BEYOND ACTION DIVERSITY: EFFECT-CLASS EXPLORATION FOR LLM AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning can improve language model agents on interactive tasks, but limited training budgets are often spent continuing rollouts from different actions that produce the same immediate environmental outcome. We aim to convert repeated environmental outcomes into more effective exploration without treating distinct language histories as interchangeable. We introduce Effect-Class Exploration (ECE), which probes candidate actions in deterministic environments with restorable states, groups identical complete one-step outcomes, and continues one uniformly sampled representative per group. Each representative keeps its own history and receives a weight equal to its group size, preserving the expected additive policy gradient; our fixed-policy analysis identifies when interaction savings outweigh the additional sampling variance. Under matched interaction budgets, a fixed node bank, and a shared training update, ECE improves Qwen2.5-7B-Instruct success rates over development-selected direct sampling by 3.2 percentage points on unseen ALFWorld tasks and 5.1 points on WebShop, and exceeds optimized parsed-action grouping by 1.9 percentage points in WebShop success rate. ECE allocates exploration by observed effects under explicit efficiency conditions, with ordinary single-trajectory evaluation.

1 INTRODUCTION

Large language model (LLM) agents support interactive problem solving in information seeking, web interaction, and embodied task execution (Yao et al., 2023). Reinforcement learning develops these capabilities through environmental feedback, including learning to interleave reasoning and search (Jin et al., 2025). Yet exploratory trajectories consume model generations and environment interactions before delayed rewards reveal their utility, making effective exploration under a fixed budget a central challenge.

Recent work makes exploration more selective and affordable. ARPO branches at tool-use steps with increased generation uncertainty (Dong et al., 2026). T^{2}PO uses uncertainty dynamics to guide reasoning and resample unproductive turns (Wang et al., 2026). Tree-GRPO shares trajectory prefixes for more rollouts within fixed token or tool-call budgets (Ji et al., 2026). These methods use branching signals and prefix reuse; distinguishing actions by environmental outcomes remains an allocation problem. Different language actions, including different parsed commands, can produce the same immediate outcome: a larger rollout tree need not offer correspondingly greater environmental diversity (Figure 1, left).

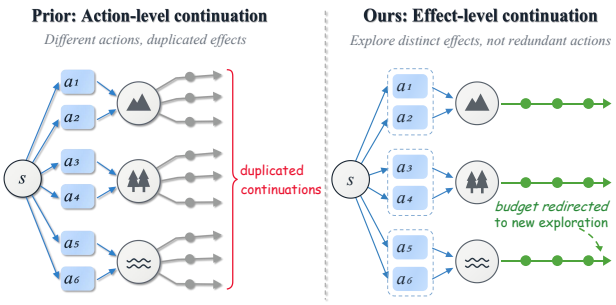


Figure 1: **Action diversity $\neq$ effect diversity. Left:** Distinct actions can repeat the same one-step effect. **Right:** ECE continues one uniformly sampled representative per effect class, saving budget for further exploration. Each representative retains its own history; equal effects need not imply identical futures.

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.

Environmental consequences already inform exploration and learning. Research on action redundancy distinguishes action entropy from transition entropy and develops objectives for diverse transitions (Baram et al., 2021). Causal effect estimation reduces redundant exploration by restricting enumerable action spaces using estimated transition effects (Liu et al., 2025). For language-based app agents, Action Semantics Learning improves generalization through supervision from interface transitions (Tang et al., 2026). GiGPO uses repeated environment states retrospectively for finer credit assignment (Feng et al., 2025). These approaches leave open how to allocate future rollout cost among sampled language actions with overlapping immediate outcomes.

The difficulty is that *environmental redundancy does not imply redundant learning signals*. LLM decisions depend on language history, so actions with the same immediate outcome can yield different subsequent behavior and returns. Completing every alternative may repeatedly pay for the same environmental branch; suppressing alternatives can discard useful training variation. Efficient allocation must balance repeated environmental outcomes against learning signals from distinct language histories.

We introduce *Effect-Class Exploration* (ECE) to allocate rollout continuations by environmental effect (Figure 1, right). In deterministic environments with restorable states, ECE probes each candidate action for one step and continues one uniformly sampled representative per distinct complete one-step outcome. Representatives retain their original language histories and receive class-size weights, preserving the expected additive policy gradient under our sampling protocol. Our fixed-policy analysis identifies when interaction savings outweigh added estimation variance, improving statistical efficiency under a fixed budget.

Our contributions are threefold:

- **An empirical diagnosis of exploration redundancy.** We document a *branching illusion* in ALFWorld and WebShop: distinct parsed actions frequently produce identical one-step effects, and wider candidate pools expose this redundancy more often.
- **A method for budgeted rollout allocation.** ECE combines unbiased additive gradient estimation with a fixed-policy criterion for improved asymptotic budget efficiency against the best fixed-width direct sampler. It retains ordinary single-trajectory execution at evaluation.
- **Evidence of gains under matched training budgets.** With a shared frozen node bank, common updates, and matched interaction budgets, ECE exceeds development-selected direct sampling in all twelve comparisons across three backbones. Online task-start validation also shows gains over direct sampling and parser grouping; further evaluations cover three host algorithms and seven search QA datasets.

2 Related Work

Budgeted rollout exploration. Tree-GRPO (Ji et al., 2026) reuses trajectory prefixes to reduce sampling cost. Adaptive methods further target exploration: ARPO branches after tool-induced entropy changes (Dong et al., 2026), T^2PO controls token generation and turn resampling using uncertainty (Wang et al., 2026), and SPARK learns to signal branching points (Wu et al., 2026). DPEPO additionally rewards action and transition diversity across parallel environments (Zhang et al., 2026a). ECE adds consolidation by exact equality of complete environmental effects, using one-step probes whose costs enter the rollout budget.

Action redundancy and semantics. Transition-entropy regularization addresses the mismatch between action diversity and environmental outcomes (Baram et al., 2021); NPM and CEE learn transition-based action filters (Zhong et al., 2024; Liu et al., 2025). ASL uses UI-transition semantics for training (Tang et al., 2026), while Harness-G tackles retrieval aliasing through a finite graph-action menu (Hou et al., 2026). These approaches change learning objectives, action support, or interfaces. ECE preserves the language-action interface and allocates continuations within a sampled pool using observed effects. It requires deterministic, restorable environments with complete effect records, without action-space enumeration or a learned effect model.

Credit assignment and history. GiGPO refines episode-relative credit through anchor-state groups (Feng et al., 2025); HGPO accounts for differing histories (He et al., 2026). RLVMR supplies

verifiable meta-reasoning rewards (Zhang et al., 2026b). ECE complements credit mechanisms by selecting candidates for continuations. It retains each representative's original history and uses multiplicity weighting to preserve the expected additive policy gradient. The guarantee allows different futures within a class and increased variance; its scope is the additive objective.

3 Preliminary Study

We use frozen Qwen2.5-3B-Instruct to examine whether more language actions yield distinct environmental branches, across 260 decision nodes per environment in ALFWorld and WebShop and six nested candidate widths ($M = 4$–32).

Parser retention (R_{parse}) measures the fraction of raw-action diversity surviving parsing; effect retention (R_{effect}) measures the fraction of parsed-action diversity surviving execution. We relate additional effect collapse among parser-distinct actions, D_{extra}, normalized by pool size, to mean remaining depth $\bar{L}$ and associated continuation mass $\mathcal{A}_{\text{extra}}$ (steps per candidate). Appendix B.1 details the protocol and metrics.

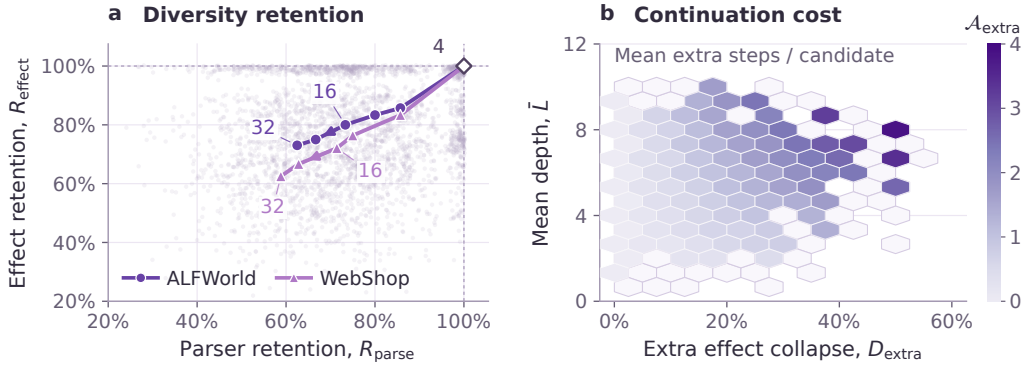


Figure 2: **Diversity compression and continuation cost. (a)** Observations and median retention curves across widths. **(b)** Mean extra continuation mass across effect-collapse and depth bins.

Figure 2(a) shows decreasing retention as candidate pools widen. At $M = 32$, median $(R_{\text{parse}}, R_{\text{effect}})$ is $(0.625, 0.730)$ on ALFWorld and $(0.588, 0.625)$ on WebShop; more than 96% of nodes in each environment contain parser-distinct actions with identical one-step effects.

Figure 2(b) associates repeated effects at greater remaining depth with larger continuation mass. Among observations with $D_{\text{extra}} > 0$, median $\mathcal{A}_{\text{extra}}/D_{\text{extra}}$ approximately doubles from the shallowest to the deepest depth quartile in both environments (Appendix C.1.2).

Branching Illusion

LLMs produce many surface-distinct and parser-distinct actions, yet distinct one-step outcomes grow more slowly. Repeated effects can carry substantial continuation cost when considerable interaction remains.

Equal one-step effects can still lead to different futures through distinct language histories, motivating effect-aware allocation that preserves histories and accounts for selection.

4 Method

Motivated by the observed effect redundancy, Effect-Class Exploration (ECE) allocates continuations by environmental outcome. It probes each candidate for one step, groups identical effects, and continues one uniformly sampled representative per class (Figure 3). Representatives retain their language histories and receive class-size weights, preserving the expected additive policy gradient. Same-class histories can have different future values, so subsampling may increase gradient vari-

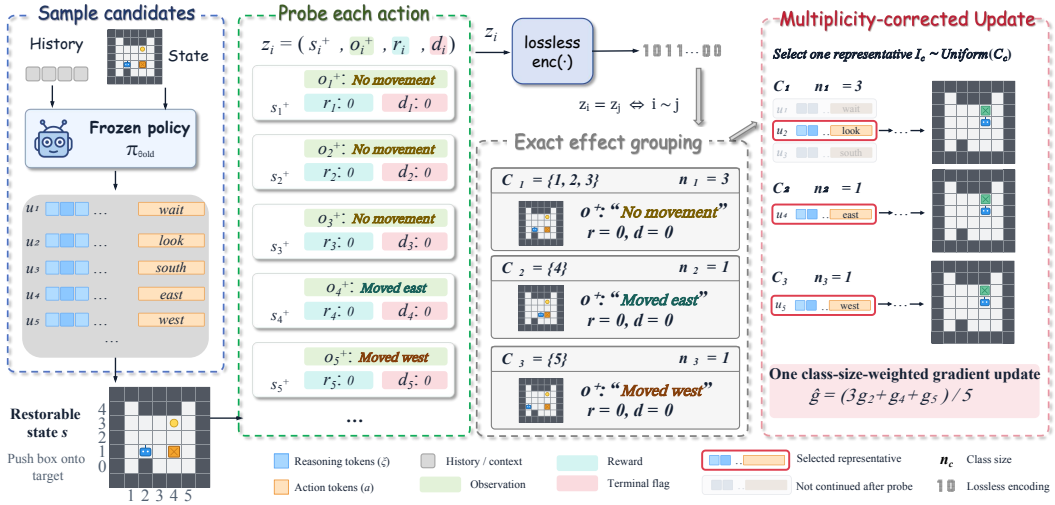


Figure 3: **Effect-Class Exploration in a restorable grid world.** A frozen policy samples $M = 5$ turns and probes each action once from the same restored state. Lossless encoding of complete records $z_i = (s_i^+, o_i^+, r_i, d_i)$ gives $K = 3$ exact-effect classes: `wait`, `look`, and the blocked `south` action share one record; `east` and `west` each form a singleton. One uniformly sampled representative per class resumes from its saved post-probe state with its own history, without repeating the probe. Selecting `look`, `east`, and `west` yields $\widehat{g} = (3g_2 + g_4 + g_5)/5$, where g_i is the full-trajectory policy-gradient contribution. Class-size weighting preserves the expected additive gradient; continuation outcomes within a class may still differ.

ance. ECE complements branching, resampling, and trajectory-tree reuse (Dong et al., 2026; Wang et al., 2026; Ji et al., 2026).

4.1 Learning Objective

We study deterministic, restorable environments with finite horizon H. A node $v = (s, h)$ pairs complete state s with policy-visible history h. The controlled study fixes a distribution μ over a task-balanced node bank collected on training tasks by a frozen reference policy before optimization; the bank also remains fixed (Appendix A.2.1). A rollout τ stops at termination or H interactions, including the initial action, with length $T(\tau) \le H$. We optimize

$$J_{\mu,H}(\theta) = \mathbb{E}_{v\sim\mu}\,\mathbb{E}_{\tau\sim p_\theta(\cdot|v)}\left[G(\tau)\right], \qquad G(\tau) = \sum_{\ell=1}^{T(\tau)} r_\ell. \tag{1}$$

Here p_θ is the rollout distribution induced by the policy and environment, and r_ℓ is the original task reward. The undiscounted return counts the probe reward once, without horizon bootstrapping. With prefixes h fixed, $J_{\mu,H}$ measures continuation value under μ. Appendix A.3 extends collection to current-policy task-start prefixes, including their token likelihoods in the gradient.

4.2 Candidate Sampling and Exact Effect Classes

Each batch freezes the current parameters as θ_{old} and samples M independent candidate turns per node:

$$u_i = (\xi_i, a_i) \overset{\text{iid}}{\sim} \pi_{\theta_{\text{old}}}(\cdot \mid h), \qquad i \in [M], \quad [M] = \{1, \dots, M\}. \tag{2}$$

Here ξ_i is generated reasoning and a_i raw action text. Candidates and continuations use this frozen policy at temperature 1, without top-k or top-p truncation, under fixed token and interaction limits.

For each probe, restore s independently and execute a_i once through deterministic $\mathcal{E}$, including parsing and invalid-action feedback:

$$z_i = \mathcal{E}(s, a_i) = \left(s_i^+, o_i^+, r_i, d_i\right), \qquad k_i = \text{enc}(z_i). \tag{3}$$

The record contains the complete next state s_i^+, full policy-visible response o_i^+, immediate reward r_i, and terminal flag d_i. The state includes all simulator and wrapper variables affecting future transitions, responses, rewards, or stopping. Deterministic, lossless, type-preserving encodings determine equality; hashes only index records. Appendix A.1 specifies the state interface, serialization, and restoration checks.

ECE partitions *candidate indices* by

$$i \sim j \quad \Longleftrightarrow \quad k_i = k_j, \qquad [M] = \bigsqcup_{c=1}^{K} \mathcal{C}_c. \tag{4}$$

Here $n_c = |\mathcal{C}_c|$, $c(i)$ identifies the class containing i, and $\rho = (M-K)/M$ is the observed duplicate-effect fraction. Repeated samples retain separate indices and their multiplicities. The post-probe history $h_i^+ = h \oplus u_i \oplus o_i^+$ appends the original turn and response in execution order. Equal effects do not imply equal language histories.

4.3 Representative Continuations

Let $\mathcal{F}$ contain the current node, all sampled turns, and their observed effects. Independently across classes, ECE samples one index uniformly:

$$I_c \sim \mathrm{Unif}(\mathcal{C}_c), \qquad q_i(\mathcal{F}) \triangleq \Pr(i \in S \mid \mathcal{F}) = \frac{1}{n_{c(i)}}, \qquad S = \{I_1, \ldots, I_K\}. \tag{5}$$

Selection uses no continuation-derived scores and is independent of continuation randomness. If $d_{I_c} = 1$, the recorded one-step trajectory is complete; otherwise, ECE resumes from saved state $s_{I_c}^+$ with the representative's own history $h_{I_c}^+$ until termination or $H-1$ further interactions (none if $H = 1$). The probe is neither repeated nor counted twice. Grouping occurs only at the sampled node; later turns use ordinary policy interaction.

4.4 Multiplicity-Corrected Policy Gradient

Let $\mathcal{I}_i$ index all policy-generated tokens y_{ik} with contexts c_{ik} in trajectory τ_i, including the initial candidate and continuation. Its autoregressive log likelihood is

$$\ell_\theta(\tau_i \mid v) \triangleq \log p_\theta(\tau_i \mid v) = \sum_{k \in \mathcal{I}_i} \log \pi_\theta(y_{ik} \mid c_{ik}). \tag{6}$$

Fixed prefixes and environment responses enter the context, but not token log probabilities. Sampling and likelihood evaluation share stopping and context rules (Appendix A.2.2).

Weighting by inverse inclusion probability n_c preserves empirical class mass (Horvitz & Thompson, 1952), giving

$$\widehat{g}_{\mathrm{ECE}}(v) = \frac{1}{M} \sum_{c=1}^{K} n_c \, G(\tau_{I_c}) \, \nabla_\theta \ell_\theta(\tau_{I_c} \mid v)|_{\theta=\theta_{\mathrm{old}}} . \tag{7}$$

Normalization uses the original pool size M, not K; returns and class sizes are held constant during differentiation.

If fully continued with its own randomness, candidate i would contribute $g_i = G(\tau_i)\nabla_\theta \ell_\theta(\tau_i \mid v)|_{\theta=\theta_{\mathrm{old}}}$. Uniform selection gives

$$\mathbb{E}[\widehat{g}_{\mathrm{ECE}}(v) \mid \mathcal{F}] = \frac{1}{M} \sum_{i=1}^{M} \mathbb{E}[g_i \mid \mathcal{F}], \qquad \mathbb{E}[\widehat{g}_{\mathrm{ECE}}(v)] = \nabla_\theta J_{\mu,H}(\theta)|_{\theta=\theta_{\mathrm{old}}} . \tag{8}$$

The outer expectation includes node sampling from μ and candidate sampling (proof and regularity conditions in Appendix D.1.1). Same-class histories can induce different continuation distributions and returns under the frozen policy. The identity holds for any partition formed before continuation sampling, regardless of the grouping criterion; representative subsampling can increase variance at fixed candidate width (Appendix D.1.2).

We average $\widehat{g}_{\mathrm{ECE}}(v)$ over N independently sampled nodes, take one gradient-ascent step at θ_{old}, and discard the batch. We use zero baseline, no weight clipping or selected-group return normalization. The unbiasedness guarantee in Eq. (8) applies to this additive score-function update (Sutton et al., 1999) (Appendix A.2.3).

4.5 Interaction Cost and Resource Allocation

Both collectors execute M initial actions; direct sampling continues all M candidates, ECE only K representatives. With L further interactions per candidate, their costs are

$$C_{\text{direct}} = M + ML, \qquad C_{\text{ECE}} = M + KL. \tag{9}$$

ECE saves $(M - K)L$ interactions. Appendices A.2.4–A.2.5 detail variable-length costs, resource accounting, complete-batch reservations, and evaluation.

Statistical efficiency. For the fixed-bank collector at a fixed policy, efficiency combines interaction cost and gradient variance. Let T, A, W denote variability across nodes, between effects, and within effects, respectively, and Q the collision-weighted within-effect variability (Appendix D.2.1).

Theorem 1 (When ECE improves budget efficiency)**.** *Fix M under the iid sampling assumptions in Appendix D.2.1 and the complete-pool reservation rule. Let $c_{\mathrm{E}}(M)$ be ECE's expected pool cost, including probes, and $\bar{\ell}$ the mean ordinary continuation length. As the interaction budget $B \to \infty$, the mean squared error (MSE) $\mathcal{R}_{\mathrm{E}}^{\mathrm{hard}}(B)$ of the gradient mean over completed pools satisfies $B\mathcal{R}_{\mathrm{E}}^{\mathrm{hard}}(B) \to \mathcal{C}_{\mathrm{E}} := c_{\mathrm{E}}(M)[T + (A + W)/M + (M - 1)Q/M]$. The best leading constant over fixed-width direct samplers is $\mathcal{C}_{\mathrm{D}}^{\star} = (1 + \bar{\ell})(T + A + W)$, attained at width one. ECE improves this constant exactly when $\mathcal{C}_{\mathrm{E}} < \mathcal{C}_{\mathrm{D}}^{\star}$; this implies smaller MSE for sufficiently large budgets against each fixed direct width.*

Proof sketch. Conditioning on effect counts gives the exact ECE pool variance. The reservation rule and uniform integrability yield the hard-budget limit of expected cost times variance. For direct sampling, this product is $(1 + \bar{\ell})(wT + A + W)$, minimized at $w = 1$. Appendix D.6 proves the criterion; Appendix D provides supporting results and comparisons. □

5 Experiments

5.1 Experimental Setup

Benchmarks and metrics. Our primary benchmarks are ALFWorld (Shridhar et al., 2021) and WebShop (Yao et al., 2022); ScienceWorld (Wang et al., 2022) and the seven-dataset Search-R1 QA suite (Jin et al., 2025) test generality. Metrics include task-start success rates, WebShop scores, ScienceWorld final progress, per-dataset and macro QA exact match, and macro token-level F1. Each evaluation follows one trajectory without branching or state restoration.

