore $\mathrm{Cov}(z)$ and σ^2_{res} are both in m^2; $\sigma_{\mathrm{res}} = 0.002$ corresponds to a two-millimetre resolution floor and can change state ordering rather than merely rescale the score. Quaternions and end-effector-relative vectors are excluded from z^{world}. Restricting the channel to this projection is an empirical design choice. The level theorem requires value to factor through the chosen projected world endpoint at the stated resolution; omitted robot state can violate that assumption. The criticality theorem retains unexplained variation in its residual term.

Finally, $\mathcal{C}_H$ reaches one additional chunk from the original state relative to L_H. The effective-depth comparison is therefore $\mathcal{C}_H$ versus L_{H+1}, while M_H is the estimator-matched mean-versus-variance control using the same successors and channel rollouts.

D.5 Distilled deployment predictor

Raw $\mathcal{C}_H$ is used as an offline teacher. Closed-loop experiments use a scalar regressor h_θ trained on $y_H(s) = \log(1 + \widehat{\mathcal{C}}_H(s)/\varsigma)$, with the configuration in Table 12.

Table 12: Distillation and calibration details for h_θ.

Item	Setting
Input	raw low-dimensional simulator observation
Preprocessing	per-feature standardisation, fitted within training fold
Architecture	$d \to 128 \to 128 \to 1$, $\tanh$ hidden activations
Parameters	23,169 at $d = 50$; 21,889 at $d = 40$
Target	$\log(1 + \widehat{\mathcal{C}}_H/\varsigma)$, median-positive teacher scale ς
Loss	mean-squared error
Optimizer	Adam, learning rate 10^{-3}, batch size 256
Training	150 epochs, no weight decay, schedule, dropout, or early stopping
Cross-fitting	5 folds over episodes; fold-local standardisation
Image-policy OOF agreement	Spearman $\rho_{\mathrm{Sp}} = 0.573$ on 3,793 states / 984 episodes
Calibration	empirical CDF per component; composition-specific top-p threshold

Spearman uses average ranks for ties. The deployed gate uses privileged low-dimensional simulator state even when the underlying policy is image-conditioned; it does not run from pixels alone. The policy implementation also supports a simulator-state branch and has no language input; these experiments are not evidence for foundation-VLA deployment. For the committed image-policy checkpoint, input dimension is 50 and $\varsigma = 0.170$. Its saved teacher targets use $H = 1$, chunk length $k = 16$, and 3,793 states; the primary offline comparison uses $H = 2$, $k = 4$. The deployed head output is the raw regression score (the code method is `logit`); no sigmoid or inverse target transform is applied before its CDF. Live evaluation episodes use reset seeds disjoint from the training corpus. The deployed fold is chosen by convention, while the calibration CDF is fit on the full offline corpus; a shifted deployment distribution therefore requires recalibration.

On the historical Stack cohort (300 resets, 5000–5299; not the 600-episode test of Section 7), h_θ alone reaches 77.7% success at nominal $p = 0.25$, the cheap gate reaches 84.3%, and their combination reaches 87.3% with 7.95 searches per episode. The targeted spend-matched cheap gate uses 8.40 searches and reaches 81.7%; the paired success difference is $+0.0567$ with 95% CI $[0.0148, 0.0985]$. This matched threshold was selected post hoc after the nominal-coverage comparison. At nominal $p = 0.25$, the paired combined-minus-cheap difference is $+0.0300$ with 95% CI $[-0.0087, +0.0687]$.

Paired deployment intervals. For N matched reset seeds, let n_+ and n_- count episodes rescued and broken by the first gate relative to the comparator. For historical runs, we report $\widehat{\delta}_{\text{succ}} = (n_+ - n_-)/N$ with the normal-approximation interval $\widehat{\delta}_{\text{succ}} \pm 1.96\sqrt{n_+ + n_-}/N$. These closed-loop intervals use episode-level discordances, not the offline episode-cluster bootstrap; the Stack comparisons use $N = 300$.