Comparison protocol. Allocation controls fix the node bank, objective, and update rule, comparing ECE with width-one, matched-width, and development-selected direct sampling, parsed-action grouping, and size-matched random partitions. Host comparisons preserve the native objectives and optimizers of GRPO, GiGPO (std) (Feng et al., 2025), and HGPO (He et al., 2026). Search QA also compares complete algorithms: Search-R1, Tree-GRPO (Ji et al., 2026), ARPO's local-search adaptation (Dong et al., 2026), and T^{2}PO (Wang et al., 2026). Methods share task splits and evaluation interfaces; matched interaction or token budgets include candidate generation, probes, and setup.

Backbones and implementation. Qwen2.5-7B-Instruct is the main backbone, with Qwen2.5-3B-Instruct and Qwen3-4B as additional backbones. Common-PG ECE uses $M = 8$ candidates, $N = 16$ nodes, temperature 1, and one multiplicity-weighted update per fresh frozen-policy batch on eight H100-80GB GPUs. Tables 1–3 report point estimates. Appendix C.3 contains four primary 7B statistical comparisons and common-pool mechanism measurements. Splits, decoding, optimization, and statistics are detailed in Appendix B.

5.2 Main Results

Table 1 compares collectors under a shared node bank, PG update, and interaction cap per backbone. Direct $w = 8$ matches ECE's width; $w = 1$ is asymptotically budget-efficient under Theorem 1's fixed-policy assumptions. ECE outperforms both fixed widths and development-selected sampling in all twelve backbone–metric comparisons, gaining 0.9–3.3 points over the parser and 2.7–7.4 over random grouping. The parser consistently outperforms random grouping, supporting the value of

	Qwen2.5-3B-Instruct				Qwen2.5-7B-Instruct				Qwen3-4B			
	ALFWorld SR		WebShop		ALFWorld SR		WebShop		ALFWorld SR		WebShop	
Rollout allocation	Seen	Unseen	Score	SR	Seen	Unseen	Score	SR	Seen	Unseen	Score	SR
Direct sampling ($w = 1$)	84.5	79.5	82.1	68.3	89.2	85.5	85.8	73.7	88.2	83.9	86.3	74.5
Direct sampling ($w = 8$)	79.0	77.0	80.1	65.6	85.6	83.5	85.3	72.4	85.7	84.0	83.8	71.0
Direct sampling (w_{dev})	83.0	78.9	81.4	68.3	90.1	88.4	87.9	75.5	90.7	88.2	86.1	74.7
Parsed-action grouping (optimized)	84.9	81.8	82.7	70.2	91.9	90.6	87.7	78.7	90.1	87.6	87.3	76.1
Size-matched random grouping	83.2	79.7	81.0	66.8	90.1	87.0	85.4	73.2	88.2	85.8	85.5	72.9
Effect-class exploration (ECE)	**88.2** (+5.2)	**85.1** (+6.1)	**84.7** (+3.3)	**71.7** (+3.4)	**92.8** (+2.7)	**91.6** (+3.2)	**89.7** (+1.8)	**80.6** (+5.1)	**91.3** (+0.6)	**89.6** (+1.4)	**89.3** (+3.2)	**77.7** (+3.0)

Table 1: **Effect-aware allocation across backbones.** Rows share a node bank, additive policy-gradient (PG) update, and interaction cap $S_{\text{int}} + 10^6$ per backbone. Scores are percentages (higher is better); SR denotes success rate. ECE increments $\Delta_{b,d} = S^{\text{ECE}}_{b,d} - S^{\text{Direct}(w_{\text{dev}})}_{b,d}$ are percentage-point gains over development-selected direct sampling, using unrounded ALFWorld means and displayed WebShop scores. Bold/underlined values mark the best/second best per backbone and metric. See Appendix B.4 for the protocol.

meaningful grouping and additional effect information. Random grouping matches class sizes but can incur different continuation costs. The four primary Holm-corrected 7B tests support ECE over tuned direct sampling in both environments and over the parser on WebShop; the ALFWorld parser-relative interval includes zero, leaving that comparison inconclusive (Table 9, Appendix C.3). These controls fix continuation states and evaluate policies from task starts. Online validation uses current-policy prefixes: ECE gains 2.9/5.2 points over direct sampling and 1.3/2.2 over the parser on ALFWorld unseen/WebShop SR (Appendix C.2).

5.3 Generality across Training Algorithms and Tasks

		ALFWorld		WebShop		ScienceWorld	
Host	Rollout allocation	Seen SR	Unseen SR	Score	SR	Score	SR
GRPO (Shao et al., 2024)	Native chain collector	79.0	76.6	81.0	67.2	64.3	46.4
	+ Parsed-action grouping	81.2	78.4	82.7	69.6	64.8	47.1
	+ ARPO-EAR (Dong et al., 2026)	82.6	79.6	83.6	71.4	66.1	49.3
	+ T^{2}PO-TDS (Wang et al., 2026)	83.8	79.9	83.9	70.9	65.9	48.2
	+ Tree-GRPO sampler (Ji et al., 2026)	82.1	79.1	82.9	70.6	64.7	48.0
	+ ECE (ours)	**84.5** (+0.7)	**80.8** (+0.9)	**84.7** (+0.8)	**72.1** (+0.7)	**66.6** (+0.5)	**50.2** (+0.9)
GiGPO (std) (Feng et al., 2025)	Native chain collector	91.9	89.1	85.8	74.2	71.2	53.6
	+ Parsed-action grouping	92.4	90.3	87.1	76.6	72.0	54.4
	+ ARPO-EAR (Dong et al., 2026)	93.1	91.5	87.8	77.3	72.9	55.8
	+ T^{2}PO-TDS (Wang et al., 2026)	92.9	90.8	88.5	77.7	71.8	54.9
	+ Tree-GRPO sampler (Ji et al., 2026)	93.8	91.0	87.4	77.0	73.8	56.7
	+ ECE (ours)	**94.3** (+0.5)	**92.3** (+0.8)	**89.1** (+0.6)	**78.6** (+0.9)	**74.3** (+0.5)	**57.5** (+0.8)
HGPO (He et al., 2026)	Native chain collector	95.0	91.8	88.3	78.6	75.9	59.5
	+ Parsed-action grouping	94.3	92.0	89.0	79.4	76.1	59.3
	+ ARPO-EAR (Dong et al., 2026)	94.5	92.8	89.5	80.3	76.5	60.1
	+ T^{2}PO-TDS (Wang et al., 2026)	95.2	92.3	90.2	80.6	75.7	59.6
	+ Tree-GRPO sampler (Ji et al., 2026)	95.0	93.3	89.4	79.8	77.2	61.2
	+ ECE (ours)	**95.5** (+0.3)	**93.8** (+0.5)	**90.8** (+0.6)	**81.4** (+0.8)	**77.6** (+0.4)	**61.9** (+0.7)

Table 2: **Exploration choices within host training algorithms.** Qwen2.5-7B-Instruct uses cap $S_{\text{int}} + 1.0 \times 10^6$, native host objectives and optimizers, and shared evaluation. ARPO-EAR and T^{2}PO-TDS adapt entropy-based adaptive rollout and turn-level dynamical sampling, respectively. Parentheses give ECE's gain over the strongest alternative per host and metric. Scores use a 0–100 scale (higher is better); bold/underlined values are best/second best per host and metric. ScienceWorld uses final progress clipped below at zero and goal completion within 30 attempted actions for SR. Protocols appear in Appendices A.4 and B.5.

Native Host Algorithms. Table 2 compares six collectors within three hosts. ECE applies each host's advantage and loss operators to multiplicity-expanded batches (Appendix A.4). It has the highest point estimates in all 18 host–metric comparisons, gaining 0.3–0.9 points over the strongest

alternative, 0.5–5.5 over native chain collection, and 1.2–3.3 over the parser. Gains over native collection are larger for GRPO (2.3–5.5 points) and GiGPO (2.4–4.4) than for HGPO (0.5–2.8), whose native collector already achieves higher scores. The smaller gains over the strongest alternative indicate benefits from other exploration methods.

A. Complete-algorithm comparison

	Macro		Single-hop EM			Multi-hop EM			
Training system	EM	F1	NQ	TriviaQA	PopQA	HotpotQA	2Wiki	MuSiQue	Bamboogle
ReAct (Yao et al., 2023)	27.7	38.6	26.8	50.2	32.4	24.1	25.6	8.9	26.2
Search-R1 (PPO) (Jin et al., 2025)	38.6	49.2	38.2	61.3	40.1	37.6	41.8	13.9	37.3
GRPO (Shao et al., 2024)	41.8	53.4	45.2	60.9	44.2	39.7	40.2	19.0	43.5
GiGPO (Feng et al., 2025)	46.8	58.9	45.8	65.1	45.9	40.7	44.8	18.4	67.1
Tree-GRPO (Ji et al., 2026)	45.0	56.4	48.2	63.5	44.9	43.4	42.5	20.7	51.8
ARPO-local (Dong et al., 2026)	47.7	58.4	46.6	65.0	47.9	44.1	**47.3**	23.4	59.4
T^2PO (QA adapt.) (Wang et al., 2026)	46.6	58.2	47.4	64.2	44.8	40.3	44.2	20.9	64.6
Common PG + ECE (ours)	**49.5**	**60.7**	**49.0**	**65.4**	**48.5**	**45.0**	47.1	**24.1**	**67.4**

B. Allocation-controlled comparison

	Macro		Single-hop EM			Multi-hop EM			
Rollout allocation	EM	F1	NQ	TriviaQA	PopQA	HotpotQA	2Wiki	MuSiQue	Bamboogle
Direct sampling ($w = 1$)	42.8	54.4	44.3	62.5	43.9	38.9	41.9	16.4	51.7
Direct sampling ($w = 8$)	42.1	54.9	43.4	62.1	43.6	39.3	39.5	16.8	50.2
Direct sampling (w_{dev})	44.7	56.3	46.1	62.8	46.1	40.2	42.5	18.4	57.0
Parsed-action grouping (optimized)	47.1	58.4	47.9	**65.7**	46.3	42.4	45.6	20.0	61.7
Size-matched random grouping	44.6	55.3	46.4	62.9	45.6	39.4	42.0	18.6	57.5
Effect-class exploration (ECE)	**49.5**	**60.7**	**49.0**	65.4	**48.5**	**45.0**	**47.1**	**24.1**	**67.4**

Table 3: **Fixed-retriever search QA.** Each trained system uses one Qwen2.5-7B-Instruct policy on the common Natural Questions–HotpotQA mixture, with setup allowance $S_{\text{int}}^{\text{QA}} = 169{,}000$ and total cap $S_{\text{int}}^{\text{QA}} + 10^6 = 1{,}169{,}000$; actual setup costs are charged per method. Panel A compares complete algorithms with native objectives; Panel B fixes the PG update and varies allocation. ECE is repeated across panels. Dataset columns report exact match (EM); macro EM and token-level F1 are unweighted means over seven datasets. Scores use a 0–100 scale; bold/underlined values are best/second best within each panel and column. Retrieval, search limits, QA adaptations, and accounting are detailed in Appendix B.2.

Fixed-Retriever Search QA. Under the common PG update (Panel B of Table 3), ECE achieves the highest macro EM (49.5) and F1 (60.7), up from 44.7/56.3 for development-selected direct sampling and 47.1/58.4 for the parser. It exceeds direct sampling and size-matched random grouping on all seven dataset EM scores, and the parser on six; TriviaQA is the exception (65.4 versus 65.7). Parser-relative EM gains reach 4.1 points on MuSiQue and 5.7 on Bamboogle, with improvements on all four multi-hop datasets. These matched-update comparisons show benefits from parsed-action grouping and additional gains from environmental effects. ECE also has the highest macro EM and F1 in Panel A, which compares complete training systems with their native objectives. It leads six dataset EM scores in this panel, with ARPO-local ahead on 2Wiki (47.3 versus 47.1).

5.4 Budget Efficiency

In Figure 4, ECE leads eleven of twelve environment–budget conditions, including all six interaction caps. Interaction-budget gains over development-selected direct sampling are 0.4–3.3 points on ALFWorld and 5.1–6.8 on WebShop; ECE also exceeds the parser at every interaction cap. On WebShop, ECE reaches 73.6% at $f = 1/2$ under the interaction cap, close to tuned direct sampling's 75.5% at $f = 1$. Under token caps, parser-relative gains peak at $f = 1/2$ (3.3 and 3.9 points, respectively). At $f = 1/4$ on ALFWorld, ECE's 69.1% trails tuned direct sampling at 72.4% and the parser at 72.1%. ECE returns to the lead on ALFWorld at both larger token caps, while retaining its lead on WebShop throughout. The changing rankings at the smallest cap show that efficiency depends on which resource limits training, even with the same allocation rule.

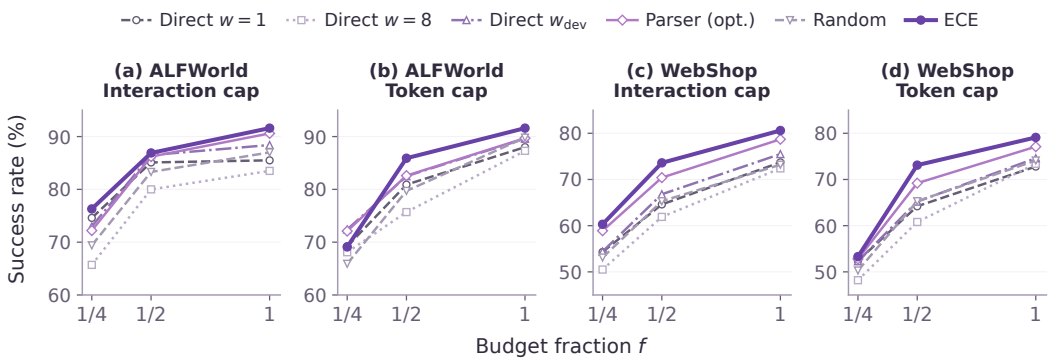


Figure 4: **Efficiency under interaction and token budgets.** Qwen2.5-7B-Instruct on ALFWorld unseen and WebShop under separate caps of $S_{\text{int}} + f\,10^6$ interactions or $S_{\text{tok}} + f\,(2 \times 10^9)$ input/output tokens. Direct width is selected at $f = 1$ on development tasks per environment and resource type, then fixed across all three caps. Markers are measured scores; lines are visual guides. See Appendix C.6 for exact scores.

5.5 Mechanism Diagnostics

Figure 5 compares expected continuation interactions saved, B_c, with the added variance penalty, R_c, for 476 repeated effect classes (224 from ALFWorld and 252 from WebShop; $n_c \in \{2, \ldots, 6\}$). The penalty R_c is the conditional variance trace introduced by sampling one class representative in a fixed projected-gradient space, divided by the node's full-direct gradient energy plus ϵ (Appendix B.11). Actions with the same immediate effect can still yield different gradient contributions through their distinct language histories. Points toward the lower right combine larger interaction savings with smaller normalized variance penalties. Parse-error and no-state-change classes account for 58.8% of records and 70.1% of summed expected savings; their median R_c values are 0.0109 and 0.0151, respectively. Task-progress classes also yield savings (median $B_c = 9.27$), with higher penalties (median $R_c = 0.0271$; 90th percentile $R_c = 0.1453$). Across all classes, the median penalty is 0.0183 and the 95th percentile is 0.1127. These descriptive results show that savings extend across effect types, with differing precision costs. Complementary effect-coverage and gradient-fidelity diagnostics appear in Appendix C.4; Appendix C.5 assesses total cost–variance.

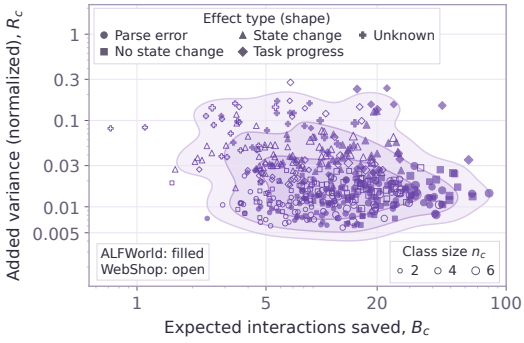


Figure 5: **Interactions saved versus added gradient variance.** Each point is a repeated effect class. Shape, fill, and area encode effect type, environment, and class size; purple contours show class density.

6 Conclusion

We studied how to allocate a limited rollout budget when distinct language actions lead to the same immediate environmental effects. Effect-Class Exploration turns this redundancy into a sampling decision, distributing continuation effort across effect classes while retaining each selected action's history and correcting its contribution to the additive learning update. Fixed-bank experiments on interactive tasks and search QA show gains over development-selected direct sampling, while the fixed-policy analysis explains how interaction cost and gradient variability jointly determine statistical efficiency. Together with online task-start validation, these findings support assessing exploration through environmental coverage and learning efficiency under matched resource budgets. An open question is how to adapt representative counts within effect classes as learning progresses, guided by the cost–variance relationship established here.

AI USE STATEMENT

We used the LLM-based assistant OpenAI Codex to assist with checking selected proofs. We also used it for literature search and synthesis, editing manuscript text for English clarity and concision, organizing sections and references, and preparing LaTeX sources and formatting equations, tables, and figure captions. The authors take responsibility for all content of this work, including its research decisions, theoretical claims, experimental results, citations, and artifacts prepared with AI assistance.

REFERENCES

F. J. Anscombe. Large-sample theory of sequential estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, 48(4):600–607, 1952. doi: 10.1017/S0305004100076386. URL https://doi.org/10.1017/S0305004100076386.

Nir Baram, Guy Tennenholtz, and Shie Mannor. Action redundancy in reinforcement learning. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pp. 376–385. PMLR, 2021. URL https://proceedings.mlr.press/v161/baram21a.html.

Guanting Dong, Hangyu Mao, Kai Ma, Licheng Bao, Yifei Chen, Zhongyuan Wang, Zhongxia Chen, Jiazhen Du, Huiyang Wang, Fuzheng Zhang, Guorui Zhou, Yutao Zhu, Ji-Rong Wen, and Zhicheng Dou. Agentic reinforced policy optimization. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/1c58e53bdf1fb045440256fd567531ae-Abstract-Conference.html.

Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-group policy optimization for LLM agent training. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/420c9f777c0b4f78d515e53cf74d58b2-Abstract-Conference.html.

Shuo He, Lang Feng, Qi Wei, Xin Cheng, Lei Feng, and Bo An. Hierarchy-of-groups policy optimization for long-horizon agentic tasks. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/2e7eb932f93d5db3b61c7a001e72fd0f-Abstract-Conference.html.

D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952. doi: 10.1080/01621459.1952.10483446. URL https://doi.org/10.1080/01621459.1952.10483446.

Yanning Hou, Haoyuan Chen, Sihang Zhou, Xiaoshu Chen, Xirui Liu, Duanyang Yuan, Lingyuan Meng, Siwei Wang, Quan Liu, and Jian Huang. Harness-G: A graph-structured harness for search agents. *arXiv preprint arXiv:2607.27652*, 2026. URL https://arxiv.org/abs/2607.27652.

Yuxiang Ji, Ziyu Ma, Yong Wang, Guanhua Chen, Xiangxiang Chu, and Liaoni Wu. Tree search for LLM agent reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/8c7304e77c832ddc70075dfee081ca6c-Abstract-Conference.html.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Ö. Arık, Dong Wang, Hamed Zamani, and Jiawei Han. Search-R1: Training LLMs to reason and leverage search engines with reinforcement learning. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=Rwhi91ideu.

Wenzhang Liu, Lianjun Jin, Lu Ren, Chaoxu Mu, and Changyin Sun. Reducing action space for deep reinforcement learning via causal effect estimation. *arXiv preprint arXiv:2501.14543*, 2025. URL https://arxiv.org/abs/2501.14543.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. URL https://arxiv.org/abs/1707.06347.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. ALFWorld: Aligning text and embodied environments for interactive learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=0IOX0YcCdTn.

Richard S. Sutton, David A. McAllester, Satinder P. Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, volume 12, 1999. URL https://papers.nips.cc/paper_files/paper/1999/hash/464d828b85b0bed98e80ade0a5c43b0f-Abstract.html.

Bohan Tang, Dezhao Luo, Jianheng Liu, Jingxuan Chen, Shaogang Gong, Jianye Hao, Jun Wang, and Kun Shao. Beyond syntax: Action semantics learning for app agents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 9444–9454, 2026. URL https://openaccess.thecvf.com/content/CVPR2026F/html/Tang_Beyond_Syntax_Action_Semantics_Learning_for_App_Agents_CVPRF_2026_paper.html.

Haixin Wang, Hejie Cui, Chenwei Zhang, Xin Liu, Shuowei Jin, Shijie Geng, Xinyang Zhang, Nasser Zalmout, Zhenyu Shi, and Yizhou Sun. T^2PO: Uncertainty-guided exploration control for stable multi-turn agentic reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2026. URL https://openreview.net/forum?id=aD1zjvdJN4.

Ruoyao Wang, Peter Jansen, Marc-Alexandre Côté, and Prithviraj Ammanabrolu. ScienceWorld: Is your agent smarter than a 5th grader? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11279–11298. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.775. URL https://aclanthology.org/2022.emnlp-main.775/.

Jinyang Wu, Shuo Yang, Yuhao Shen, Shuai Zhang, Zhengqi Wen, and Jianhua Tao. SPARK: Strategic policy-aware exploration via dynamic branching for long-horizon agentic learning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 23981–24004. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1100. URL https://aclanthology.org/2026.acl-long.1100/.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. WebShop: Towards scalable real-world web interaction with grounded language agents. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://papers.nips.cc/paper_files/paper/2022/hash/82ad13ec01f9fe44c01cb91814fd7b8c-Abstract-Conference.html.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

JunShuo Zhang, Chengrui Huang, Feng Guo, Zihan Li, Ke Shi, Menghua Jiang, Jiguo Yu, Shuo Shang, and Shen Gao. DPEPO: Diverse parallel exploration policy optimization for LLM-based agents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 46367–46389. Association for Computational Linguistics, 2026a. doi: 10.18653/v1/2026.acl-long.2151. URL https://aclanthology.org/2026.acl-long.2151/.

Zijing Zhang, Ziyang Chen, Mingxiao Li, Zhaopeng Tu, and Xiaolong Li. RLVMR: Reinforcement learning with verifiable meta-reasoning rewards for robust long-horizon agents. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/baa41b7368408670c6a14e06a04420d9-Abstract-Conference.html.

Dianyu Zhong, Yiqin Yang, and Qianchuan Zhao. No prior mask: Eliminate redundant action for deep reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 17078–17086, 2024. doi: 10.1609/aaai.v38i15.29652. URL https://ojs.aaai.org/index.php/AAAI/article/view/29652.

Appendix Contents

A Technical Details **14**

A.1 Environment Interface and Exact Matching 14
A.2 Collection, Updates, and Resource Accounting 15
A.3 Online single-branch collection 17
A.4 Host-adapter conventions 18

B Experimental Details **20**

B.1 Preliminary-Study Protocol and Metrics 20
B.2 Benchmarks, splits, and evaluation interfaces 21
B.3 Task, efficiency, and mechanism metrics 22
B.4 Allocation-controlled comparisons 23
B.5 Host-adapter and complete-algorithm comparisons 24
B.6 Implementation, budget admission, and statistics 25
B.7 Budget-Efficiency Protocol 26
B.8 Statistical Testing and Rollout Accounting 26
B.9 Effect-Coverage and Gradient-Fidelity Protocol 27
B.10 Frozen-Checkpoint Cost–Variance Protocol 28
B.11 Class-Level Continuation Savings and Representative Variance 28

C Additional Experimental Results **30**

C.1 Additional Preliminary-Study Results 30
C.2 Online Validation from Task Starts 30
C.3 Statistical Tests and Rollout Efficiency 31
C.4 Effect Coverage and Gradient Fidelity 32
C.5 Cost-Normalized Gradient Variance 33
C.6 Budget-Efficiency Results 34
C.7 Environment-Adapter Verification 35

D Proofs **36**

D.1 Estimator Properties 36
D.2 Population Model, Variance, and Cost Identities 37
D.3 Expected-Budget Efficiency and Grouping Comparisons 42
D.4 Asymptotic Efficiency under a Hard Budget 47
D.5 Finite-Budget Confidence Bounds 52
D.6 Proof of the Main Budget-Efficiency Criterion 55

A TECHNICAL DETAILS

A.1 ENVIRONMENT INTERFACE AND EXACT MATCHING

A.1.1 STATE AND EXECUTION CONTRACT

An admissible environment adapter exposes state capture, independent restoration, and deterministic single-turn execution. A snapshot binds the task instance and simulator version to all mutable simulator and wrapper state used by future execution. Its required contents are specified in Table 4. Immutable task data may be stored once and referenced by an immutable identifier; the reference must resolve to exactly the same data on restore.

Component	Required contents
Task binding	Task identity, simulator configuration, static world or catalog data, and task conditions used by the reward and termination functions.
World state	All mutable entities, attributes, relations, inventories, containers, and other variables consulted by the transition function.
Session state	Active session, page or view, navigation and search state, result ordering, selected items, and options used by subsequent interface calls.
Progress and feedback	Interaction counters, remaining allowances, reward bookkeeping, terminal flags, parser and wrapper state, and error status.
Runtime dependencies	Logical clocks, random-generator state, generated identifiers, caches, and external-store contents whenever they can affect future execution.

Table 4: Snapshot requirements for exact-effect collection. Each applicable component includes its full state, including fields hidden from the policy. Equality of a screen summary, page title, inventory, or product identifier alone is insufficient.

The full response is stored separately in the effect record, together with the reward and terminal flag. Response equality includes text formatting, item order, identifiers, action echoes, and error messages. For structured responses, it includes the complete structure returned by the interface. The policy receives only its standard observations. The collector uses privileged state to restore branches and compare effects; representative selection uses equality and class membership, without ranking classes by reward magnitude.

Determinism is required conditional on the captured state. A pseudorandom generator is compatible with this contract only when its state is captured and replayed exactly. Uncontrolled clocks, remote mutations, and uncaptured randomness violate the contract. Controlling these sources requires state capture and replay in addition to response agreement.

A.1.2 SERIALIZATION AND RESTORATION CHECKS

The encoding of (s_i^+, o_i^+, r_i, d_i) preserves types and values, with explicit field boundaries. String contents and sequence order are retained. Scalar representations are lossless; floating-point values are not rounded for matching. Mapping keys can be ordered canonically only when their iteration order has no effect on the environment or the policy-visible response. No semantic normalization removes identifiers, timestamps, whitespace, failure types, or action echoes. A digest locates candidate matches, followed by equality of the complete encodings, so a digest collision cannot merge unequal records.

Restoring a snapshot must reproduce its encoded state and the outputs of every admissible subsequent action sequence. The validation procedure checks this contract with repeated single actions, short action sequences, invalid actions, and transitions reaching termination. It also executes independent branches in different orders to detect shared mutable state. Each check compares the complete state, response, reward, and terminal flag at every step. Source review of the adapter's field mapping and exclusion rules checks state completeness and complements the finite replay tests.

Invalid model actions follow the environment's fixed error transition and consume one interaction attempt, without a retry-until-valid loop. They remain eligible for grouping according to their full

observed effects. Infrastructure faults are recorded separately. A restoration mismatch halts collection before an update from the affected batch, and all costs already incurred remain in the ledger. Successful branches from an incomplete batch are not substituted for its missing trajectories.

A.2 Collection, Updates, and Resource Accounting

A.2.1 Fixed Node Distribution

A frozen reference policy generates trajectories on the training tasks using a fixed seed schedule. For each task, a fixed quota of nonterminal decision nodes is sampled uniformly without replacement from these trajectories. The quota and reference-collection limits are fixed before comparing allocation methods. Tasks that cannot supply the quota are identified before policy optimization, and the resulting task set and exclusions are shared by every method. Neither node inclusion nor its sampling weight depends on effect multiplicity or continuation return.

Each bank record contains the immutable task binding, restorable state s, and complete policy history h. The same quota per task makes the uniform distribution over bank records task-balanced. The bank and this distribution μ are frozen. For each admitted update, N nodes are drawn independently with replacement. Evaluation tasks do not enter the bank. Reference rollout generation, snapshot validation, and bank storage are recorded as setup costs and charged consistently to all methods.

A.2.2 Token Likelihoods and the Update

The rollout policy samples from the full softmax at temperature 1. Sampling and likelihood evaluation both disable dropout. The tokenizer, prompt construction, context-window rule, end-of-turn tokens, maximum generated tokens per turn, and action parser remain fixed across methods. Every selected candidate retains its own original turn and response. Any context-window truncation uses the same deterministic rule in sampling and log-probability evaluation, and the complete original history remains available in the collection record.

The token mask in Eq. (6) includes the selected candidate's reasoning and action tokens and all tokens sampled in its continuation, including sampled stopping tokens. It excludes the fixed prefix, environment outputs, padding, and deterministically inserted delimiters. If a token cap stops generation, the likelihood contains exactly the tokens sampled before the cap, without an additional stopping token. The resulting turn is handled by the fixed parser, including its invalid-action behavior. Recomputed token probabilities use the actual contexts and the same θ_{old} that generated the batch.

Let $v_1, \ldots, v_N$ be the batch nodes, with classes and representatives indexed by (j, c). The scalar objective used for one gradient-ascent step is

$$\mathcal{L}_{\mathrm{batch}}(\theta) = \frac{1}{NM} \sum_{j=1}^{N} \sum_{c=1}^{K_j} \mathrm{stopgrad}(n_{jc} G(\tau_{jI_c}))\, \ell_\theta(\tau_{jI_c} \mid v_j). \tag{10}$$

Starting at $\theta = \theta_{\mathrm{old}}$, take the single update $\theta \leftarrow \theta_{\mathrm{old}} + \eta \nabla_\theta \mathcal{L}_{\mathrm{batch}}(\theta_{\mathrm{old}})$, where $\eta > 0$ is the learning rate. The gradient is summed over generated tokens, not averaged by trajectory length. We do not subtract a return baseline, standardize selected returns, clip class weights, or divide by $\sum_j K_j$. All trajectories are discarded after this update. Thus the implemented gradient is exactly the average of the node estimators in Eq. (7).

A.2.3 Complete Batch Procedure

Algorithm 1 specifies the update, including terminal candidates and the resource admission rule. One interaction is one attempted agent action, including an invalid attempt. Environment terminal flags take precedence over further generation; the remaining interaction counter enforces the common cap H. The outer loop invokes this procedure only when its reservation succeeds.

Algorithm 1 One ECE collection and learning update

Require: Current policy θ, fixed node distribution μ, width $M \geq 1$, node count $N \geq 1$, horizon $H \geq 1$, learning rate η, resource ledger

Reserve the worst-case cost of a complete batch (Appendix A.2.4); return without sampling if unavailable
Freeze $\theta_{\text{old}} \leftarrow \theta$ and draw $v_1, \ldots, v_N \overset{\text{iid}}{\sim} \mu$
for $j = 1, \ldots, N$ **do**
    Read $(s, h) \leftarrow v_j$
    for $i = 1, \ldots, M$ **do**
        Sample $u_i = (\xi_i, a_i)$ under $\pi_{\theta_{\text{old}}}$ conditioned on h
        Restore an independent copy of s and execute a_i once
        Save $z_i = (s_i^+, o_i^+, r_i, d_i)$, $h_i^+ = h \oplus u_i \oplus o_i^+$, and incurred costs
    end for
    Partition candidate indices by exact equality of $\text{enc}(z_i)$
    for each class $\mathcal{C}_c$ **do**
        Sample I_c uniformly in $\mathcal{C}_c$; set $n_c \leftarrow |\mathcal{C}_c|$
        Initialize τ_{jI_c} with the recorded probe and its reward
        if $d_{I_c} = 0$ **and** $H > 1$ **then**
            Restore $s_{I_c}^+$ and resume its original history $h_{I_c}^+$
            Use $\pi_{\theta_{\text{old}}}$ until the first terminal transition, capped at $H - 1$ further interactions
            Append the continuation to τ_{jI_c} without re-executing the probe
        end if
        Store the complete trajectory, its undiscounted return, and multiplicity n_c
    end for
end for
At θ_{old}, differentiate $\mathcal{L}_{\text{batch}}$ in Eq. (10)
Take one gradient-ascent step, discard the batch, charge actual costs, and release unused reservations

A.2.4 Resource Reservations and Variable-Length Rollouts

For a given candidate pool, couple direct sampling and ECE by assigning every candidate the same potential continuation in both methods. Let $\lambda_i \in \{0, \ldots, H-1\}$ be the number of further interactions in candidate i's continuation; it is zero after a terminal probe. For that coupling,

$$C_{\text{direct}} = M + \sum_{i=1}^{M} \lambda_i, \qquad C_{\text{ECE}} = M + \sum_{c=1}^{K} \lambda_{I_c} \leq C_{\text{direct}}. \tag{11}$$

This pathwise comparison concerns interaction counts with common candidates and continuations. Averaging over selection gives a conditional saving $\sum_c (1 - 1/n_c) \sum_{i \in \mathcal{C}_c} \mathbb{E}[\lambda_i \mid \mathcal{F}] \geq 0$. Independent training runs need not realize this coupling and incur different system overheads. When $K = M$, ECE samples every candidate and saves no continuation interactions; when $H = 1$, both methods incur exactly M interactions.

Before any candidates for an update are sampled, the ledger reserves NMH interaction attempts. Let U and V be the fixed upper bounds on generated and input tokens per model turn, and let N_{call} be the enforced maximum number of underlying calls per attempted action. The corresponding collection bounds are $NMHU$ generated tokens, $NMHV$ input tokens, and $NMHN_{\text{call}}$ underlying calls. These bounds include the M initial generations and probes at each node. A controlled resource must have an enforced finite bound for this admission rule to apply. A batch starts only if all reservations fit the remaining budget. All representatives are completed once a batch starts; there is no truncation based on remaining budget, early return estimates, or trajectory length.

Unused reservations are released only after the batch is complete. Savings fund complete subsequent batches drawn from the fixed distribution μ, without reward-dependent node selection. When a full new reservation cannot be met, training stops and records the unspent balance. This conservative rule prevents a partially completed update from favoring short trajectories. Comparisons use the same admission rule and report actual spending together with the unspent balance under each resource cap.

The ledger separately records reference-bank construction, candidate generation, all probes, continuation interactions, input and output tokens, underlying calls, snapshot capture and restoration, serialization and comparison, likelihood recomputation, optimization time, and wall-clock time. Shared-prefix caching and batching are available to all compared collectors. Interaction-capped and token-capped runs are separate comparisons. Wall-clock efficiency is a distinct system-level measure.

A.2.5 Allocation Controls and Evaluation

The full-width direct collector draws M independent turns at each node and continues all of them. Its gradient uses the ordinary width-M average with the same return, token mask, and single-update protocol. A parsed-action control groups the same candidate indices by the parser's complete structured command, with every execution-relevant argument retained, then samples one representative per group and uses that group's actual multiplicity.

To isolate the grouping criterion from the number of continuations, a random partition control uses each observed ECE pool's K and class sizes $(n_1, \ldots, n_K)$. It uniformly permutes the candidate indices and partitions the permutation into blocks of those sizes, independently of continuation randomness. One representative is sampled uniformly from each block, with block-size weighting. ECE and this control incur the same candidate and probe costs and select the same number of trajectories; different continuation lengths are charged as realized. ECE, direct sampling, parsed-action grouping, and size-matched random grouping share the fixed node bank, admission rule, and resource caps. Common-pool comparisons share candidates, while complete training runs spend their available resources on fresh batches from μ.

Final policy evaluation starts from the original task initialization and uses the standard observation interface and a common finite interaction cap, without branching. The collector's snapshots and effect partitions are not available during evaluation. Task-start performance and total resource use assess learning efficiency; duplicate-effect fractions and the identities above only characterize collection and estimation.

A.3 Online single-branch collection

Online-ECE applies the same allocation rule at nodes visited by the current policy, without constructing a frozen node bank. Let ν be the uniform distribution over training tasks. The online objective is the task-start return

$$J^{\mathrm{start}}_{\nu,H}(\theta) = \mathbb{E}_{x\sim\nu}\,\mathbb{E}_{\sigma\sim p_\theta(\cdot|x)}[G(\sigma)]. \tag{12}$$

Each batch uses one frozen policy $\pi_{\theta_{\mathrm{old}}}$; fresh prefixes are generated after every update. Before generating a prefix, we independently sample a branching depth a, uniformly from $\{0, 4, 8\}$ for ALFWorld or $\{0, 2, 4\}$ for WebShop. The policy generates a prefix p until a attempted actions or the first terminal transition. The episode horizons are $H = 50$ and $H = 15$, respectively, including invalid actions.

Task-start gradient. If the prefix is terminal, it supplies one ordinary complete-trajectory policy-gradient contribution. Otherwise, its final node $v(p)$ is the branching point. ECE samples $M = 8$ independent candidate turns, probes them from independent restorations of this node, and groups their complete one-step effects as in Section 4.2. It selects one representative I_c uniformly from each class $\mathcal{C}_c$, independently across classes and of continuation randomness. Representatives resume from their saved post-probe states with their own histories, without repeating the probe or branching again. The probe is action $a + 1$, leaving at most $H - a - 1$ further interactions before the original episode horizon. A terminal probe completes its trajectory without further actions.

For $\sigma_{I_c} = p \oplus \tau_{I_c}$, the complete task-start trajectory, the estimator is

$$\widehat{g}_{\mathrm{online}} = \frac{1}{M}\sum_{c=1}^{K} n_c\, G(\sigma_{I_c})\ \nabla_\theta \log p_\theta(\sigma_{I_c} \mid x)|_{\theta=\theta_{\mathrm{old}}}\,. \tag{13}$$

The likelihood includes sampled prefix tokens, candidate reasoning and action tokens, and continuation tokens. Environment responses have no score terms. Each complete return includes prefix

rewards once and retains the undiscounted task rewards and stopping rules of Section 4.1. The prefix is generated by the policy being differentiated, so its score is included in Eq. (13); no separate prefix update is added. Returns and class sizes are held constant during differentiation.

Conditional on the prefix, candidate turns, and observed effects, inverse inclusion weighting recovers the mean contribution of all potential complete trajectories, including their prefix scores. Averaging over candidate sampling, prefix generation, and tasks therefore recovers $\nabla_\theta J^{\mathrm{start}}_{\nu,H}(\theta_{\mathrm{old}})$. The branching depth is independent of actions, rewards, and policy parameters; terminal prefixes contribute their complete trajectory directly. This identity applies to one frozen-policy update under the regularity conditions in Appendix D.1.1. Convergence and budget efficiency of the evolving online process require separate analysis because its visited-state distribution changes during training; the constants in Theorem 1 assume fixed μ.

Controls and update. Online direct sampling shares the prefix protocol and completes $w = 4$ independent continuations. The optimized parser generates $M = 8$ candidate turns, groups their complete parsed commands before execution, and probes only uniformly selected representatives. It uses actual group-size weights and the same complete-trajectory score as ECE. All methods average over the original number of task-start pools, retaining terminal prefixes as complete trajectories. Each fresh batch supplies one additive policy-gradient step, with zero baseline, no selected-group normalization or weight clipping, and no trajectory reuse. Generation, continuation, and likelihood evaluation use the same frozen policy and token protocol in Appendix A.2.2, with the online prefix included in the generated-token mask.

Resource accounting. Each pool executes its prefix once. For a nonterminal prefix, let $|p|$ be its number of attempted actions and L_i the number of interactions after candidate i's initial action. The realized costs are

$$C_{\mathrm{online,ECE}} = |p| + M + \sum_{c=1}^{K} L_{I_c}, \qquad C_{\mathrm{online,direct}} = |p| + w + \sum_{i=1}^{w} L_i. \tag{14}$$

A terminal prefix costs only $|p|$. Parser costs include its prefix, selected probes, and continuations; omitted probes incur no interaction charge, while all generated candidates incur token charges. Restored counters preserve the remaining episode allowance, so every complete trajectory uses at most H attempted actions.

For candidate width $m \geq 1$, the complete-pool bound satisfies $a + m(H - a) \leq mH$. All methods reserve the uniform upper bound mH per pool before sampling tasks, depths, or prefixes, along with finite token and underlying-call allowances. This keeps admission independent of the sampled depth and outcomes. Admitted pools run to their defined stopping points; no pool is stopped early because of its realized reward or length. Unused reservations are released after the complete batch. The ledger charges prefix interactions once and counts all input and generated tokens, including prefix tokens reused as model inputs, along with executed probes, continuations, and actual setup costs. Online collection incurs no node-bank construction cost.