Gate composition. Let F_t, F_g, F_C be right-continuous empirical CDFs on the calibration corpus. For a component with calibration values $v_1, \ldots, v_{N_{\text{cal}}}$, $F(x) = N_{\text{cal}}^{-1} \sum_i \mathbf{1}\{v_i \le x\}$. Writing the elapsed timestep explicitly, the combined score and gate are

$$R_{\text{comb}}(s,t) = F_t(-t) + F_g(u_g(s)) + F_C(h_\theta(s)), \qquad g_{\text{comb}}(s,t) = \mathbf{1}\{R_{\text{comb}}(s,t) \ge \tau_p\}.$$

Here τ_p is that composition's empirical $(1-p)$ quantile, computed with linear interpolation. The cheap gate omits F_C, the head-only gate uses F_C, and clock+head omits F_g; each has its own τ_p. Offline average ranks and online right-continuous CDFs handle ties differently. Nominal p need not equal calibration coverage or realized online search frequency. The compact notation $g(s)$ elsewhere includes elapsed time among the gate inputs.

Image-policy diagnostics and random control. The historical 13-regime correlation 0.989 uses realized search count multiplied by offline conditional benefit. Excluding the shared-schedule random control leaves 12 regimes with correlations 0.993 (spend-adjusted), 0.937 (fully offline additive), and 0.972 (realized calls alone). The offline additive denominator is all 1,000 collected episodes, including episodes without a retained probe. Thresholds are read from saved gate configurations. These are reanalyses of existing artifacts, not new rollouts. The fully offline additive estimate is $\widehat{\mathcal{B}}_{\text{off}}(g) = N_{\text{ep}}^{-1} \sum_i w_i \widehat{\Delta}_i g_i$, where g_i is offline selection and N_{ep} counts all collected episodes. It ignores the state-distribution shift caused by repeated gating.

Corrected episode-randomized control. The archived random gate keyed decisions by base seed and timestep only, reusing one sampled schedule. Corrected runs also key by reset seed (`per_episode_v1`), keeping the recorded environment, policy, planner and base-seed configurations fixed. Each rerun has the same 300 reset IDs (5000–5299) as its historical comparator. Fixed-schedule outcomes are not relabelled as corrected data. Table 13 reports the correction separately from the new held-out test. For the image policy, combined $p = 0.25$ versus corrected random is $+4.33$ points (95% normal-discordance CI $[-0.33, 9.00]$), an unresolved difference. The twelve non-random regimes in the correlation diagnostic are unchanged.

D.6 Teacher controls and veto sensitivity

The Stack teacher comparison uses 4,460 states from 250 episodes, five episode folds (split seed 3), training seeds 100, 200 and 300, and the same $50 \to 128 \to 128 \to 1$ architecture and optimizer

Table 13: Corrected random allocation at $p = 0.25$, 300 historical reset IDs per setting. Each cell gives success (%) / mean searches per episode. Combined is the historical $p = 0.25$ gate, not the new 600-episode test.

Setting	Legacy random	Corrected random	Historical combined
Stack	68.67 / 11.33	66.33 / 10.33	87.33 / 7.95
Lift	62.67 / 10.60	68.67 / 8.32	94.67 / 9.48
Image policy	47.67 / 3.06	50.67 / 2.53	55.00 / 2.29

as the criticality head. The deployed log target is $\log(1 + x/\operatorname{median}(x_{x>0}))$ for each nonnegative teacher. The alternative target is the teacher's average rank divided by $N - 1$. The target transform (median scale or ranks) and observation standardization are both fitted within each training fold, so no held-out value enters a training target. Predictions are out of fold, and the corpus itself is distinct from closed-loop test resets.

Table 14: Teacher sensitivity: mean offline GapClosed over three training seeds. All tested coverages and both target transforms are shown. These are point estimates conditional on the offline corpus; the closed-loop frontier is in Appendix D.8.

		Head only			Cheap + head		
Target	Teacher	.10	.25	.50	.10	.25	.50
Log	C	.092	.298	.755	.382	.639	.853
Log	M	-.003	.070	.424	.358	.627	.898
Log	random	-.076	-.106	-.040	.161	.348	.745
Rank	C	.130	.344	.791	.377	.630	.908
Rank	M	.008	.079	.557	.387	.640	.865
Rank	$\widehat{\Delta}$	.149	.261	.462	.281	.473	.891
Rank	random	-.086	-.134	-.069	.145	.339	.720
Cheap alone					.250	.476	.904

Table 2 reports sample SD across seeds. For the paired criticality-minus-mean difference, one episode-cluster bootstrap (4,000 resamples, seed 7) resamples the same episodes for all three head seeds, averages the per-seed differences within each draw, and takes percentile 95% intervals. With the deployed log target at $p = 0.25$, the standalone difference is $+0.227$ $[0.101, 0.361]$ (seed SD 0.078) and the difference inside the cheap gate is $+0.012$ $[-0.088, 0.101]$; with rank targets they are $+0.265$ $[0.141, 0.416]$ and -0.010 $[-0.097, 0.083]$. These intervals condition on the fitted heads; the seed SD describes variation across training initializations on the same corpus. The rank-target $\widehat{\Delta}$ head uses outcome supervision and is a separate comparator, not an oracle upper bound.

The Stack closed-loop mean-head control uses the same 600 reset IDs 20000–20599, environment, base policy and best-of-eight evaluator as the criticality test. Each teacher has its own CDF and $p = 0.25$ threshold. Criticality has 499 successes, mean has 492, and cheap has 457; criticality versus mean has 34 rescues and 27 breaks. Its paired percentile interval uses the protocol of Appendix D.7 (10,000 matched-ID resamples, seed 20260924), giving $[-1.33, 3.67]$ points. The mean-head gate uses 8.6483 calls per episode; it was not independently pilot-matched to the criticality gate. The three-teacher offline comparison therefore complements, rather than replaces, the closed-loop control.

The image-policy mean-head control repeats this on the LIBERO test resets 20000–20599. Its head differs from the deployed one only in the teacher (same labels, folds and training seed). Calibration clock and gripper columns are bit-identical to the criticality arm, and the stored head column matches the mean head's forward pass to 10^{-15}. The mean head reaches 348/600 (58.00%) at 2.48 calls per episode, against 355 (59.17%) at 2.26 for criticality and 330 (55.00%) at 2.84 for cheap. Criticality versus mean has 29 rescues and 22 breaks: $+1.17$ points $[-1.17, 3.50]$, with -0.228 calls per episode $[-0.355, -0.095]$. Mean versus cheap is $+3.00$ points $[0.67, 5.50]$. The call saving is uneven: on episodes both heads solve, criticality calls search 28% less, and on episodes both fail it calls 33% more. Episode lengths are not stored, so this split does not separate allocation from episode length. Appendix D.8 extends this comparison to three budgets and three head seeds.