Task manifests, initial checkpoints, branching-depth schedules, and evaluation seeds are fixed before training and shared across online collectors. Training-seed and task identities define the pairing for uncertainty estimation. Hyperparameter selection uses development tasks and is charged separately from evaluation. Evaluation starts from the original task initialization and follows one ordinary trajectory without branching.

A.4 Host-adapter conventions

For the host comparisons in Table 2, ECE supplies a multiplicity-weighted batch to the host's native advantage, credit, and loss operators. Multiplicities enter both the empirical statistics and the loss reduction, including statistics used in nonlinear normalization and clipping.

Multiplicity-expanded batch. For a batch of N sampled nodes, each with candidate width M and $\sum_c n_{jc} = M$, define

$$\mathcal{D}_{\mathrm{exp}} = \biguplus_{j,c} \{\tau_{jI_c}\}^{\uplus n_{jc}}. \tag{15}$$

Repeated copies encode empirical mass only. They share the selected trajectory's actual history, rewards, old log probabilities, and value predictions. Expansion preserves node-level group boundaries and the selected trajectory's native grouping metadata. We implement it using weights, without tensor duplication or stochastic augmentation of virtual copies.

Let $\mathcal{H}_h$ be the unmodified advantage and credit operator of host h, and let $\mathcal{L}_h$ be its loss reduction, including its original clipping, KL, entropy, value-loss, and token-mask conventions. The adapter is defined by

$$\begin{aligned} A_h^{\mathrm{ECE}} &= \mathcal{H}_h(\mathcal{D}_{\mathrm{exp}}), \\ \mathcal{L}_h^{\mathrm{ECE}}(\theta) &= \mathcal{L}_h(\theta; \mathcal{D}_{\mathrm{exp}}, A_h^{\mathrm{ECE}}). \end{aligned} \tag{16}$$

We record the exact host commit and configuration defining these operators. For GiGPO and HGPO, the released step-, state-, and history-grouping operators act on the expanded empirical measure with their original grouping keys and normalizers.

GRPO reward statistics. For one original node-level group of width M, with $\sum_c n_c = M$, the multiplicity-weighted statistics are

$$\begin{aligned} \bar{G}_w &= \frac{1}{M} \sum_c n_c G(\tau_{I_c}), \\ s_w^2 &= \frac{1}{M} \sum_c n_c \big(G(\tau_{I_c}) - \bar{G}_w\big)^2, \\ A_c &= \frac{G(\tau_{I_c}) - \bar{G}_w}{\sqrt{s_w^2 + \epsilon_{\mathrm{adv}}}}. \end{aligned} \tag{17}$$

The variance uses divisor M, without a Bessel correction. The run configuration fixes the numerical stabilizer and treatment of zero-variance groups. These statistics use the weighted executed representatives, so s_w describes their expanded empirical distribution. Replacing n_c/M with $1/K$ gives a different, equally weighted representative distribution and changes the adapter.

A trajectory-normalized clipped-surrogate example. Let $\mathcal{I}_{jc}$ contain the actual generated tokens of representative (j, c). The token probability ratio is

$$r_{jck}(\theta) = \frac{\pi_\theta(y_{jck} \mid h_{jck})}{\pi_{\theta_{\mathrm{old}}}(y_{jck} \mid h_{jck})}. \tag{18}$$

For a host that averages token losses within trajectories and then averages trajectories, its clipped actor surrogate is

$$\mathcal{L}_{\mathrm{actor}}^{\mathrm{ECE}} = \frac{1}{NM} \sum_{j,c} \frac{n_{jc}}{|\mathcal{I}_{jc}|} \sum_{k \in \mathcal{I}_{jc}} \min\{r_{jck} A_{jck}, \mathrm{clip}(r_{jck}, 1-\epsilon_h, 1+\epsilon_h) A_{jck}\}. \tag{19}$$

This expression is written in maximization form; a minimizing actor loss uses its negative. Advantages and multiplicities are held constant in the actor gradient. Multiplicity weights the surrogate, leaving probability ratios and clipping thresholds unchanged.

For global token normalization, remove the per-trajectory factors $1/|\mathcal{I}_{jc}|$ from Eq. (19) and replace the outer denominator NM with $\sum_{j,c} n_{jc}|\mathcal{I}_{jc}|$. The weighted token-surrogate sum is thus divided by the weighted token count. The adapter follows the host's original choice of reduction.

PPO's selected-trajectory value targets and GAE use that trajectory's actual rewards, values, and stopping flags. Any batch-level advantage whitening uses the same weighted empirical measure. KL, entropy, and value terms retain the host's original coefficients and corresponding expanded-batch reductions. Native optimizer steps, trajectory reuse, and gradient clipping are documented separately in the run configuration.

These conventions supply a native host objective with weighted observed trajectories. The additive-PG unbiasedness identity remains specific to the additive objective; nonlinear host objectives are evaluated empirically.

B Experimental Details

B.1 Preliminary-Study Protocol and Metrics

B.1.1 Model, nodes, and sampling

The preliminary policy is a frozen Qwen2.5-3B-Instruct model, with the same decoding protocol in ALFWorld and WebShop. Inference uses BF16 with dropout disabled. Candidate turns are sampled at temperature 1.0 and top-p 1.0, without top-k truncation or a repetition penalty, with at most 256 generated tokens per turn. Model parameters remain fixed, and prompts contain no effect-class or ECE-specific information.

For each environment, we first collect frozen-policy trajectories on official training tasks and select 260 nonterminal decision nodes from at least 65 distinct tasks. Candidate nodes are stratified into four intervals according to their relative position in a trajectory, with at most one node per task per interval, to limit the contribution of long trajectories. Selection does not depend on action diversity, effect multiplicity, reward, or eventual task success.

At each node $v = (s, h)$, we independently sample a fixed pool of $M_{\max} = 32$ candidate turns from the same policy history h and use its prefixes to form nested pools of widths $M \in \{4, 8, 12, 16, 24, 32\}$. Each candidate is probed once from the same restored state s. All candidates retain their original language histories and are independently continued, including candidates with identical one-step effects, until termination or a common interaction horizon $H = 12$, including the probe. Let $L_i \in \{0, \dots, 11\}$ count candidate i's actual continuation interactions after the probe; $L_i = 0$ if the probe terminates the episode. Effect information enters only the analysis after all trajectories have been collected.

B.1.2 Observations and diagnostics

The six widths yield 3,120 node–width observations. An observation is identified by the tuple `(env, node, M)`; the same node therefore appears at multiple widths. The summaries in Appendix C.1 describe these repeated observations without assuming that all rows are independent.

For a pool of M sampled candidates, U_{raw} counts distinct raw language actions, U_{parse} counts distinct parser-normalized actions, and U_{eff} counts distinct one-step environmental effects. The effect of candidate i is

$$z_i = \left(s_i^+, o_i^+, r_i, d_i\right), \tag{20}$$

where s_i^+ is the next environment state, o_i^+ the returned observation, r_i the immediate reward, and d_i the terminal flag. Parser and effect retention are

$$R_{\text{parse}} = \frac{U_{\text{parse}}}{U_{\text{raw}}}, \qquad R_{\text{effect}} = \frac{U_{\text{eff}}}{U_{\text{parse}}}. \tag{21}$$

The first ratio measures diversity retained after parsing; the second measures diversity retained after environment execution, relative to parser-level diversity. Additional effect collapse is

$$D_{\text{extra}} = \frac{U_{\text{parse}} - U_{\text{eff}}}{M}. \tag{22}$$

It counts parser-level distinctions that disappear after execution, normalized by the sampled pool size. All observations satisfy $1 \le U_{\text{eff}} \le U_{\text{parse}} \le U_{\text{raw}} \le M$. Mean remaining depth is $\bar{L} = M^{-1} \sum_{i=1}^{M} L_i$.

To measure the continuation cost associated with redundancy beyond parsing, let $\mathcal{P}$ partition candidates by their complete parsed actions and $\mathcal{Z}$ partition them by their complete one-step effects z_i. For any grouping $\mathcal{G}$, define its duplicate continuation mass by

$$\mathcal{R}(\mathcal{G}) = \sum_{g \in \mathcal{G}} (|g| - 1)\bar{L}_g, \qquad \bar{L}_g = \frac{1}{|g|} \sum_{i \in g} L_i. \tag{23}$$

Here, $|g| - 1$ counts members beyond one representative and $\bar{L}_g$ is their group's mean observed continuation length. The extra mass revealed by effect grouping is

$$\boxed{\mathcal{A}_{\text{extra}} = \frac{\max\{\mathcal{R}(\mathcal{Z}) - \mathcal{R}(\mathcal{P}), 0\}}{M}}. \tag{24}$$

Its unit is environment steps per sampled candidate. It measures how much additional continuation mass is associated with effect duplicates beyond the duplication identified by parsed actions. The statistic is computed from fully executed trajectories and measures potential savings. It allows candidates with equal one-step effects to have different future values.

B.1.3 AGGREGATION ACROSS CANDIDATE WIDTHS

Table 6 reports all six widths. Within each environment–width group, each retention ratio is computed per node and then summarized by its median. The retention pairs are thus *coordinate-wise medians* of the node-level ratios. The prevalence of additional collapse is the percentage of the 260 nodes with $D_{\text{extra}} > 0$. The continuation-mass column is the median over all 260 nodes, including those with zero additional collapse.

B.1.4 DEPTH STRATIFICATION

For each environment separately, we retain observations with $D_{\text{extra}} > 0$, pool all six widths, and divide $\bar{L}$ into four quantile bins using `pandas.qcut`. Zero-collapse observations are excluded because $\mathcal{A}_{\text{extra}}/D_{\text{extra}}$ is undefined; positive-collapse observations with zero continuation mass remain included. This leaves 1,133 observations in ALFWorld and 1,218 in WebShop. Equal depth values stay in the same bin, giving slightly unequal counts. The three internal cutpoints are $(4.375, 5.53125, 6.6875)$ for ALFWorld and $(31/6, 6.3125, 7.3671875)$ for WebShop. Bins include their upper boundary, and the first bin also includes the minimum observed depth.

B.1.5 FIGURE CONSTRUCTION AND INTERPRETATION

In Figure 2(a), gray points show all node–width observations at the stored display coordinates, which are displaced for visibility. Colored lines connect coordinate-wise medians calculated from the original metric columns. Arrows indicate increasing M; numeric labels identify selected widths, and the diamond marks the common $M = 4$ median. Display-coordinate displacement does not affect any summary.

Panel (b) pools both environments and all widths in fixed hexagonal cells over $(D_{\text{extra}}, \bar{L})$, using `hexbin` with `gridsize=(12,9)` and `extent=(0,0.6,0,12)`. The coordinates and cell means use the original metric columns. Cells with at least four observations are colored by their arithmetic mean $\mathcal{A}_{\text{extra}}$, on a scale from 0 to 4 steps per candidate; cells with one to three observations have gray outlines. There are 132 occupied cells: 90 colored cells contain 3,052 observations, and 42 outline-only cells contain the remaining 68. Panel (b) has no displaced scatter layer.

These diagnostics describe one-step redundancy and its associated interaction costs. Learning benefits and wall-clock efficiency require separate evaluation because candidates with equal immediate effects can retain different histories and future behavior.

B.2 BENCHMARKS, SPLITS, AND EVALUATION INTERFACES

Interactive environments. ALFWorld (Shridhar et al., 2021) uses the text-based simulator. We retain its six task types and report the standard seen-environment and unseen-environment evaluation sets separately, containing 140 and 134 tasks, respectively. WebShop (Yao et al., 2022) uses the local simulator, a fixed product catalog, and the released instruction splits. ScienceWorld (Wang et al., 2022) extends evaluation to scientific experimentation and longer dependencies between actions. For every benchmark, task identifiers, task variations, and all derived nodes remain within their original split. When a release has no dedicated development set, we hold out a fixed 10% of training-task identifiers. Neither successful trajectories nor high-redundancy tasks are selected for the test set.

The horizons in Table 5 define the harmonized interface used in the reported comparisons. The standard ALFWorld unseen split measures generalization to new environments within the same task types.

Search-augmented question answering. We follow the Search-R1 data and retrieval protocol (Jin et al., 2025), training on the released Natural Questions–HotpotQA mixture and reporting results separately on Natural Questions, TriviaQA, PopQA, HotpotQA, 2WikiMultiHopQA, MuSiQue, and

Table 5: Harmonized task-start evaluation protocol. The horizon counts attempted actions, including invalid actions. Continuations inherit the snapshot's remaining episode allowance.

Environment	Horizon	Reported task metrics
ALFWorld	50	Success rate: seen / unseen
WebShop	15	Success rate; task score
ScienceWorld	30	Success rate; final progress score
Search-augmented QA	4	Exact match; token-level F1

Bamboogle. We fix the local Wikipedia corpus, E5 retriever, index, and tie-breaking order, returning three passages per query. Each episode permits at most three search calls and one final-answer turn. The same corpus and retrieval responses are available to every method. ARPO-local and T^{2}PO (QA adapt.) use this shared training mixture, fixed local retriever, and search-call limit, while retaining their respective training objectives (Dong et al., 2026; Wang et al., 2026). The exact-effect comparison uses deterministic local retrieval with restorable state; live-web search is outside this setting.

Shared QA setup allowance. Panels A and B of Table 3 share one allowance, $S_{\text{int}}^{\text{QA}} = 169{,}000$, determined by the largest measured setup cost across their twelve trained methods (T^{2}PO, QA adaptation). Their common total interaction cap is $B_{\text{total}} = 169{,}000 + 10^6 = 1{,}169{,}000$. A method with actual setup cost s_m has $B_{\text{total}} - s_m$ interactions available for subsequent collection and training. The ledger below reports setup costs and remaining allowances in thousands of interactions.

Panel A: trained algorithms			Panel B: common PG		
System	Setup	Remaining	Collector	Setup	Remaining
Search-R1 (PPO)	61	1,108	Direct ($w = 1$)	91	1,078
GRPO	64	1,105	Direct ($w = 8$)	91	1,078
GiGPO	75	1,094	Direct (w_{dev})	96	1,073
Tree-GRPO	128	1,041	Optimized parser	104	1,065
ARPO-local	143	1,026	Random grouping	102	1,067
T^{2}PO (QA adapt.)	169	1,000	ECE	115	1,054

For example, ECE retains $1{,}054{,}000$ interactions after setup, while T^{2}PO (QA adapt.) retains $1{,}000{,}000$. These are available upper bounds; complete-batch admission can leave an unspent balance.

B.3 Task, efficiency, and mechanism metrics

Task performance. Success rate is the percentage of tasks satisfying the benchmark's success predicate within the interaction limit. WebShop additionally reports 100 times its mean native terminal reward in $[0, 1]$; ScienceWorld reports mean final evaluator progress after clipping negative terminal scores to zero. QA exact match uses normalized final-answer equality against any accepted reference; token-level F1 takes the maximum over reference answers. We report per-dataset EM and unweighted macro-averages of EM and F1. Malformed final answers and unsuccessful terminations remain failures. Each evaluation trial executes one unbranched trajectory from the original task start, without state restoration or effect-based selection. We use temperature 0.4 for interactive environments and greedy decoding for QA, consistently across methods, with a single attempt and no best-of-n or evaluator-guided selection. Independent evaluation seeds estimate mean single-trajectory success; their outcomes are averaged, without selecting the best attempt.

ScienceWorld reporting and success. Native ScienceWorld scores are retained unchanged during training and trajectory collection. For reporting final progress only, episode i contributes $\text{Progress}_i = \max(0, \text{nativeFinalScore}_i)$. Clipping is applied only to the final evaluator score; individual rewards and intermediate scores retain their native values. SR is the percentage of episodes in which the benchmark goal is completed within 30 attempted actions, including invalid actions. Timeout, invalid-action termination, and other unsuccessful terminal states remain failures.

Budgeted performance. Figure 4 reports task-start success at the three prescribed caps $f \in \{1/4, 1/2, 1\}$. Interaction-capped and token-capped experiments are separate runs. A cap is an upper bound on actual spending; the complete-batch admission rule can leave an unspent balance.

Mechanism validation. Appendix C.3 reports a common-pool comparison of continuations, distinct effects actually continued, and charged interaction and input/output-token costs. Effect coverage is measured locally within each task and node. Invalid actions remain in the candidate distribution and consume interaction attempts. The preliminary study measures retention and continuation mass under the protocol in Appendix B.1. Appendix C.4 additionally compares ECE and random grouping by relative interaction cost and gradient-direction fidelity under a frozen-policy diagnostic width sweep.

B.4 Allocation-controlled comparisons

Main-result comparison protocol. In Table 1, the common interaction cap is $S_{\text{int}} + 10^6$. The direct-sampling width w_{dev} is selected on development tasks separately for each backbone and environment from $\{1, 2, 4, 8, 16\}$; grouping methods use $M = 8$. The selected widths for Table 1 are:

Backbone	ALFWorld w_{dev}	WebShop w_{dev}
Qwen2.5-3B-Instruct	2	2
Qwen2.5-7B-Instruct	4	4
Qwen3-4B	4	2

Evaluation uses one ordinary trajectory, with horizons 50 and 15 for ALFWorld and WebShop, respectively. The shared node bank, optimized parser, and size-matched random control follow the protocols below.

ALFWorld evaluation aggregation. All ALFWorld scores in Table 1 average three training seeds, each evaluated with three evaluation seeds on all 140 seen and 134 unseen tasks. Each task–evaluation-seed pair uses one ordinary trajectory, giving 420 seen and 402 unseen trials per training seed. We average unrounded training-seed success rates and compute ALFWorld contrasts before rounding to one decimal. For example, the Qwen3-4B ECE training seeds yield 382, 383, and 385 seen successes out of 420 trials each, and 358, 360, and 363 unseen successes out of 402 trials each. With equal denominators, averaging these rates equals pooling the counts: $1150/1260 \approx 91.2698\%$ seen and $1081/1206 \approx 89.6352\%$ unseen, reported as 91.3 and 89.6.

Shared objective and node distribution. For each backbone and training split, a frozen copy of the initial policy constructs a task-balanced node bank before optimization. We collect up to four reference episodes per task and sample four nonterminal node records uniformly without replacement from the resulting records. Tasks unable to supply this quota are identified before learning; bank exclusions are shared across allocation methods and reported, while the benchmark evaluation set remains unchanged. The bank is frozen across methods and training seeds. Each update draws $N = 16$ nodes independently with replacement. Node selection is independent of effect multiplicity, return, and eventual success. Saved interaction counters are restored, so branching does not grant a fresh episode allowance.

Direct and grouping controls. Direct sampling completes w independent trajectories per node, with no extra candidate-generation or probing stage. We include $w = 1$, $w = M$, and a validation-selected width from $\{1, 2, 4, 8, 16\}$. Width one provides the budget-efficient reference identified by the fixed-policy analysis. Parsed-action grouping uses the complete structured command, retaining every execution-relevant argument. Each grouping control samples one index uniformly per group and uses that group's actual multiplicity. Size-matched random grouping uniformly permutes the candidate indices and partitions them into blocks with ECE's observed class-size multiset. It then uses uniform representative selection and block-size weighting.

The common-pool validation in Appendix C.3 fixes candidate turns and their potential continuations to isolate grouping from proposal quality. Independent learning runs use fresh candidates from their

own current policies. Random grouping matches ECE's class-size multiset and number of selected representatives; continuation lengths and the number of nonterminal representatives can differ. The optimized parser control groups before execution and probes only selected representatives; omitted probes are not charged.

ECE update. ECE uses $M = 8$ candidates per node by default and uniformly samples one representative from every observed effect class; there is no additional cap on the number of selected classes. All selected trajectories retain their own original reasoning, action, and observation history, and resume from saved post-probe states. Grouping occurs only at the sampled node, and the initial probe is never repeated. The update is

$$\theta \leftarrow \theta_{\text{old}} + \frac{\eta}{NM} \sum_{j=1}^{N} \sum_{c=1}^{K_j} \text{stopgrad}(n_{jc} G(\tau_{jI_c})) \nabla_\theta \log p_\theta(\tau_{jI_c} \mid v_j)\big|_{\theta=\theta_{\text{old}}}. \tag{25}$$

Each fresh batch is collected under a frozen policy. We use the undiscounted native task return, zero baseline, and a single plain gradient-ascent step per batch. The controlled update excludes selected-group normalization, class-weight clipping, gradient clipping, KL penalties, momentum, and trajectory reuse. Generated-token log probabilities are summed over each trajectory. Fixed prefixes, environment outputs, and inserted delimiters are masked out.