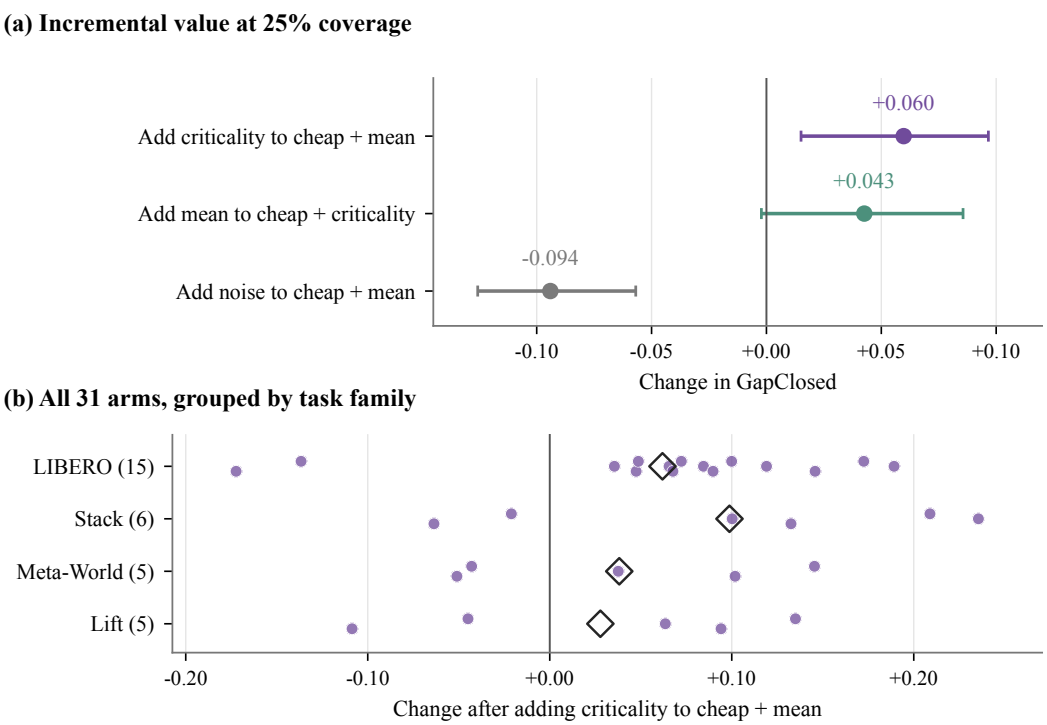

Figure 7: **What criticality adds beyond the cheap cues and successor mean.** (a) Changes in offline GapClosed at $p = 0.25$ using raw estimated channel statistics, with 95% bootstrap intervals over 17 corpus clusters. (b) All 31 arm-level criticality increments, grouped descriptively by task family. Intervals condition on the saved arm estimates; these are not deployed-head or matched-compute results.

Incremental value over the successor mean. Figure 7 shows all three contrasts and the arm-level variation. We committed one contrast to version control before its first run (the commit timestamp precedes it; not externally verified), with no subgroups: the oracle-statistic GapClosed of the rank-sum gate clock + gripper + $\widehat{M}_H$ + $\widehat{\mathcal{C}}_H$ minus that of clock + gripper + $\widehat{M}_H$, at $p = 0.25$ over all 31 arms. It is $+0.0597$, 17-corpus interval $[0.0151, 0.0966]$, positive on 23/31 arms. Because a rank sum dilutes its components, an uninformative column is not neutral: adding a fixed noise column instead gives -0.0941 $[-0.1257, -0.0569]$, so criticality exceeds the null by $+0.1538$ $[0.1045, 0.1962]$. The reverse increment, $\widehat{M}_H$ added to a gate already containing $\widehat{\mathcal{C}}_H$, is $+0.0426$ $[-0.0022, 0.0856]$. At $p = 0.10$ the increment is $+0.031$ $[-0.023, 0.077]$ and at $p = 0.50$ it is $+0.093$ $[0.006, 0.198]$. On the Stack BC d1, $H = 2$ arm used for the Stack deployment the increment at $p = 0.25$ is negative on all eight estimator seeds (mean -0.042, SD 0.032), consistent with the unresolved closed-loop difference above. Across those eight seeds criticality's pairwise Spearman correlation is 0.883 (successor mean 0.996), and the Jaccard overlap of the selected sets is 0.702, 0.741 and 0.809 at $p = 0.10, 0.25, 0.50$, above the 0.606 of the veto set itself.

Veto aggregation sensitivity. The 31-arm mean shares of net benefit given up are 0.1990 (criticality), 0.2578 (successor mean), 0.2349 (cheap) and 0.3338 (current-state level). The successor-mean-minus-criticality difference is 0.0587; a bootstrap of the 17 corpus clusters gives $[0.0084, 0.1203]$. This sensitivity resamples saved arm estimates together within their corpus and retains arm weighting; it does not resample episodes or paired labels within corpora and does not establish generalization to independent tasks. On the ten matched-depth pairs $\mathcal{C}_1$ versus L_2, the analogous difference is 0.0426 with interval $[-0.0122, 0.1078]$. On the same ten pairs the deeper level itself gives up less: L_1 minus L_2 is 0.122 $[0.087, 0.161]$, positive on 10/10 arms (paired sign-flip test, $P = 0.0021$). We therefore use the depth-matched level L_{H+1} for the controlled comparison. Using the positive

part of noisy benefit estimates instead of net benefit gives shares given up of 0.3029 for criticality, 0.3470 for successor mean and 0.2274 for cheap. Because truncation amplifies label noise, this is a sensitivity diagnostic rather than a replacement ground-truth measure.