B.5 Host-adapter and complete-algorithm comparisons

Table 2 compares the native chain collector, parsed-action grouping, ARPO-EAR, T^{2}PO-TDS, the Tree-GRPO sampler, and ECE within GRPO, GiGPO (std), and HGPO. These are exploration components within each host, retaining its native training objective and optimizer under a shared backbone, resource cap, and evaluation interface. Here ARPO-EAR denotes the entropy-based adaptive rollout component of Agentic Reinforced Policy Optimization (Dong et al., 2026): it uses entropy changes after tool feedback to direct additional branching. T^{2}PO-TDS denotes turn-level dynamical sampling from Token- and Turn-level Policy Optimization (Wang et al., 2026): a small change in the turn-level uncertainty signal triggers regeneration at the same state, subject to a resampling limit. We transplant these collection rules into each host while preserving its native objective, optimizer, and credit assignment, under the common environment interface and resource ledger. The displayed ECE increments use the strongest non-ECE collector separately for each host and metric. The allocation controls in Table 1 use the common additive PG update, to which the unbiasedness guarantee applies. Appendix A.4 specifies the multiplicity-expanded batch, reward statistics, and loss reductions used by the ECE adapter.

Table 2 training and evaluation. Every row uses `Qwen/Qwen2.5-7B-Instruct` at the largest interaction budget, $S_{\text{int}} + 1.0 \times 10^6$. We use three training seeds and three independent evaluation seeds per checkpoint, with evaluation temperature 0.4. Each evaluation trial follows one trajectory without test-time branching. The horizons are 50, 15, and 30 attempted actions for ALFWorld, WebShop, and ScienceWorld, respectively. Repeated evaluation outcomes are averaged, without selecting the best attempt.

Method selection and labeling. Baseline implementations include the agentic versions of PPO and GRPO used by GiGPO, together with GiGPO and HGPO (Schulman et al., 2017; Shao et al., 2024; Feng et al., 2025; He et al., 2026). ReAct (Yao et al., 2023) provides a non-updating reference. T^{2}PO (Wang et al., 2026) is included in the search QA comparison. Search-R1, Tree-GRPO, and the local-search adaptation of ARPO provide the closest search-based comparisons (Jin et al., 2025; Ji et al., 2026; Dong et al., 2026). Native losses, exploration triggers, and credit-assignment mechanisms are retained; modified environments, horizons, or cold starts are identified as harmonized adaptations. Exploration-component comparisons and complete-algorithm comparisons are labeled separately.

External supervision and reward augmentation. Primary task-only training starts from the same released checkpoint without task-specific expert demonstrations. For methods using a warm start, initialization weights and training-task supervision are matched within each comparison; distinct

external-supervision regimes are accounted for separately. Teacher inference, demonstration collection, and warm-start optimization enter the cost ledger. Native auxiliary rewards and format penalties are recorded separately from the shared benchmark evaluation reward.

Common infrastructure and tuning. All comparator scores come from reruns under the common evaluation interface. Within a panel, methods share the task split, actor initialization, retrieval resources, observation content, parser semantics, interaction limits, and total input/output limits. Method-specific memory handling and control tags are documented. Any in-context demonstrations are shared and drawn only from training tasks; no method receives additional task answers or privileged observations. Prefix caching and batching are available to every collector. Host algorithms retain their native optimizers; complete-algorithm baselines retain both their own optimizers and update schedules. Cost comparisons use the measured resource ledger, accounting for the different update schedules and trajectory counts. We initialize baseline hyperparameters from their released recipes and allocate the same development-search resource budget per method. The reported host comparison uses GiGPO (std). Test scores never select widths, hyperparameters, or checkpoints.

B.6 Implementation, budget admission, and statistics

Backbones and numerical settings. The checkpoint identifiers are `Qwen/Qwen2.5-3B-Instruct`, `Qwen/Qwen2.5-7B-Instruct`, and `Qwen/Qwen3-4B`. Qwen2.5-7B is the primary comparison backbone; the other two extend the allocation-controlled results in Table 1. We use full-parameter tuning, BF16 model computation, FP32 master weights and loss/gradient accumulation, parameter sharding, and activation checkpointing on the same eight H100-80GB GPU allocation. The nominal controlled-update learning rate is 10^{-6}; validation considers $\{10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}\}$ under an equal tuning budget. The host algorithms in Table 2 and the complete-algorithm baselines retain their native optimizers.

Collection disables dropout and uses temperature 1, top-$p = 1$, no top-k truncation, and no repetition penalty. Input context is capped at 8,192 tokens; generated turns are capped at 512 tokens for interactive environments and 1,024 for search QA. Qwen3's thinking mode is enabled across compared runs, without inserting T^{2}PO's forced thinking termination into ECE. Sampling and likelihood recomputation use identical context construction, tokenizer, and stopping rules. Oversized histories are retained in the record; a fixed oldest-turn truncation rule bounds the policy input, prioritizing the instruction and current observation. An oversized current observation uses the same fixed head–tail truncation for every method; its full untruncated response remains in the effect record. Model revisions, environment commits, parser versions, split manifests, reward definitions, and truncation rules accompany each run.

Complete-resource accounting. The resource-cap grid is $S_{\mathrm{int}} + \{1/4, 1/2, 1\} \times 10^6$ interactions and, in separate experiments, $S_{\mathrm{tok}} + \{1/4, 1/2, 1\} \times 2 \cdot 10^9$ inference tokens. Here S is a common setup allowance fixed from the largest measured setup cost among methods in the corresponding comparison before training. For search QA, this maximum is taken jointly over both panels of Table 3, as detailed in Appendix B.2. Each method is charged its actual setup cost. Inference-token accounting includes both input and generated tokens, including discarded candidates. Logical input tokens are charged even when prefix caching is used. The cap grid is fixed after development-only feasibility checks.

For ECE, admission reserves NMH attempted interactions and $NMH(V + U)$ inference tokens, using the enforced input/output caps V, U and the appropriate maximum horizon H. Analogous valid upper bounds are used for other collectors, including their finite resampling limits. A batch is admitted before observing its candidates and is then completed; unused reservations are released afterward. Admission is independent of the lengths of the trajectories subsequently collected. Training stops when another complete batch cannot be admitted. Each run retains its setup cost, actual spending, and unspent balance in the resource ledger. Savings fund fresh updates under the same task-sampling rule. Collection and system overheads remain in the ledger defined in Appendix A.2.4; evaluation and tuning are accounted for separately. The reported efficiency comparisons measure interactions and tokens.

Restoration audit. Every exact-effect adapter undergoes source-level state-field review, repeated-action replay, short-sequence replay, and branch-order independence tests. Comparisons include full returned observations, rewards, terminal flags, and all simulator/wrapper state relevant to future execution. Effect records retain IDs, action echoes, error text, ordering, and whitespace. Hashes index records; merging requires equality of their complete lossless encodings. Invalid actions consume an attempt without retry-until-valid handling. A restoration mismatch invalidates the affected collection batch and halts the run; its incurred costs remain recorded. Every adapter used in an exact-effect result, including ScienceWorld and local-retrieval QA, requires this audit. Table 13 summarizes the completed state-field audits and replay checks for all four adapters.

Uncertainty and preliminary-study protocol. Tables 1–3 and Figure 4 primarily report point estimates. Table 9 reports the four primary 7B common-PG contrasts at the largest interaction cap: ECE versus tuned direct sampling and optimized parser on ALFWorld unseen SR and WebShop SR. Their 95% hierarchical-bootstrap intervals are marginal intervals for SR differences, and significance uses Holm-adjusted $p < 0.05$ over this four-test family. Random grouping is descriptive; the ALFWorld parser contrast is inconclusive. Statistical inference is restricted to these four comparisons. Appendix B.8 gives the common-pool task-resampling protocol. The preliminary study follows the separate frozen-policy protocol in Appendix B.1.

B.7 Budget-Efficiency Protocol

The two resource constraints use the same allocation controls. The full interaction-budget column reuses the 7B results in Table 1, including the ALFWorld training-seed success counts. Direct widths are selected once on development tasks at $f = 1$, separately for each environment and resource type, then fixed over $f \in \{1/4, 1/2, 1\}$: $w^{\mathrm{int}}_{\mathrm{dev}} = 4$ in both environments; $w^{\mathrm{tok}}_{\mathrm{dev}} = 2$ for ALFWorld and 4 for WebShop.

At every interaction cap, ALFWorld averages three training seeds, each evaluated with three evaluation seeds on all 134 unseen tasks (402 trials per training seed; 1206 total). Means and contrasts use unrounded success rates before one-decimal reporting. Caps include the common setup allowance; token counts include inputs and generated tokens, including discarded candidates. Lines join only the three reported fractions, without smoothing or interpolation-based estimates.

B.8 Statistical Testing and Rollout Accounting

Primary comparison scope. Table 9 analyzes the Qwen2.5-7B-Instruct common-PG results at the largest interaction cap, with the same node bank and evaluation interfaces as Table 1. The four primary comparisons are ECE versus development-selected direct sampling and optimized parsed-action grouping, each on ALFWorld unseen SR and WebShop SR. Size-matched random grouping remains a descriptive control and is not part of this test family. The environments retain horizons 50 and 15, and each evaluation trial uses one ordinary trajectory from the task start. ALFWorld aggregates three training seeds, each with three evaluation seeds over all 134 unseen tasks.

Uncertainty and multiplicity. We report 95% hierarchical-bootstrap intervals for ECE-minus-baseline SR differences in percentage points, alongside raw and Holm-adjusted p values. Confidence intervals give marginal coverage for each comparison. Holm's step-down correction orders the four raw values as $p_{(1)} \le \cdots \le p_{(4)}$ and sets $p^{\mathrm{adj}}_{(i)} = \min\{1, \max_{j \le i}(5-j)p_{(j)}\}$. The sorted raw values are 0.0008, 0.004, 0.018, and 0.090; the corresponding adjusted values are 0.0032, 0.012, 0.036, and 0.090. Significance requires adjusted $p < 0.05$. The ALFWorld parser contrast has an interval spanning zero and an adjusted $p = 0.090$, leaving its gain beyond parser grouping inconclusive. All intervals in this table describe ECE-minus-baseline contrasts.

Paired common-pool accounting. For each environment, we sample 32 training tasks uniformly from those able to supply four nonterminal nodes in the fixed initial-policy bank, then sample four records per task without replacement. Quota-based exclusions do not condition on effects or returns. The resulting 128 nodes each use one pool of eight candidates from the frozen initial Qwen2.5-7B-Instruct policy and the remaining allowance of the recorded episode. Candidate generation uses the learning-study decoding and token caps. For each candidate, we record its full effect and one po-

tential continuation with its own history and independent random stream. Different allocation rules reuse the same potential continuation when they select that candidate. Representative and partition draws are independent of future returns and lengths. Table 10 fixes $M = 8$. The separate diagnostic in Appendix B.9 varies widths and representative draws while reusing potential continuations.

The direct control continues every candidate. Parsed grouping selects one index uniformly per complete parsed command. ECE selects one index per exact-effect class; random grouping uses a uniformly permuted partition with the same class-size multiset as ECE. The number of started continuations counts selected candidates with at least one post-probe interaction. The distinct-effects column counts their unique complete effect records. Terminal probes and exhausted horizons contribute no started continuation. Random grouping matches selected representative counts, but need not match started continuations when terminal outcomes are mixed within a random block.

Let L_i be the number of post-probe interactions in candidate i's potential continuation, let P be the number of parsed commands, and let S_A be the representatives selected by allocation rule A. The charged interaction counts are $C_D = M + \sum_{i=1}^{M} L_i$, $C_P = P + \sum_{i \in S_P} L_i$, and $C_A = M + \sum_{i \in S_A} L_i$ for $A \in \{\mathrm{random}, \mathrm{ECE}\}$. The parser's unused diagnostic probes are not charged to its collector. Token accounting includes all eight candidate generations and their logical input tokens, plus the selected continuations' input/output tokens, including tokens of discarded candidates. Fixed prefixes count even when cached; setup is recorded separately from the per-pool costs. All costs of collecting potential trajectories and auditing unselected branches remain in a separate diagnostic ledger. The table gives paired collector costs computed from fully collected potential trajectories.

Table 10 reports per-pool means with 95% bootstrap intervals from 10,000 task-level resamples, retaining each task's four nodes and all methods together. These marginal intervals describe each method's mean. The equal direct/ECE effect counts reflect ECE retaining one representative of every continuable effect class. Task-start performance and gradient diagnostics separately assess the learning and precision consequences of this allocation.

B.9 Effect-Coverage and Gradient-Fidelity Protocol

Candidate pools and allocation draws. The experiment uses the frozen initial Qwen2.5-7B-Instruct policy and the task/node sampling protocol of Appendix B.8: 32 training tasks per environment, with four nonterminal nodes per task. The diagnostic widths are $M \in \{4, 6, 8, 10, 12, 16\}$. Within each node and width, both allocation rules share the sampled candidates and their potential continuations. ECE uses complete one-step effects; random grouping uniformly permutes candidate indices into blocks with the same class-size multiset. Six representative/partition draws reuse this continuation bank. The width sweep is performed solely for this frozen-policy diagnostic.

Gradient-direction fidelity. At the frozen parameters θ_0, candidate i contributes $g_i = G(\tau_i)\, \nabla_\theta \ell_\theta(\tau_i \mid v)|_{\theta_0}$, using the full rollout and the common additive-PG token/history convention. The return and group multiplicities are treated as constants when differentiating. For allocation A with blocks $\mathcal{B}_c$, sizes n_c, and uniform representatives I_c, the estimates are

$$\widehat{g}_{\mathrm{D}} = \frac{1}{M} \sum_{i=1}^{M} g_i, \qquad \widehat{g}_A = \frac{1}{M} \sum_{c=1}^{K} n_c g_{I_c}. \tag{26}$$

All estimates use the same fixed Rademacher map $\mathsf{P} \in \mathbb{R}^{4096 \times d}$, where d is the policy parameter dimension. The map is used only for analysis. Cosine fidelity, defined for nonzero mapped vectors, is

$$F_A = \cos\big(\mathsf{P}\widehat{g}_A, \mathsf{P}\widehat{g}_{\mathrm{D}}\big). \tag{27}$$

The reference is the finite direct estimate for each pool and width. Fidelity therefore measures agreement with a pool-specific direction; population MSE and the variance components require separate diagnostics.

Interaction accounting. Let L_i denote candidate i's post-probe continuation interactions and S_A its allocation's selected representatives. Costs follow the paired accounting rule of Table 10:

$$C_{\mathrm{D}} = M + \sum_{i=1}^{M} L_i, \qquad C_A = M + \sum_{i \in S_A} L_i, \qquad \widetilde{C}_A = \frac{C_A}{C_{\mathrm{D}}}. \tag{28}$$

Thus the full-direct reference is $(\widetilde{C}, F) = (1, 1)$. The width-level statistic is the mean of per-record cost ratios. Averaging the ratios separately can give a different value from dividing the mean costs in Table 10. Setup and diagnostic collection overhead remain separate from the per-pool collector costs.

Aggregate comparison. Figure 6 displays 24 width-level aggregate estimates: two environments, six widths, and two allocation rules. For each width and method, we aggregate draws and nodes within each task and report the equally weighted task means in the heatmap. We report costs to five decimal places to resolve small differences between collectors. Task-bootstrap 95% intervals are included in the accompanying data. They describe marginal uncertainty for each mean. The heatmap displays aggregate costs and cosine fidelity; class counts, correlations, and theoretical variance components require different measurements.

B.10 Frozen-Checkpoint Cost–Variance Protocol

Common target and independent pools. At each checkpoint t, all collectors use the same frozen policy θ_t, fixed node distribution μ, and additive-PG convention. Their common expected pool estimate is $g_t = \nabla J_{\mu,H}(\theta_t)$. The analysis map P is fixed across collectors and checkpoints and is not used for training or representative selection. Each pair $Y_{a,t}, Y'_{a,t}$ consists of independent complete-pool draws, including node, candidate, representative, and continuation randomness. Nodes are drawn independently from μ and may coincide by chance; holding a node fixed for both draws would omit between-node variance. The independent-pair identity gives

$$\tfrac{1}{2}\mathbb{E}\|\mathsf{P}Y_{a,t} - \mathsf{P}Y'_{a,t}\|_2^2 = \operatorname{tr}\operatorname{Cov}(\mathsf{P}Y_{a,t}) = \mathbb{E}\|\mathsf{P}(Y_{a,t} - g_t)\|_2^2. \tag{29}$$

The last equality follows from the common additive-PG expectation.

Estimation and interaction accounting. For R independent pool pairs and a separate set of S diagnostic cost observations, we use

$$\widehat{V}^{\mathsf{P}}_{a,t} = \frac{1}{2R}\sum_{r=1}^{R}\|\mathsf{P}Y^{(r)}_{a,t} - \mathsf{P}{Y'}^{(r)}_{a,t}\|_2^2, \quad \widehat{c}_{a,t} = \frac{1}{S}\sum_{s=1}^{S} C^{(s)}_{a,t}, \quad \widehat{\Xi}_{a,t} = \frac{\widehat{c}_{a,t}\widehat{V}^{\mathsf{P}}_{a,t}}{\widehat{c}_{\mathrm{D},1,t}\widehat{V}^{\mathsf{P}}_{\mathrm{D},1,t}}. \tag{30}$$

Here $C_{a,t}$ follows the collector's complete-pool interaction ledger: direct sampling pays for all candidate probes and continuations; ECE and size-matched random grouping pay for all probes and selected continuations; the optimized parser pays for selected parsed-action representatives and their continuations. Diagnostic replication is kept separate from these per-pool costs. The independent sample sets allow dependence between a collector's realized cost and gradient. The ratio is a plug-in estimate and can have finite-sample bias.

Connection to the theorem. For a fixed linear sketch, the fixed-policy cost–variance argument applies to $\mathsf{P}Y$ and its projected moments. Hence $c_a V_a^{\mathsf{P}}$ is the leading asymptotic projected-MSE constant under the theorem's complete-pool assumptions. This diagnostic estimates the product directly; its components T, A, W, Q remain unestimated. The independent-pair identity includes between-node variability and estimates variance without estimating the population gradient explicitly. It also differs from the same-pool cosine statistic in Appendix B.9.

B.11 Class-Level Continuation Savings and Representative Variance

Expected continuation savings. Condition on a candidate pool of width M and its potential continuations. For a repeated effect class $\mathcal{C}_c$ of size $n_c \geq 2$, let L_i denote candidate i's post-probe continuation interactions. The probe cost is common to both allocations and cancels. Continuing all class members under direct sampling, versus one uniformly sampled member I_c under ECE, gives

$$B_c = \sum_{i\in\mathcal{C}_c} L_i - \mathbb{E}_{I_c\sim\mathrm{Unif}(\mathcal{C}_c)}[L_{I_c}] = \left(1 - \frac{1}{n_c}\right)\sum_{i\in\mathcal{C}_c} L_i. \tag{31}$$

This is the empirical class-level counterpart of the conditional expected saving in Appendix A.2.4. If continuation lengths are first averaged over repeated draws, the same linear expression estimates the corresponding expected saving. This averages over continuation randomness in addition to representative selection.

Representative-variance penalty. Use the full additive trajectory-gradient contribution $g_i = G(\tau_i)\nabla_\theta \ell_\theta(\tau_i \mid v)$ at the frozen parameters, with the token/history convention of Appendix B.9. Let $\widetilde{g}_i = \mathsf{P} g_i$ use the same fixed $\mathsf{P} \in \mathbb{R}^{4096\times d}$ Rademacher map, and define

$$\bar{g}_c = \frac{1}{n_c}\sum_{i\in\mathcal{C}_c} \widetilde{g}_i, \qquad \widetilde{g}_{\mathrm{D}} = \frac{1}{M}\sum_{j=1}^{M} \widetilde{g}_j. \tag{32}$$

The projected full-class contribution is $(n_c/M)\bar{g}_c$. Its ECE replacement, $(n_c/M)\widetilde{g}_{I_c}$, has the same conditional mean. Its exact conditional variance trace is

$$V_c^{\mathrm{rep}} = \mathbb{E}_{I_c}\left[\left\|\frac{n_c}{M}(\widetilde{g}_{I_c} - \bar{g}_c)\right\|_2^2\right] = \frac{n_c^2}{M^2}\frac{1}{n_c}\sum_{i\in\mathcal{C}_c}\|\widetilde{g}_i - \bar{g}_c\|_2^2. \tag{33}$$

Uniform selection from the fixed finite class gives divisor n_c; the sample-variance correction $n_c - 1$ is inapplicable here. For a fixed positive stabilizer ϵ, the plotted quantity is

$$R_c = \frac{V_c^{\mathrm{rep}}}{\|\widetilde{g}_{\mathrm{D}}\|_2^2 + \epsilon}. \tag{34}$$

This dimensionless penalty normalizes representative variance by the node's full-direct gradient energy. With independent representative draws across classes, their V_c^{rep} values sum to the conditional projected MSE relative to this pool's full-direct contribution. This conditional, pool-specific quantity differs from population gradient MSE and the theoretical within-effect variance component. A small full-direct energy can inflate R_c even when the absolute gradient discrepancy is small.