D.7 Pilot selection and held-out deployment tests

The local protocol and pilot selection artifacts specify LIBERO object:3 with the frozen image-conditioned flow-matching checkpoint (chunk length 16, ten sampling steps), the existing fold-0 criticality head, and the same best-of-eight privileged planner (eight evaluator rollouts per candidate). The task was selected using earlier offline evidence; it is not an unseen-task evaluation. The local protocol documents prospective intent, but its timestamp is not an independently verified pre-registration.

Pilot resets are 10000–10059. Combined coverage is fixed at $p = 0.25$; the cheap gate is selected from $\{0.15, 0.20, 0.25, 0.30, 0.35\}$ by nearest mean search count. The recorded selection uses calls, not success. The selected $p = 0.30$ gives 2.8167 calls versus 2.7000 for combined, within the 5% pilot tolerance. The saved hashes match all six pilot files. Component CDF arrays and head specification are identical between pilot, test and timing; combined and cheap thresholds are 1.912998 and 1.242658.

All three test gates have 600 unique reset IDs 20000–20599 and matching environment, policy, planner and base-seed configurations. Success counts are 355 (combined), 330 (cheap) and 229 (never). We resample matched episode IDs jointly across gates 10,000 times (seed 20260924) and take percentile 95% intervals. Combined minus cheap has 41 rescues and 16 breaks; its success interval is $[1.67, 6.67]$ percentage points and its search-count difference is -0.582 calls per episode ($[-0.698, -0.467]$). The test call counts therefore do not preserve the pilot match. Both search allocation and episode length affect realized calls; these results do not identify an effect at controlled equal spend.

Completed Stack replication. The frozen Stack BC ensemble uses the same $(8, 8)$ privileged planner, chunk length 4 and episode horizon 250. Combined $p = 0.25$ is fixed. The 60-episode pilot (reset IDs 10000–10059) extends the cheap grid once at step 0.01 over 0.21–0.24, as documented in the local selection record; $p = 0.22$ minimizes call mismatch: 8.30 versus 8.25 calls (0.61%). Test uses 600 disjoint reset IDs 20000–20599. All three gates have the same unique IDs and environment, policy, planner and base-seed configuration. Pilot and test component calibration arrays match exactly; selected gate thresholds and all ten pilot-file hashes match the selection record.

Table 15: Completed Stack replication on 600 paired held-out resets. These are the new pilot-selected gates, separate from historical Table 13.

Gate	Successes	Success (%)	Calls / episode
Never	271/600	45.17	0.00
Cheap ($p = 0.22$)	457/600	76.17	8.32
Combined ($p = 0.25$)	499/600	83.17	8.39

Combined rescues 77 episodes and breaks 35. The paired success difference is $+7.00$ points, with percentile bootstrap 95% CI $[3.67, 10.33]$ (10,000 matched-ID resamples, seed 20260924). Mean calls differ by $+0.070$ ($[-0.382, 0.540]$); their ratio is 1.0084 ($[0.9545, 1.0655]$). The observed means are close, but this ratio interval is not contained in $[0.95, 1.05]$: equivalence within 5% is not established. Nor do call counts establish equal wall-clock cost. This replication supports the combined gate on a second setting; it does not isolate criticality from alternative same-input learned teachers.

D.8 Image-policy success–calls frontier

Two protocols were committed to version control before their episodes ran (commit timestamps precede the first episode; they are not externally verified). The first (`docs/protocol_frontier_vla.md`) runs the three gates at nominal $p = 0.10$, 0.25 and 0.50 on the test resets 20000–20599, reusing the $p = 0.25$ head arms and *never*. It registered two predictions: both head gates have higher

success than cheap at every p, and the criticality head uses fewer calls than the mean head at every p. The first holds at $p = 0.10$ and 0.25 but not at 0.50, where all three gates reach about 68% and the head gates use about one call fewer per episode. The second holds only at $p = 0.25$. Table 16 gives all points and Table 17 the paired contrasts (10,000 matched-ID resamples, seed 20260924). Nominal p does not equalize calls, so contrasts at a nominal p compare different spends.

Table 16: Image-policy frontier: success (%) / realized calls per episode on 600 paired resets, head seed 100. *Never* reaches 38.17%; the paper's cheap $p = 0.30$ arm reaches 55.00 / 2.837.