Data and visualization. Our class-level dataset contains 476 effect classes, including 130 parse-error, 150 no-state-change, 98 state-change, 68 task-progress, and 30 unknown classes. All recorded B_c and R_c values are positive; we plot every class without clipping its values or adding jitter. Both axes use logarithmic scales. Marker area, normalized for the geometric area of each marker shape, is proportional to n_c. The pooled Gaussian KDE weights each class equally and uses Scott's bandwidth in $(\log_{10} B_c, \log_{10} R_c)$ coordinates. Its three highest-density contours enclose approximately 50%, 80%, and 95% of the fitted mass, evaluated on a 400×400 grid padded by five marginal kernel standard deviations. The contours describe the density of classes; they carry no task-level confidence interpretation.

C Additional Experimental Results

C.1 Additional Preliminary-Study Results

Sampling, metrics, and aggregation are specified in Appendix B.1.

C.1.1 Summaries across candidate widths

Wider pools show lower median retention and more widespread additional collapse in both environments; median continuation mass need not increase monotonically with width.

Width	ALFWorld				WebShop			
M	Parser R_{parse}	Effect R_{effect}	Collapse (%)	Mass $\mathcal{A}_{\text{extra}}$	Parser R_{parse}	Effect R_{effect}	Collapse (%)	Mass $\mathcal{A}_{\text{extra}}$
4	1.000	1.000	25.4	0.000	1.000	1.000	33.1	0.000
8	0.857	0.857	58.8	0.446	0.857	0.833	70.0	0.724
12	0.800	0.833	76.5	0.542	0.750	0.764	83.8	0.788
16	0.733	0.800	86.5	0.635	0.714	0.721	90.4	0.882
24	0.667	0.750	92.3	0.634	0.628	0.667	94.2	0.876
32	0.625	0.730	96.2	0.577	0.588	0.625	96.9	0.806

Table 6: **Width-wise summaries.** Retentions and $\mathcal{A}_{\text{extra}}$ are medians; collapse is the percentage of nodes with $D_{\text{extra}} > 0$. Each environment–width group contains 260 nodes. Continuation mass is in steps per sampled candidate.

C.1.2 Depth-stratified continuation mass

Table 7 summarizes the resulting groups. The median normalized continuation mass rises from the shallowest to the deepest quartile in both environments. This comparison describes a descriptive association with remaining depth across the sampled observations. All trajectories were executed, so this comparison measures potential continuation savings.

Depth	ALFWorld			WebShop		
Quartile	Count N	Median $\bar{L}$	Median $\mathcal{A}_{\text{extra}}/D_{\text{extra}}$	Count N	Median $\bar{L}$	Median $\mathcal{A}_{\text{extra}}/D_{\text{extra}}$
Q1	285	3.72	3.60	307	4.25	4.05
Q2	283	4.97	4.98	303	5.78	5.51
Q3	288	6.16	6.17	303	6.75	6.86
Q4	277	7.46	7.58	305	7.88	7.95

Table 7: **Depth quartiles** among observations with $D_{\text{extra}} > 0$, formed separately by environment after pooling candidate widths. N counts node–width observations; Q1 is the shallowest quartile.

C.2 Online Validation from Task Starts

To test whether the allocation benefit depends on a frozen node bank, we compare direct sampling, optimized parsed-action grouping, and ECE under the online single-branch protocol in Appendix A.3. Each pool starts from a training task; a branching depth is sampled before prefix generation, independently of actions and rewards. The collector branches at the node reached by its current frozen policy. Prefix interactions, candidate generation, executed probes, and continuations all enter the resource ledger. Evaluation remains single-trajectory execution from the original task start.

State protocol	Collector	ALFWorld unseen SR	WebShop SR
Fixed bank	Direct ($w_{\text{dev}} = 4$)	88.4	75.5
	Optimized parser ($M = 8$)	<u>90.6</u>	<u>78.7</u>
	ECE ($M = 8$)	**91.6**	**80.6**
Online	Direct ($w = 4$)	89.1	76.0
	Optimized parser ($M = 8$)	<u>90.7</u>	<u>79.0</u>
	ECE ($M = 8$)	**92.0**	**81.2**

Table 8: **Online validation from task starts.** Qwen2.5-7B-Instruct at the full interaction cap $S_{\text{int}} + 10^6$, using a common additive PG update and single-trajectory evaluation. Fixed-bank scores reuse Table 1; online collectors generate fresh prefixes with their current policy. Collector comparisons are within each training-state protocol. SR is reported as a percentage. Best and second-best values are bold and underlined within each state protocol and metric.

Within the online block, ECE exceeds direct sampling by 2.9 and 5.2 percentage points, and the optimized parser by 1.3 and 2.2 points, on ALFWorld unseen and WebShop SR, respectively. These point estimates support an allocation benefit in the tested single-branch setting. The fixed-bank block provides context: the protocols optimize different objectives and visit different training states, so comparisons are made within each block. The online contrasts are separate from the four fixed-bank primary tests in Table 9.

C.3 Statistical Tests and Rollout Efficiency

Table 9 reports four primary tests for the 7B common-PG results at $S_{\text{int}} + 10^6$; random grouping remains descriptive. Table 10 compares allocation on frozen-policy common pools. Appendix B.8 gives the protocols; Appendix C.7 reports adapter checks.

C.3.1 Primary comparisons

Environment	Baseline	Δ SR	95% CI	p_{raw}	p_{Holm}	Conclusion
ALFWorld	Direct (w_{dev})	$+3.2$	$[1.3, 5.2]$	0.004	0.012	Significant
	Optimized parser	$+1.0$	$[-0.2, 2.1]$	0.090	0.090	Inconclusive
WebShop	Direct (w_{dev})	$+5.1$	$[2.9, 7.3]$	0.0008	0.0032	Significant
	Optimized parser	$+1.9$	$[0.4, 3.4]$	0.018	0.036	Significant

Table 9: **Primary statistical comparisons.** Qwen2.5-7B-Instruct, common PG update, and interaction cap $S_{\text{int}} + 10^6$, corresponding to Table 1. Δ is ECE minus the listed baseline in percentage points, using ALFWorld unseen-environment SR and WebShop SR. Intervals are marginal 95% hierarchical-bootstrap intervals for these differences. Holm correction covers exactly the four comparisons shown; significance requires $p_{\text{Holm}} < 0.05$.

ECE improves over development-selected direct sampling on both environments after Holm correction ($p_{\text{adj}} = 0.012$ and 0.0032). Its WebShop gain over the optimized parser is also significant ($+1.9$ points, 95% CI $[0.4, 3.4]$, $p_{\text{adj}} = 0.036$). On ALFWorld, the parser-relative interval includes zero ($+1.0$ point, 95% CI $[-0.2, 2.1]$, $p_{\text{adj}} = 0.090$). Thus, the evidence for a gain beyond parser grouping is resolved on WebShop and inconclusive on ALFWorld; equivalence on ALFWorld remains undetermined.

C.3.2 HGPO contrast uncertainty

HGPO's ECE-minus-strongest-alternative gains are assessed using 95% confidence intervals. The comparator is T^{2}PO-TDS for ALFWorld seen SR and both WebShop metrics, and the Tree-GRPO sampler for ALFWorld unseen SR and both ScienceWorld metrics. Five intervals cross zero; the WebShop SR interval is $[0.0, 1.7]$ at the reported precision. The positive point estimates therefore

remain statistically unresolved. These contrasts have no reported adjusted p values and are separate from the four primary common-PG comparisons in Table 9.

C.3.3 PAIRED ROLLOUT EFFICIENCY

Environment	Allocation	Started continuations ↓	Distinct effects continued ↑	Charged interactions ↓	Input + output tokens (10^3) ↓
ALFWorld	Direct ($w = 8$)	7.55 [7.38, 7.72]	**5.55** [5.25, 5.83]	64.8 [59.8, 70.1]	98.6 [90.5, 107.3]
	Optimized parser	6.31 [6.01, 6.60]	5.49 [5.19, 5.78]	51.8 [47.4, 56.6]	79.9 [73.1, 87.4]
	Size-matched random	**5.46** [5.17, 5.74]	4.39 [4.07, 4.69]	47.0 [43.0, 51.4]	72.0 [65.6, 79.0]
	ECE ($M = 8$)	5.55 [5.25, 5.83]	**5.55** [5.25, 5.83]	**46.6** [42.7, 50.9]	**71.5** [65.2, 78.3]
WebShop	Direct ($w = 8$)	7.46 [7.27, 7.64]	**5.18** [4.86, 5.49]	43.5 [40.0, 47.2]	132.0 [119.0, 145.6]
	Optimized parser	6.04 [5.71, 6.36]	5.11 [4.79, 5.42]	34.6 [31.7, 37.7]	108.2 [97.8, 119.6]
	Size-matched random	**5.10** [4.81, 5.39]	4.00 [3.68, 4.31]	**32.1** [29.4, 35.0]	**97.8** [88.2, 108.4]
	ECE ($M = 8$)	5.18 [4.86, 5.49]	**5.18** [4.86, 5.49]	32.4 [29.7, 35.4]	99.0 [89.3, 109.7]

Table 10: **Common-pool mechanism check.** Frozen initial Qwen2.5-7B-Instruct, $M = 8$, and 128 nodes from 32 tasks per environment. Cells report means [95% task-bootstrap CI]; bold/underlining mark first/second by the arrows within each environment. Only post-probe interactions start continuations. The parser probes only representatives; other methods probe all eight candidates. Paired costs exclude separately logged setup and diagnostic overhead (Appendix B.8).

ECE preserves direct sampling's 5.55/5.18 distinct effects per pool on ALFWorld/WebShop with 28.1%/25.5% fewer interactions and 27.5%/25.0% fewer tokens. Relative to the optimized parser, interaction costs fall 10.0%/6.4% with slightly greater effect coverage. ECE continues 26.4%/29.5% more distinct effects than random grouping at similar cost; WebShop random is slightly cheaper (32.1 versus 32.4 interactions; 97.8k versus 99.0k tokens). These estimates describe local allocation, with marginal uncertainty intervals for each method.

C.4 EFFECT COVERAGE AND GRADIENT FIDELITY

To isolate the grouping signal beyond class-size matching, we compare ECE and random grouping on common candidate pools from the frozen initial Qwen2.5-7B-Instruct policy. Each environment contributes 128 nodes from 32 training tasks. We vary $M \in \{4, 6, 8, 10, 12, 16\}$ and repeat representative/partition sampling six times, reusing each candidate's potential continuation. Fidelity is cosine agreement with the full-direct gradient estimate after a shared, fixed 4,096-dimensional Rademacher map; cost is normalized to direct sampling. No policy is trained or selected in this diagnostic. Figure 6 compares the reported width-level aggregate estimates; the definitions and aggregation procedure are given in Appendix B.9.

The common-pool audit provides a coverage check at $M = 8$ (Table 10, Appendix C.3). ECE continues 26.4%/29.5% more distinct effects than size-matched random grouping on ALFWorld/WebShop at similar cost; the random control is slightly cheaper on WebShop. Thus, similar collection costs can support different environmental coverage.

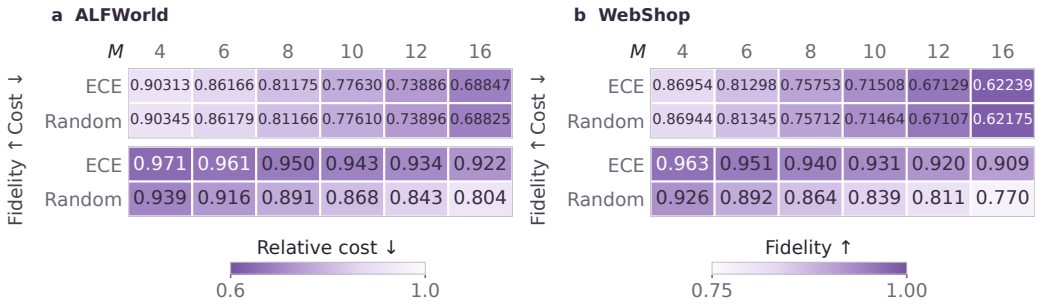


Figure 6: **Cost–gradient-fidelity trade-off.** Panels compare ECE and size-matched random grouping on (a) ALFWorld and (b) WebShop across candidate widths M. Cells report mean relative interaction cost (five decimal places) and gradient fidelity (three decimal places). Darker purple indicates lower cost or higher fidelity, with separate metric scales shared across environments. The full-direct reference is 1 for both metrics.

At $M = 8$, ALFWorld costs are nearly equal (0.81175/0.81166 for ECE/random), while ECE fidelity is 0.950 versus 0.891 for random grouping. WebShop shows the same pattern: costs are 0.75753/0.75712 and fidelities 0.940/0.864 for ECE/random. At $M = 16$, the fidelity gaps widen to 0.118 and 0.139, respectively. At similar cost, effect identities retain more of the finite-pool gradient direction than size matching alone.

The absolute differences between ECE and random mean relative costs range from 8.93×10^{-5} to 6.43×10^{-4} across the twelve environment–width pairs. Their near agreement is an empirical result. At a fixed pool, both rules pay for M probes and select K representatives, but the selected continuation lengths can differ. Conditional expected continuation cost is $\sum_c n_c^{-1} \sum_{i \in \mathcal{C}_c} L_i$ for ECE and $(K/M) \sum_i L_i$ for uniform size-matched random grouping. Matching the class-size multiset therefore does not enforce equal realized or expected interaction costs.

C.5 Cost-Normalized Gradient Variance

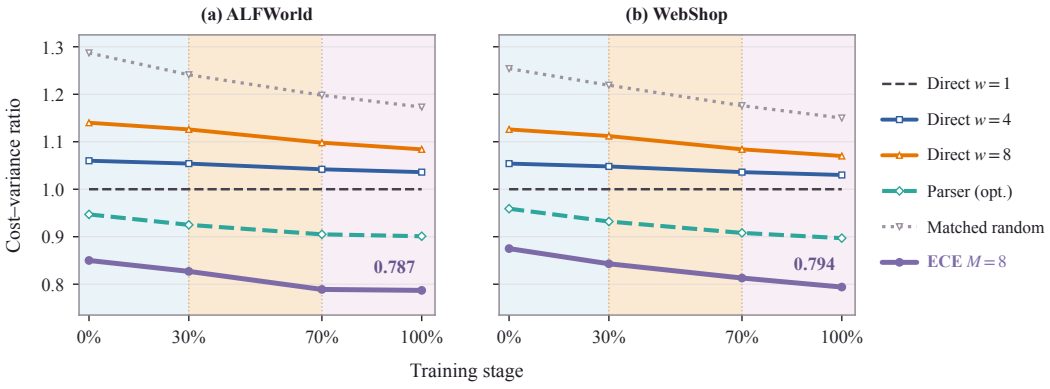


Figure 7: **Cost-normalized gradient MSE across training stages.** Mean interaction cost times projected gradient variance, relative to Direct $w = 1$ at each frozen checkpoint. Lower is better; values below one favor the collector. Markers denote four measured checkpoints, lines join them, and bands separate their intervals.

To test the cost–variance criterion of Theorem 1, we compare collectors under a common frozen policy, fixed node distribution, and original additive-PG estimator. For independent complete-pool estimates Y_a, Y_a' from collector a, including independent node draws, define

$$V_a^{\mathsf{P}} = \tfrac{1}{2}\mathbb{E}\|\mathsf{P}Y_a - \mathsf{P}Y_a'\|_2^2, \qquad \Xi_a = \frac{c_a V_a^{\mathsf{P}}}{c_{\mathrm{D},1} V_{\mathrm{D},1}^{\mathsf{P}}}. \tag{35}$$

Here P is a shared, fixed analysis-only linear sketch and c_a is mean interaction cost per complete pool. Independent pairs include node and rollout variance; under additive PG, this equals MSE around the projected population gradient. Cost and variance use independent diagnostic samples (Section B.10).

ECE's ratio stays below Direct $w = 1$ at all four stages (Figure 7); its cost–variance reduction grows from 15.0% to 21.3% on ALFWorld and 12.5% to 20.6% on WebShop. The parser also outperforms Direct, but trails ECE at every stage; size-matched random grouping stays above one despite compressing branches. These point estimates favor ECE on the theorem's cost–variance criterion in the fixed projected space. Checkpoint-specific normalization measures relative improvement.

Environment	Collector	0%	30%	70%	100%
ALFWorld	Direct ($w = 1$)	1.000	1.000	1.000	1.000
	Direct ($w = 4$)	1.060	1.054	1.042	1.036
	Direct ($w = 8$)	1.140	1.126	1.098	1.084
	Optimized parser	<u>0.947</u>	<u>0.925</u>	<u>0.905</u>	<u>0.901</u>
	Size-matched random	1.287	1.241	1.198	1.173
	ECE ($M = 8$)	**0.850**	**0.827**	**0.789**	**0.787**
WebShop	Direct ($w = 1$)	1.000	1.000	1.000	1.000
	Direct ($w = 4$)	1.054	1.048	1.036	1.030
	Direct ($w = 8$)	1.126	1.112	1.084	1.070
	Optimized parser	<u>0.959</u>	<u>0.932</u>	<u>0.908</u>	<u>0.897</u>
	Size-matched random	1.254	1.219	1.176	1.150
	ECE ($M = 8$)	**0.875**	**0.843**	**0.813**	**0.794**

Table 11: **Cost–variance ratios at frozen checkpoints.** Entries are plug-in point estimates $\widehat{\Xi}_a$ (Equation (30)), relative to Direct $w = 1$ at the same checkpoint and environment. Lower is better in the fixed projected-gradient space. Bold/underlined values are best/second best per environment and stage.

Table 11 reports point estimates for all collectors and checkpoints. Direct $w = 1$ is the reference at each checkpoint; the $w = 4$ and $w = 8$ controls provide additional width comparisons. Values below one favor the collector's measured projected criterion. The comparison is descriptive and limited to the fixed projected space at frozen checkpoints. Statistical significance, full-dimensional MSE, and training convergence require additional evidence. Each checkpoint has its own normalization, so these ratios measure relative performance and leave absolute cost and variance unspecified.

C.6 Budget-Efficiency Results

Table 12 gives the numerical results underlying Figure 4 for Qwen2.5-7B-Instruct. Appendix B.7 specifies width selection, seed aggregation, and resource accounting.

ECE has the highest success rate in eleven of the twelve environment–cap conditions, including all six interaction-budget comparisons. Under interaction caps, its advantage over development-selected direct sampling is 0.4–3.3 percentage points on ALFWorld and 5.1–6.8 on WebShop. The exception is the smallest token cap on ALFWorld: ECE scores 69.1%, versus 72.4% for tuned direct sampling and 72.1% for the parser. At the two larger token caps, ECE leads both environments. Its parser-relative token-budget gains peak at $f = 1/2$, reaching 3.3 points on ALFWorld and 3.9 on WebShop. These differences show that a method's ranking depends on the resource constraint and available budget.

Environment	Rollout allocation	Interaction budget: f			Token budget: f		
		$1/4$	$1/2$	1	$1/4$	$1/2$	1
ALFWorld	Direct sampling ($w = 1$)	<u>74.6</u>	85.1	85.5	69.2	80.9	88.0
	Direct sampling ($w = 8$)	65.7	80.0	83.5	68.1	75.7	87.3
	Direct sampling (w_{dev})	73.0	<u>86.5</u>	88.4	**72.4**	82.5	89.6
	Parsed-action grouping (optimized)	72.2	86.2	<u>90.6</u>	<u>72.1</u>	<u>82.6</u>	89.7
	Size-matched random grouping	69.4	83.3	87.0	65.9	79.7	<u>89.8</u>
	Effect-class exploration (ECE)	**76.3**	**86.9**	**91.6**	69.1	**85.9**	**91.6**
WebShop	Direct sampling ($w = 1$)	54.3	64.6	73.7	52.4	64.2	72.8
	Direct sampling ($w = 8$)	50.5	61.9	72.4	48.2	60.8	73.4
	Direct sampling (w_{dev})	54.4	66.8	75.5	52.3	65.2	74.6
	Parsed-action grouping (optimized)	<u>58.9</u>	<u>70.4</u>	<u>78.7</u>	<u>52.8</u>	<u>69.2</u>	<u>77.1</u>
	Size-matched random grouping	53.1	65.3	73.2	50.3	65.3	74.1
	Effect-class exploration (ECE)	**60.3**	**73.6**	**80.6**	**53.3**	**73.1**	**79.1**

Table 12: **Success rates under interaction and token budgets** for Qwen2.5-7B-Instruct. Budget fractions are $f \in \{1/4, 1/2, 1\}$, with caps $S_{\mathrm{int}} + f\,10^6$ interactions and $S_{\mathrm{tok}} + f\,(2 \times 10^9)$ input/output tokens. Scores are ALFWorld unseen-environment and WebShop success rates (%); bold/underlined values mark the best/second best within each environment and cap. The full interaction-budget column corresponds to the 7B results in Table 1.

C.7 Environment-Adapter Verification

Table 13 summarizes one complete state-field audit for each of the four adapters, together with 4,608 step-replay tests and 1,152 branch-order tests in total, with no observed failures. Each adapter also has 128 invalid-action and 128 terminal-case records. The source review checks state completeness, while the replay results establish agreement on the tested cases. Any restoration mismatch invalidates the affected collection batch until the adapter is repaired and its results rechecked.

Adapter	State-field review record	Step replay tested / failed	Branch-order tested / failed	Case counts invalid / terminal
ALFWorld	1 complete audit	1,024 / 0	256 / 0	128 / 128
WebShop	1 complete audit	1,024 / 0	256 / 0	128 / 128
ScienceWorld	1 complete audit	1,536 / 0	384 / 0	128 / 128
Local-retrieval QA	1 complete audit	1,024 / 0	256 / 0	128 / 128

Table 13: **Exact-effect adapter verification.** Step replay checks complete state, response, reward, and termination at every step of single actions and short sequences. The final column gives counts of invalid-action and terminal-case records, respectively. Finite replay and branch-order checks complement the source-level audit of state completeness.

D PROOFS

D.1 ESTIMATOR PROPERTIES

D.1.1 CONDITIONAL UNBIASEDNESS

Fix a collection batch and its parameters θ_{old}. The node distribution μ, environment dynamics, reward function, parsing rules, and stopping rules are independent of the parameters being differentiated. Assume the policy likelihood is differentiable with parameter-independent support in a neighborhood of θ_{old}, and that differentiation can be interchanged with the finite-horizon expectation. Finite token and interaction caps, bounded rewards, and locally integrable score functions are sufficient. For variance statements below, the corresponding second moments are also assumed finite.

Write $\mathcal{F} = \sigma(v, u_1, \dots, u_M, z_1, \dots, z_M)$. For a randomized partition, condition additionally on the realized partition. Conditional on $\mathcal{F}$, associate each candidate i with an independent random stream for its potential continuation from (s_i^+, h_i^+). Only selected continuations need to be executed. The representative random variables I_c are mutually independent given $\mathcal{F}$ and independent of these streams. In particular, conditioning on selection leaves the continuation law of candidate i unchanged.

Let $\delta_i = \mathbf{1}\{i \in S\}$. Since every class is nonempty, $q_i(\mathcal{F}) = 1/n_{c(i)} > 0$. The estimator in Eq. (7) can be written as $M^{-1}\sum_i \delta_i g_i / q_i(\mathcal{F})$. Consequently,

$$\begin{aligned}\mathbb{E}[\widehat{g}_{\text{ECE}}(v) \mid \mathcal{F}] &= \frac{1}{M}\sum_{i=1}^{M} \frac{\mathbb{E}[\delta_i g_i \mid \mathcal{F}]}{q_i(\mathcal{F})} \\ &= \frac{1}{M}\sum_{i=1}^{M} \frac{\mathbb{E}[\delta_i \mid \mathcal{F}]\,\mathbb{E}[g_i \mid \mathcal{F}]}{q_i(\mathcal{F})} = \frac{1}{M}\sum_{i=1}^{M} \mathbb{E}[g_i \mid \mathcal{F}].\end{aligned} \tag{36}$$

All g_i and all inclusion probabilities in this identity are conditioned on the same candidate information. Averaging over candidate generation and continuation randomness at a fixed node recovers the ordinary rollout expectation. Averaging over μ then gives

$$\begin{aligned}\mathbb{E}[\widehat{g}_{\text{ECE}}(v)] &= \mathbb{E}_{v\sim\mu}\mathbb{E}_{\tau\sim p_{\theta_{\text{old}}}(\cdot|v)}\left[G(\tau)\,\nabla_\theta \log p_\theta(\tau \mid v)|_{\theta=\theta_{\text{old}}}\right] \\ &= \nabla_\theta \mathbb{E}_{v\sim\mu}\mathbb{E}_{\tau\sim p_\theta(\cdot|v)}[G(\tau)]\big|_{\theta=\theta_{\text{old}}}.\end{aligned} \tag{37}$$

The prefix-generating reference policy contributes no score term because μ is fixed.

The proof applies to each update conditional on its newly frozen parameters. It requires a fresh batch, one update, fixed class-size weights, and normalization by the original candidate count. A misspecified effect key can merge unequal effects while leaving the inclusion-weighted estimator unbiased for the resulting partition. Incorrect restoration changes the continuation law and can invalidate Eq. (36).

D.1.2 VARIANCE AND THE ROLE OF GROUPING

The precision cost can be stated for any fixed direction $a \in \mathbb{R}^{\dim(\theta)}$. Conditional on $\mathcal{F}$, write $X_i = a^\top g_i$, $m_i = \mathbb{E}[X_i \mid \mathcal{F}]$, and $\sigma_i^2 = \text{Var}(X_i \mid \mathcal{F})$. For class c, define $\bar{m}_c = n_c^{-1}\sum_{i\in\mathcal{C}_c} m_i$ and $v_c = n_c^{-1}\sum_{i\in\mathcal{C}_c}(m_i - \bar{m}_c)^2$. The law of total variance yields $\text{Var}(X_{I_c} \mid \mathcal{F}) = n_c^{-1}\sum_{i\in\mathcal{C}_c}\sigma_i^2 + v_c$. Hence

$$\text{Var}\big(a^\top \widehat{g}_{\text{ECE}} \mid \mathcal{F}\big) = \frac{1}{M^2}\sum_{c=1}^{K}\left(n_c \sum_{i\in\mathcal{C}_c}\sigma_i^2 + n_c^2 v_c\right). \tag{38}$$

For the full estimator $\widehat{g}_{\text{full}} = M^{-1}\sum_{i=1}^{M} g_i$, independent continuations give $\text{Var}(a^\top \widehat{g}_{\text{full}} \mid \mathcal{F}) = M^{-2}\sum_i \sigma_i^2$. Therefore,

$$\begin{aligned}&\text{Var}(a^\top \widehat{g}_{\text{ECE}} \mid \mathcal{F}) - \text{Var}(a^\top \widehat{g}_{\text{full}} \mid \mathcal{F}) \\ &\qquad = \frac{1}{M^2}\sum_{c=1}^{K}\left((n_c - 1)\sum_{i\in\mathcal{C}_c}\sigma_i^2 + n_c^2 v_c\right) \geq 0.\end{aligned} \tag{39}$$

The first term reflects fewer independent continuation samples; the second reflects differences between the conditional means of candidate histories. Even identical continuation means do not eliminate the first term. For two deterministic contributions $X_1 = -1$ and $X_2 = 1$ in one class, the full mean is always zero, while the ECE estimate is equally likely to be -1 and 1 and has variance one.

Effect equality alone imposes no bound on either term. Its utility depends on the interaction savings, the variation within the resulting classes, and how the saved resources are used. Equation (39) compares estimators at the same node and candidate width. Training performance under a total resource cap also depends on the allocation of the saved budget. The comparisons in Appendix A.2.5 separate the effects of subsampling, grouping, and resource allocation.

D.2 Population Model, Variance, and Cost Identities

D.2.1 Population model and exact quantities

All results fix the policy θ and node distribution μ. Independent pools draw nodes iid from μ; conditional on a node, candidates are iid from $\pi_\theta(\cdot \mid h)$. For one candidate, write $v \sim \mu$ and $U \sim \pi_\theta(\cdot \mid h)$. Let Z be its exact one-step effect and let ω denote independent continuation randomness. Write

$$X = X(v, U, \omega) = G(\tau)\nabla_\theta \log p_\theta(\tau \mid v), \qquad g_\star = \nabla_\theta J_{\mu,H}(\theta). \tag{40}$$

We assume finite second moments, parameter-independent policy support, and domination sufficient to differentiate the rollout expectation. All continuations use the same frozen policy and preserve their original histories. Conditional on a node, candidates and their potential continuation streams are independent. Conditional on the complete candidate information, representative indices are mutually independent across occupied classes and uniform within each class. Their selection randomness is independent of all continuation streams. The effect support $\mathcal{Z}(v)$ is finite; only positive-probability effects are included. These assumptions also allow different histories within an effect class to have different values and gradients.

For $e \in \mathcal{Z}(v)$, define

$$\begin{aligned} &p_e(v) = \Pr(Z = e \mid v), \quad m_e(v) = \mathbb{E}[X \mid v, Z = e], \\ &m(v) = \mathbb{E}[X \mid v], \qquad\qquad w_e(v) = \mathbb{E}[\|X - m_e(v)\|^2 \mid v, Z = e]. \end{aligned} \tag{41}$$

All norms are Euclidean. Define the node, between-effect, within-effect, and collision-weighted within-effect quantities by

$$\begin{aligned} T &= \mathbb{E}_v \|m(v) - g_\star\|^2, \\ A &= \mathbb{E}_v \sum_e p_e(v)\|m_e(v) - m(v)\|^2, \\ W &= \mathbb{E}_v \sum_e p_e(v) w_e(v), \qquad Q = \mathbb{E}_v \sum_e p_e(v)^2 w_e(v). \end{aligned} \tag{42}$$

In particular, $w_e(v)$ includes both history heterogeneity and continuation noise; it need not vanish when effects agree.

For a direct-sampling width $w \geq 1$ and an ECE width $M \geq 1$, the single-node estimators are

$$\widehat{g}_{\mathrm{D},w} = \frac{1}{w}\sum_{i=1}^{w} X_i, \qquad \widehat{g}_{\mathrm{E},M} = \frac{1}{M}\sum_{e:N_e>0} N_e X_{I_e}, \tag{43}$$

where N_e is the number of ECE candidates with effect e, and I_e is a uniform index among these candidates. Different nodes use independent copies of the entire collection procedure.

Let $\Lambda \in \{0, \ldots, H-1\}$ be a potential continuation length, $\ell_e(v) = \mathbb{E}[\Lambda \mid v, Z = e]$, and $\bar{\ell} = \mathbb{E}_v \sum_e p_e(v)\ell_e(v)$. Costs count environment interactions, including every initial probe. They need not be independent of gradient contributions.

Theorem 2 (Exact expected-budget risk)**.** *Let $\widehat{g}_{\mathrm{D},w}$ average all w candidate gradients at one node, and let $\widehat{g}_{\mathrm{E},M}$ be the width-M ECE estimator in Eq. (7). Both are unbiased for $g_\star$. Their exact single-pool squared-error risks are $V_{\mathrm{D}}(w) = T + (A+W)/w$ and $V_{\mathrm{E}}(M) = T + (A+W)/M +$*

$(M-1)Q/M$. *Let c_a be collector a's expected interaction cost, given exactly in Eq.* (45)*. For fixed widths and $B \geq \max_a c_a$, averaging the deterministic number $n_a(B) = \lfloor B/c_a \rfloor$ of independent complete pools has exact risk $\mathcal{R}_a(B) = V_a/\lfloor B/c_a \rfloor$. Consequently, $B\mathcal{R}_a(B) \to c_a V_a$. For $\varepsilon > 0$, among positive deterministic pool counts, the minimum expected cost for risk at most ε^2 is $c_a \max\{1, \lceil V_a/\varepsilon^2 \rceil\}$.*

For reference, the exact single-pool risks and expected costs in Theorem 2 are

$$\begin{aligned} V_{\mathrm{D}}(w) &:= \mathbb{E}\|\widehat{g}_{\mathrm{D},w} - g_\star\|^2 = T + \frac{A+W}{w}, \\ V_{\mathrm{E}}(M) &:= \mathbb{E}\|\widehat{g}_{\mathrm{E},M} - g_\star\|^2 = T + \frac{A+W}{M} + \frac{M-1}{M}Q. \end{aligned} \tag{44}$$

$$\begin{aligned} c_{\mathrm{D}}(w) &= w(1+\bar{\ell}), \\ c_{\mathrm{E}}(M) &= M + \mathbb{E}_v \sum_e \big[1 - (1-p_e(v))^M\big]\ell_e(v). \end{aligned} \tag{45}$$

For $a \in \{\mathrm{D}, \mathrm{E}\}$, fix the collector width and write V_a, c_a for these quantities. Under an expected budget $B \geq \max_a c_a$, the count $n_a(B) = \lfloor B/c_a \rfloor$ is deterministic. The exact finite-budget comparison and accuracy cost are

$$\mathcal{R}_a(B) := \mathbb{E}\left\| \frac{1}{n_a(B)} \sum_{j=1}^{n_a(B)} \widehat{g}_{a,j} - g_\star \right\|^2 = \frac{V_a}{\lfloor B/c_a \rfloor}, \tag{46}$$

$$\mathcal{R}_{\mathrm{E}}(B) < \mathcal{R}_{\mathrm{D}}(B) \quad \Longleftrightarrow \quad V_{\mathrm{E}}\lfloor B/c_{\mathrm{D}} \rfloor < V_{\mathrm{D}}\lfloor B/c_{\mathrm{E}} \rfloor.$$

$$\mathcal{B}_a(\varepsilon) = c_a \max\left\{1, \left\lceil \frac{V_a}{\varepsilon^2} \right\rceil \right\}, \qquad \varepsilon > 0. \tag{47}$$

Here $\varepsilon > 0$, and the accuracy cost is minimized only over positive deterministic counts of the stated collector. For $B > c_a$, a convenient rounding envelope is

$$\frac{c_a V_a}{B} \leq \mathcal{R}_a(B) \leq \frac{c_a V_a}{B - c_a}, \qquad \lim_{B\to\infty} B\mathcal{R}_a(B) = c_a V_a. \tag{48}$$

The equality for the leading constant is exact. If $\Gamma = c_{\mathrm{D}}V_{\mathrm{D}} - c_{\mathrm{E}}V_{\mathrm{E}} > 0$, a sufficient, generally conservative finite-budget threshold is

$$B > \max\left\{c_{\mathrm{D}}, c_{\mathrm{E}}, \frac{c_{\mathrm{E}}c_{\mathrm{D}}V_{\mathrm{D}}}{\Gamma}\right\}. \tag{49}$$

The floor comparison in Eq. (46) gives the exact criterion; the threshold above is a simpler sufficient bound. For common deterministic continuation length $L > 0$, put $\kappa = \mathbb{E}K_M$. For $M > 1$, the necessary-and-sufficient leading-constant condition in Corollary 1 expands to

$$\begin{aligned} L(M-\kappa)(A+W) > M\big[M - 1 + L(\kappa-1)\big]T \\ + (M + L\kappa)(M-1)Q. \end{aligned} \tag{50}$$

D.2.2 Conditional orthogonality and the target gradient

All expectations in this appendix are evaluated at the fixed parameter θ. Unexecuted continuations are represented as potential outcomes on an enlarged probability space. The following orthogonality identity underlies the variance calculations.

Lemma 1 (Conditional Pythagorean identity)**.** *Let $X \in \mathbb{R}^d$ be square-integrable, let $\mathcal{H}$ be a σ-field, and let b be an $\mathcal{H}$-measurable, square-integrable vector. Then*

$$\mathbb{E}\|X - b\|^2 = \mathbb{E}\|X - \mathbb{E}[X \mid \mathcal{H}]\|^2 + \mathbb{E}\|\mathbb{E}[X \mid \mathcal{H}] - b\|^2. \tag{51}$$

In particular, conditional expectation minimizes squared error over all $\mathcal{H}$-measurable predictors.

Proof. Set $m_{\mathcal{H}} = \mathbb{E}[X \mid \mathcal{H}]$. Conditional Jensen and Cauchy–Schwarz ensure that all terms below are integrable:

$$\begin{aligned} \mathbb{E}\|m_{\mathcal{H}}\|^2 \le \mathbb{E}\mathbb{E}[\|X\|^2 \mid \mathcal{H}] = \mathbb{E}\|X\|^2 < \infty, \\ \mathbb{E}|\langle X - m_{\mathcal{H}}, m_{\mathcal{H}} - b\rangle| \le \big(\mathbb{E}\|X - m_{\mathcal{H}}\|^2\big)^{1/2}\big(\mathbb{E}\|m_{\mathcal{H}} - b\|^2\big)^{1/2} < \infty. \end{aligned} \tag{52}$$

Measurability and the tower property give the exact cancellation

$$\mathbb{E}\langle X - m_{\mathcal{H}}, m_{\mathcal{H}} - b\rangle = \mathbb{E}\,\langle \mathbb{E}[X - m_{\mathcal{H}} \mid \mathcal{H}], m_{\mathcal{H}} - b\rangle = 0. \tag{53}$$

Expanding the squared norm of $(X - m_{\mathcal{H}}) + (m_{\mathcal{H}} - b)$ proves Eq. (51). The minimizing property follows from the nonnegativity of its second term. □

The score-function identity is justified by the differentiability and domination assumptions in Appendix D.2.1. Since μ, the environment, and the return function do not depend on the differentiated parameter, we have

$$\begin{aligned} \mathbb{E}X &= \int \mu(dv) \int G(\tau)\nabla_\theta \log p_\theta(\tau \mid v) p_\theta(\tau \mid v)\, d\tau \\ &= \int \mu(dv) \int G(\tau)\nabla_\theta p_\theta(\tau \mid v)\, d\tau \\ &= \nabla_\theta \int \mu(dv) \int G(\tau) p_\theta(\tau \mid v)\, d\tau = g_\star. \end{aligned} \tag{54}$$

Here integration includes summation for discrete trajectories. The likelihood includes all policy-generated tokens, while the fixed prefix and environment observations contribute no score terms.

To separate the two sources of within-effect uncertainty, write $b(v, U) = \mathbb{E}[X \mid v, U]$. Since Z is determined by (v, U), conditional application of Lemma 1 gives

$$\begin{aligned} w_e(v) = \mathbb{E}[\|b(v,U) - m_e(v)\|^2 \mid v, Z = e] \\ + \mathbb{E}[\|X - b(v,U)\|^2 \mid v, Z = e]. \end{aligned} \tag{55}$$

The first term measures variation across original candidate histories; the second measures randomness in their continuations. Neither is controlled solely by equality of the one-step effect.

D.2.3 Exchangeability within observed effect classes

Lemma 2 (Conditioning on effect labels)**.** *Fix v and condition on the ordered labels $Z_1, \dots, Z_M$. Let $\mathcal{H}_M = \sigma(v, Z_1, \dots, Z_M)$. The selected gradients in different occupied classes are conditionally independent, and a representative of class e has conditional mean $m_e(v)$ and conditional squared-error variance $w_e(v)$. Consequently,*

$$\begin{aligned} \mathbb{E}[\widehat{g}_{\mathrm{E},M} \mid \mathcal{H}_M] = \frac{1}{M}\sum_e N_e m_e(v), \\ \mathbb{E}\Big[\|\widehat{g}_{\mathrm{E},M} - \mathbb{E}[\widehat{g}_{\mathrm{E},M} \mid \mathcal{H}_M]\|^2 \Big| \mathcal{H}_M\Big] = \frac{1}{M^2}\sum_e N_e^2 w_e(v). \end{aligned} \tag{56}$$

Proof. Conditional on v, the pairs (Z_i, X_i) are independent copies. For measurable sets $D_1, \dots, D_M$ and a positive-probability label configuration $z_1, \dots, z_M$, independence yields

$$\begin{aligned} &\Pr(X_i \in D_i \text{ for all } i \mid v, Z_i = z_i \text{ for all } i) \\ &= \frac{\prod_{i=1}^M \Pr(X_i \in D_i, Z_i = z_i \mid v)}{\prod_{i=1}^M \Pr(Z_i = z_i \mid v)} \\ &= \prod_{i=1}^M \Pr(X_i \in D_i \mid v, Z_i = z_i). \end{aligned} \tag{57}$$

For a general node distribution the calculation is understood through regular conditional laws, for μ-almost every v. Thus gradients in an occupied class are iid under their effect- conditional law, even though their individual realized histories differ.