Gate	$p = 0.10$	$p = 0.25$	$p = 0.50$
Cheap	39.83 / 0.248	50.67 / 1.752	68.67 / 6.017
Cheap + $\widehat{\mathcal{C}}$ head	48.00 / 0.787	59.17 / 2.255	68.17 / 4.957
Cheap + $\widehat{M}$ head	44.83 / 0.622	58.00 / 2.483	67.83 / 5.033

Table 17: Paired frontier contrasts, head seed 100: success (points) and calls per episode, 95% intervals.

p	Contrast	Success	Calls
0.10	$\widehat{\mathcal{C}} - \widehat{M}$	$+3.17$ $[+0.83, +5.50]$	$+0.165$ $[+0.085, +0.247]$
0.10	$\widehat{\mathcal{C}}$ − cheap	$+8.17$ $[+5.67, +10.67]$	$+0.538$ $[+0.478, +0.600]$
0.10	$\widehat{M}$ − cheap	$+5.00$ $[+3.17, +7.00]$	$+0.373$ $[+0.313, +0.435]$
0.25	$\widehat{\mathcal{C}} - \widehat{M}$	$+1.17$ $[-1.17, +3.50]$	-0.228 $[-0.355, -0.095]$
0.25	$\widehat{\mathcal{C}}$ − cheap	$+8.50$ $[+5.67, +11.33]$	$+0.503$ $[+0.380, +0.623]$
0.25	$\widehat{M}$ − cheap	$+7.33$ $[+4.83, +9.84]$	$+0.732$ $[+0.567, +0.890]$
0.50	$\widehat{\mathcal{C}} - \widehat{M}$	$+0.33$ $[-1.33, +2.00]$	-0.077 $[-0.202, +0.057]$
0.50	$\widehat{\mathcal{C}}$ − cheap	-0.50 $[-3.17, +2.17]$	-1.060 $[-1.188, -0.927]$
0.50	$\widehat{M}$ − cheap	-0.83 $[-3.50, +1.83]$	-0.983 $[-1.127, -0.843]$

With one head per teacher the criticality–mean contrast cannot be told from head-seed variation, so an addendum (`docs/protocol_frontier_vla_seeds.md`), committed before it ran and stated as motivated by the result above, retrained both heads at seeds 200 and 300 and reran $p = 0.10$ and 0.25. Retraining at seed 100 reproduces the deployed criticality head bit-identically, so the new heads differ only in the seed. The registered result pools the three seeds: one matched-ID resample (10,000 draws) averages the per-seed differences inside each draw (Table 18). These intervals resample resets with the three fitted heads held fixed; they do not describe a population of head seeds. Both registered expectations hold: at $p = 0.10$ criticality has higher success at calls not distinguishable from the mean head's, and at $p = 0.25$ fewer calls at success not distinguishable from it; every seed has the same sign on both. The $p = 0.10$ lower bound is close to zero, and the result covers one task.

Table 18: Criticality minus mean head, pooled over head seeds 100, 200, 300; per-seed values in seed order.

p	Success (points)	Per seed	Calls	Per seed
0.10	$+2.17$ $[+0.17, +4.11]$	$+3.17, +2.33, +1.00$	$+0.048$ $[-0.023, +0.120]$	$+0.165, +0.120, -0.140$
0.25	$+0.67$ $[-1.44, +2.83]$	$+1.17, +1.00, -0.17$	-0.257 $[-0.365, -0.144]$	$-0.228, -0.115, -0.428$

D.9 Compute accounting

On the primary Stack configuration (chunk length 4, episode horizon 250), the full scoring profile (criticality together with the current-state level, successor mean and one-step influence proxy computed alongside it) costs 6,360 environment steps and 1,608 policy samples per state. By count from the estimator, criticality alone needs 6,240 steps and 1,560 samples: the first chunk of each outer rollout plus the two halves at each successor. Search cost depends on decision time because each of the eight candidates is evaluated with eight rollouts to episode termination. Instrumented costs are

reported in Table 19; Horvitz–Thompson weighting over the primary corpus gives 9,006 environment steps per search invocation, so the scoring profile costs 0.71 of one privileged search call on average (0.69 for criticality alone).

Table 19: Measured environment-step cost of one privileged search call. The full scoring profile costs 6,360 steps at every decision time.

Decision time t	Search steps	Profile/search
20	13,396	0.47
60	11,240	0.57
120	7,516	0.85
180	3,436	1.85
HT-weighted mean	9,006	0.71

These measurements show that the raw scoring profile is expensive relative to the search it gates. Stack's sample veto share is 27.0%, whereas the weighted skipped share is 30.63%. Even under equal search cost per decision, the latter would allow gate cost of only 0.3063 average calls, below the measured 0.71. With timestep-dependent costs, the correct break-even condition is instead

$$\mathbb{E}[c_{\text{gate}}(S)] < \mathbb{E}[\mathbf{1}\{S \in V\} c_{\text{search}}(S)].$$

A cross-arm mean skipped share cannot be combined with one Stack cost ratio to infer aggregate compute savings. Historical deployment artifacts contain success and call counts; the separate timed runs below measure cost directly and are not inferred from those counts.

Image-policy timing cohort. There are 100 resets per gate (30000–30099), disjoint from pilot and test, after three excluded warm-up resets (29000–29002). Runs use one worker, CUDA synchronization, and an NVIDIA RTX PRO 6000 Blackwell Workstation Edition GPU (PyTorch 2.11.0+cu128, NumPy 1.26.4). Only one repeat is saved, in the fixed order never, combined, clock+head, cheap; drift across gate order is not estimated. Configuration and calibration match the test gates, apart from the specified clock+head ablation, which omits the gripper term.

Table 20: Image-policy timing on a separate 100-episode cohort. Decision latency is gate plus planner/policy time, excluding environment stepping. Online time includes stepping and excludes startup. Latencies are medians; calls and online seconds are episode means. No success comparison is drawn from this timing cohort.

Gate	Gate (ms)	Unsearched (ms)	Searched (s)	Calls	Online (s)
Never	0.023	2.77	n/a	0.00	0.47
Cheap ($p = 0.30$)	27.142	29.41	21.76	2.70	54.19
Combined ($p = 0.25$)	26.881	29.14	21.96	2.28	46.01
Clock+head ($p = 0.25$)	0.075	2.80	22.53	3.15	68.04

The combined gate's 1,021 decisions include 793 unsearched decisions; the corresponding counts are 1,027/757 for cheap, 978/663 for clock+head, and 1,120/1,120 for never. Removing the twelve-sample gripper cue makes unsearched decisions cheaper, but changes gate selection and search count: the clock+head run has greater total online time. It is not an equal-allocation measurement of the isolated head cost. Planner execution accounts for 98.5–99.4% of online time in the three gated conditions. Thus small head cost, unsearched-decision latency and total online cost are distinct quantities. The snapshot contains no always-search timing run; no measured reduction relative to always-search is claimed.

D.10 Reproducibility status

Table 21 distinguishes quantities regenerated by the results pipeline from measurements or checkpoints that require additional care.

Summary reproduction does not establish end-to-end reproducibility: training metadata and the extra-pool recipe remain incomplete, and rollout replay requires the base image-policy checkpoint.

Table 21: Reproducibility status of reported artifacts and measurements.

Item	Status	Note
Ranking, veto, historical summaries	automatic	regenerated by legacy results script, which still reads historical random-control paths
New deployment results	bundled audit	completed outcomes, timing and corrected controls checked separately
Score/label joins	automatic	verified by artifact fingerprint
Compute accounting	scripted	re-measured separately; not in main results pipeline
Tie diagnostic	saved-array audit	separate stratified evaluator redraw; not label-time evaluations
Image-policy distilled head	scripted	retraining at seed 100 reproduces the deployed head bit-identically
Image-policy frontier, head seeds	scripted	one script each; checks reset IDs, calibration columns and head identity
Larger Stack distilled head	partial	checkpoint uses an extra state pool not fully rebuildable
Stack mean-head control	bundled audit	same reset IDs, calibration and planner as the criticality gate
Incremental, depth, degeneracy, exact-mass analyses	scripted	offline re-analyses of saved arrays; one script and one JSON each

The audit records input hashes, keeps historical, corrected-random and held-out artifacts apart, and bundles outcome arrays, timing records and the bootstrap script.