Uniform selection remains independent of the within-class gradients conditional on the labels. Indeed, conditional on the full candidate information its probability is the same $1/N_e$ for every class member, and selection is independent of continuation streams. Therefore, for bounded measurable functions f_e,

$$\begin{aligned}
&\mathbb{E}\left[\prod_{e:N_e>0} f_e(X_{I_e})\middle|\mathcal{H}_M\right] \\
&= \sum_{(i_e):Z_{i_e}=e} \left(\prod_{e:N_e>0}\frac{1}{N_e}\right)\mathbb{E}\left[\prod_{e:N_e>0} f_e(X_{i_e})\middle|\mathcal{H}_M\right] \\
&= \sum_{(i_e):Z_{i_e}=e}\prod_{e:N_e>0}\frac{\mathbb{E}[f_e(X)\mid v, Z=e]}{N_e} \\
&= \prod_{e:N_e>0}\mathbb{E}[f_e(X)\mid v, Z=e].
\end{aligned} \tag{58}$$

This proves conditional independence and identifies each marginal law. Let $\eta_e = X_{I_e} - m_e(v)$. Their conditional means vanish, so

$$\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1}{M}\sum_e N_e\eta_e\right\|^2\middle|\mathcal{H}_M\right] \\
&= \frac{1}{M^2}\sum_e N_e^2\mathbb{E}[\|\eta_e\|^2\mid\mathcal{H}_M] + \frac{2}{M^2}\sum_{e<f} N_eN_f\mathbb{E}[\langle\eta_e,\eta_f\rangle\mid\mathcal{H}_M] \\
&= \frac{1}{M^2}\sum_e N_e^2 w_e(v).
\end{aligned} \tag{59}$$

This proves Eq. (56). □

D.2.4 Proof of the exact variance and cost identities

We prove the first part of Theorem 2. At a fixed node abbreviate $A(v) = \sum_e p_e(v)\|m_e(v) - m(v)\|^2$, $W(v) = \sum_e p_e(v)w_e(v)$, and $Q(v) = \sum_e p_e(v)^2 w_e(v)$. The common conditional mean in Lemma 2 is the mean of the iid vectors $m_{Z_i}(v)$. Hence

$$\begin{aligned}
\mathbb{E}[\widehat{g}_{\mathrm{E},M}\mid v] &= \frac{1}{M}\sum_{i=1}^M \mathbb{E}[m_{Z_i}(v)\mid v] = m(v), \\
\mathbb{E}\left[\left\|\frac{1}{M}\sum_{i=1}^M m_{Z_i}(v) - m(v)\right\|^2\middle|v\right] &= \frac{1}{M^2}\sum_{i=1}^M \mathbb{E}[\|m_{Z_i}(v)-m(v)\|^2\mid v] = \frac{A(v)}{M}.
\end{aligned} \tag{60}$$

Writing $N_e = \sum_i \mathbf{1}\{Z_i = e\}$ and separating diagonal and off-diagonal terms gives the exact occupancy moments

$$\begin{aligned}
\mathbb{E}[N_e\mid v] &= Mp_e(v), \\
\mathbb{E}[N_e^2\mid v] &= \sum_{i=1}^M p_e(v) + \sum_{i\neq j} p_e(v)^2 = Mp_e(v) + M(M-1)p_e(v)^2.
\end{aligned} \tag{61}$$

Apply conditional orthogonality first to $\mathcal{H}_M$ and then to the node. Equations (56)– (61) imply

$$\begin{aligned}\mathbb{E}[\|\widehat{g}_{\mathrm{E},M} - m(v)\|^2 \mid v] &= \frac{A(v)}{M} + \frac{1}{M^2}\sum_e \mathbb{E}[N_e^2 \mid v] w_e(v) \\ &= \frac{A(v)}{M} + \sum_e \left(\frac{p_e(v)}{M} + \frac{M-1}{M} p_e(v)^2\right) w_e(v) \\ &= \frac{A(v)+W(v)}{M} + \frac{M-1}{M} Q(v), \\ V_{\mathrm{E}}(M) &= \mathbb{E}\|m(v) - g_\star\|^2 + \mathbb{E}_v \mathbb{E}[\|\widehat{g}_{\mathrm{E},M} - m(v)\|^2 \mid v] \\ &= T + \frac{A+W}{M} + \frac{M-1}{M} Q.\end{aligned} \tag{62}$$

For direct sampling, conditional independence and one further application of Lemma 1 give

$$\begin{aligned}\mathbb{E}[\|X - m(v)\|^2 \mid v] &= A(v) + W(v), \\ \mathbb{E}[\widehat{g}_{\mathrm{D},w} \mid v] &= m(v), \\ V_{\mathrm{D}}(w) &= T + \frac{1}{w^2}\mathbb{E}_v \sum_{i=1}^{w} \mathbb{E}[\|X_i - m(v)\|^2 \mid v] = T + \frac{A+W}{w}.\end{aligned} \tag{63}$$

Equation (54) and the conditional means prove unbiasedness. The exact variance difference satisfies

$$\begin{aligned}0 \le Q = \mathbb{E}_v \sum_e p_e(v)^2 w_e(v) &\le \mathbb{E}_v \sum_e p_e(v) w_e(v) = W, \\ 0 \le V_{\mathrm{E}}(M) - V_{\mathrm{D}}(M) = \frac{M-1}{M} Q &\le \frac{M-1}{M} W.\end{aligned} \tag{64}$$

For $M > 1$, the variance difference is zero exactly when $w_e(v) = 0$ for every positive-probability effect, for μ-almost every node. The upper endpoint is attained when each node has one effect class.

For costs, attach the pair (X_i, Λ_i) to each potential continuation. The factorization in Eq. (57) applies to these pairs, allowing dependence between X_i and Λ_i. Thus, on $\{N_e > 0\}$, $\mathbb{E}[\Lambda_{I_e} \mid \mathcal{H}_M] = \ell_e(v)$. Consequently,

$$\begin{aligned}\mathbb{E}[C_{\mathrm{E}} \mid v] &= M + \mathbb{E}\left[\sum_{e:N_e>0} \Lambda_{I_e} \,\middle|\, v\right] \\ &= M + \sum_e \Pr(N_e > 0 \mid v)\ell_e(v) \\ &= M + \sum_e [1 - (1 - p_e(v))^M]\ell_e(v), \\ \mathbb{E}[C_{\mathrm{D},w} \mid v] &= w + \sum_{i=1}^{w} \mathbb{E}[\Lambda_i \mid v] = w + w\sum_e p_e(v)\ell_e(v).\end{aligned} \tag{65}$$

Averaging over v proves Eq. (45). Class occupancy also gives

$$\begin{aligned}0 \le 1 - (1-p)^M = p\sum_{r=0}^{M-1}(1-p)^r \le Mp, \qquad 0 \le p \le 1, \\ M \le c_{\mathrm{E}}(M) \le M + M\mathbb{E}_v \sum_e p_e(v)\ell_e(v) = c_{\mathrm{D}}(M) \le MH.\end{aligned} \tag{66}$$

The risk comparison additionally depends on the variance cost of representative sampling.

D.3 Expected-Budget Efficiency and Grouping Comparisons

D.3.1 Proof of the finite expected-budget statements

For either fixed collector let $\xi_j = \widehat{g}_{a,j} - g_\star$. The vectors ξ_j are iid, mean zero, and have second moment V_a. For a deterministic $n \geq 1$,

$$\begin{aligned}\mathbb{E}\left\|\frac{1}{n}\sum_{j=1}^{n}\xi_j\right\|^2 &= \frac{1}{n^2}\sum_{j=1}^{n}\mathbb{E}\|\xi_j\|^2 + \frac{2}{n^2}\sum_{i<j}\mathbb{E}\langle\xi_i,\xi_j\rangle \\ &= \frac{nV_a}{n^2} + \frac{2}{n^2}\sum_{i<j}\langle\mathbb{E}\xi_i,\mathbb{E}\xi_j\rangle = \frac{V_a}{n}.\end{aligned} \tag{67}$$

Linearity of expectation gives total expected cost nc_a, regardless of dependence between a pool's estimate and its cost. Therefore the largest admissible deterministic pool count is $n_a(B) = \lfloor B/c_a \rfloor$, proving Eq. (46). Cross-multiplication is valid because both counts are positive, and yields its exact comparison.

Similarly, $V_a/n \leq \varepsilon^2$ is equivalent to $n \geq V_a/\varepsilon^2$. Restricting to positive integers proves Eq. (47), including the case $V_a = 0$. For $B > c_a$, elementary rounding gives

$$\begin{aligned}\frac{B}{c_a} - 1 < \left\lfloor \frac{B}{c_a} \right\rfloor \leq \frac{B}{c_a}, \\ \frac{c_a}{B} \leq \frac{1}{\lfloor B/c_a \rfloor} \leq \frac{c_a}{B - c_a}, \\ c_a V_a \leq B\mathcal{R}_a(B) \leq \frac{B}{B - c_a} c_a V_a \longrightarrow c_a V_a.\end{aligned} \tag{68}$$

This proves Eq. (48). If $\Gamma > 0$ and Eq. (49) holds, then

$$\begin{aligned}B\Gamma > c_{\mathrm{E}} c_{\mathrm{D}} V_{\mathrm{D}}, \\ B c_{\mathrm{E}} V_{\mathrm{E}} < (B - c_{\mathrm{E}}) c_{\mathrm{D}} V_{\mathrm{D}}, \\ \mathcal{R}_{\mathrm{E}}(B) \leq \frac{c_{\mathrm{E}} V_{\mathrm{E}}}{B - c_{\mathrm{E}}} < \frac{c_{\mathrm{D}} V_{\mathrm{D}}}{B} \leq \mathcal{R}_{\mathrm{D}}(B).\end{aligned} \tag{69}$$

If $\Gamma < 0$, the limit in Eq. (48) implies the opposite ordering for all sufficiently large budgets. If $\Gamma = 0$, leading constants agree and the exact finite-budget comparison retains the floor terms. This completes the proof of Theorem 2.

D.3.2 Sharp comparisons, equality cases, and attainable examples

Corollary 1 (Comparison with the best fixed-width direct sampler)**.** *Let $\bar{\ell}$ be the mean continuation length of one ordinary candidate. Among all fixed direct-sampling widths, width one minimizes the leading interaction-budget constant, which equals $(1+\bar{\ell})(T+A+W)$. Thus ECE improves on this entire direct-sampling family in leading constant exactly when $c_{\mathrm{E}}(M)V_{\mathrm{E}}(M) < (1+\bar{\ell})(T+A+W)$. For common continuation length L, the fully expanded condition is Eq.* (50)*.*

Proof of Corollary 1. For arbitrary bounded continuation lengths, direct sampling satisfies

$$\begin{aligned}c_{\mathrm{D}}(w)V_{\mathrm{D}}(w) &= w(1+\bar{\ell})\left(T + \frac{A+W}{w}\right) \\ &= (1+\bar{\ell})(wT + A + W) \\ &\geq (1+\bar{\ell})(T + A + W) = c_{\mathrm{D}}(1)V_{\mathrm{D}}(1).\end{aligned} \tag{70}$$

Equality holds for all widths when $T = 0$, and otherwise only for $w = 1$. Let $\Gamma_1 = c_{\mathrm{D}}(1)V_{\mathrm{D}}(1) - c_{\mathrm{E}}(M)V_{\mathrm{E}}(M)$. Specialize now to a common deterministic continuation length $L > 0$, so $c_{\mathrm{E}}(M) =$

$M + L\kappa$. Keeping every term in the expansion gives

$$\begin{aligned} M\Gamma_1 &= M(1+L)(T+A+W) \\ &\quad - (M+L\kappa)\big[MT + A + W + (M-1)Q\big] \\ &= \big[M(1+L) - M(M+L\kappa)\big]T \\ &\quad + \big[M(1+L) - (M+L\kappa)\big](A+W) \\ &\quad - (M+L\kappa)(M-1)Q \\ &= L(M-\kappa)(A+W) \\ &\quad - M\big[M - 1 + L(\kappa-1)\big]T - (M+L\kappa)(M-1)Q. \end{aligned} \tag{71}$$

Thus $\Gamma_1 > 0$ is equivalent to Eq. (50). If $W = 0$, then $Q = 0$ by Eq. (64); the noiseless special case $T = W = 0$, $A > 0$, $\kappa < M$ also follows. This minimizes the population leading constant over fixed direct widths. Finite-budget comparisons retain the integer-rounding terms.

At the same width M, the analogous exact expansion is

$$\begin{aligned} &M\big[c_{\mathrm{D}}(M)V_{\mathrm{D}}(M) - c_{\mathrm{E}}(M)V_{\mathrm{E}}(M)\big] \\ &\quad = L(M-\kappa)(MT + A + W) - (M+L\kappa)(M-1)Q. \end{aligned} \tag{72}$$

In the one-effect case, $\kappa = 1$, $A = 0$, and $Q = W$, so for $M > 1$ the same-width improvement threshold reduces exactly to

$$\begin{aligned} c_{\mathrm{D}}(M)V_{\mathrm{D}}(M) - c_{\mathrm{E}}(M)V_{\mathrm{E}}(M) &= (M-1)(LT - W), \\ c_{\mathrm{E}}(M)V_{\mathrm{E}}(M) - c_{\mathrm{D}}(1)V_{\mathrm{D}}(1) &= (M-1)(T+W) \geq 0. \end{aligned} \tag{73}$$

The second identity separates improvement at a matched width from improvement over the best fixed-width direct sampler.

Proposition 1 (Compression alone is insufficient)**.** *Suppose the node is fixed, all effect-conditional gradient means agree, and all effects have the same within-effect variance $\sigma^2 > 0$. Assume a common continuation length $L \geq 0$ and $M > 1$. With $s_2 = \sum_e p_e^2$, ECE satisfies*

$$\frac{c_{\mathrm{E}}(M)V_{\mathrm{E}}(M)}{c_{\mathrm{D}}(M)V_{\mathrm{D}}(M)} \geq 1 + \frac{(M-1)s_2}{1+L} > 1. \tag{74}$$

The first inequality is an equality for one effect class.

Proof. Let $q_e = \Pr(N_e > 0) = 1 - (1-p_e)^M$. Because $N_e = N_e \mathbf{1}\{N_e > 0\}$, Cauchy–Schwarz yields

$$\begin{aligned} (Mp_e)^2 &= \big(\mathbb{E}[N_e \mathbf{1}\{N_e > 0\}]\big)^2 \\ &\leq \mathbb{E}[N_e^2]\,\mathbb{E}[\mathbf{1}\{N_e > 0\}^2] \\ &= \big[Mp_e + M(M-1)p_e^2\big]q_e. \end{aligned} \tag{75}$$

A second, weighted Cauchy–Schwarz inequality gives

$$\begin{aligned} \kappa = \sum_e q_e &\geq M \sum_e \frac{p_e}{1 + (M-1)p_e} \\ &\geq M \frac{(\sum_e p_e)^2}{\sum_e p_e[1 + (M-1)p_e]} \\ &= \frac{M}{1 + (M-1)s_2}. \end{aligned} \tag{76}$$

Here $T = A = 0$, $W = \sigma^2$, and $Q = s_2\sigma^2$. Substitution into the exact identities produces

$$\begin{aligned} \frac{c_{\mathrm{E}}V_{\mathrm{E}}}{c_{\mathrm{D}}V_{\mathrm{D}}} &= \frac{(M+L\kappa)[1 + (M-1)s_2]}{M(1+L)} \\ &\geq \frac{M[1 + (M-1)s_2] + LM}{M(1+L)} \\ &= 1 + \frac{(M-1)s_2}{1+L} > 1. \end{aligned} \tag{77}$$

For a single effect, $N_1 = M$, $s_2 = \kappa = 1$, and both Cauchy–Schwarz steps are equalities. The homoscedastic assumption is essential: heteroscedastic effects can have a much smaller collision-weighted variance Q than this example permits. □

Attainability by a policy-gradient model. A finite deterministic environment realizes the preceding phase boundaries. Fix an integer $L \geq 1$ and let the node encode $S \in \{-1, 1\}$ with equal probabilities. The first, parameter-independent turn selects an effect $Z \in \{1, \ldots, J\}$ uniformly. The environment records Z, so these are exact effect classes. After $L-1$ parameter-independent filler actions, the policy samples $Y \in \{-1, 1\}$ with $\Pr_\theta(Y = 1) = (1 + e^{-\theta})^{-1}$. All rewards are zero except the terminal reward $G = 2Y\nu(S, Z) + 2\sigma$, where $\nu(S, Z) = m_0 + \alpha S + \beta f(Z)$, $m_0 \neq 0$, $\sigma \geq 0$, and f has uniform mean zero and variance one when $J \geq 2$. For $J = 1$, set $\beta = f(1) = 0$. All variables are bounded and the policy has smooth, parameter-independent support. At $\theta = 0$,

$$\begin{aligned} \nabla_\theta \log \Pr_\theta(Y)\big|_{\theta=0} &= Y/2, \\ X = G(Y/2) &= m_0 + \alpha S + \beta f(Z) + \sigma Y, \qquad g_\star = m_0, \\ T &= \alpha^2, \qquad A = \beta^2, \qquad W = \sigma^2, \qquad Q = \sigma^2/J, \\ \kappa &= J\big[1 - (1 - 1/J)^M\big]. \end{aligned} \tag{78}$$

Thus $J = 1$ realizes both sides and the equality case of $LT = W$. Taking $\alpha = \sigma = 0$, $\beta \neq 0$, and $J \geq 2$ realizes $T = W = 0$, $A > 0$, and $\kappa < M$ for $M > 1$, which satisfies the stronger comparison in Corollary 1. The construction uses bounded, possibly signed rewards, as allowed by the theorem. This example establishes attainability of the moment identities and phase conditions.

D.3.3 Why effect grouping can improve on matched-size random grouping

Proposition 2 (Effect grouping versus matched-size random grouping)**.** *Suppose every node has J equiprobable effects and $M \geq 2$. Apply the random-partition control in Appendix A.2.5 to the same candidate pool: uniformly permute its indices, partition into blocks with the observed ECE class sizes, and select one uniform representative per block with the same size weights. Its single-pool risk $V_{\mathrm{R}}(M)$ satisfies the exact identity $V_{\mathrm{R}}(M) - V_{\mathrm{E}}(M) = (M-2)(M+1)A/(M^2J)$. With a common deterministic continuation length, the two collectors also have the same realized pool cost. Thus ECE has smaller expected-budget MSE at every common positive pool count whenever $M \geq 3$ and $A > 0$.*

Proof of Proposition 2. The random-partition control uses exactly the observed ECE class sizes, as specified in Appendix A.2.5. We first analyze this control conditionally on the original node, candidates, effect labels, and all potential full-rollout contributions $X_1, \ldots, X_M$. These potential contributions retain their original histories and continuation streams; the random partition does not change them. Write $n_1, \ldots, n_K$ for the positive observed class sizes, and set

$$\bar{X} = \frac{1}{M}\sum_{i=1}^{M} X_i, \qquad S = \sum_{i=1}^{M} \|X_i - \bar{X}\|^2, \qquad D = \sum_{c=1}^{K} n_c^2 - M. \tag{79}$$

The remaining randomness consists of a uniform permutation of the candidate indices and independent uniform representative positions within its blocks of sizes $n_1, \ldots, n_K$.

Let B_c be the positions in block c, and let I_c be its selected candidate index. For any distinct indices $i_1, \ldots, i_K$, averaging over the representative positions gives

$$\begin{aligned} \Pr(I_1 = i_1, \ldots, I_K = i_K \mid X_{1:M}, Z_{1:M}, v) &= \sum_{s_1 \in B_1, \ldots, s_K \in B_K} \left(\prod_{c=1}^{K} \frac{1}{n_c}\right) \frac{(M-K)!}{M!} \\ &= \frac{(M-K)!}{M!}. \end{aligned} \tag{80}$$

Thus the ordered representatives are a uniform sample without replacement of K distinct candidate indices. Put $x_i = X_i - \bar{X}$. Since $\sum_i x_i = 0$, their diagonal and off-diagonal moments satisfy

$$\begin{aligned} \mathbb{E}[x_{I_c} \mid X_{1:M}, Z_{1:M}, v] &= 0, \\ \mathbb{E}[\|x_{I_c}\|^2 \mid X_{1:M}, Z_{1:M}, v] &= S/M, \\ \mathbb{E}[\langle x_{I_c}, x_{I_d}\rangle \mid X_{1:M}, Z_{1:M}, v] &= \frac{1}{M(M-1)} \sum_{i \neq j} \langle x_i, x_j \rangle \\ &= -\frac{S}{M(M-1)}, \qquad c \neq d. \end{aligned} \tag{81}$$

The estimator $\widehat{g}_{\mathrm{R},M} = M^{-1}\sum_c n_c X_{I_c}$ therefore has conditional mean $\bar{X}$, and its conditional variance is exactly

$$\begin{aligned}
&\mathbb{E}[\|\widehat{g}_{\mathrm{R},M} - \bar{X}\|^2 \mid X_{1:M}, Z_{1:M}, v] \\
&= \frac{1}{M^2}\left\{\frac{S}{M}\sum_c n_c^2 - \frac{S}{M(M-1)}\sum_{c\neq d} n_c n_d\right\} \\
&= \frac{S}{M^3(M-1)}\left\{(M-1)\sum_c n_c^2 - \left(M^2 - \sum_c n_c^2\right)\right\} \\
&= \frac{DS}{M^2(M-1)}.
\end{aligned} \tag{82}$$

The negative off-diagonal term accounts for dependence between representatives sampled without replacement from the fixed pool.

We next average over the potential pool at a fixed node v. Write $A(v) = \sum_e p_e(v)\|m_e(v) - m(v)\|^2$ and $W(v) = \sum_e p_e(v) w_e(v)$. Under the uniform-effect assumption, $p_e(v) = 1/J$ for every one of the J effects, so $Q(v) = W(v)/J$; the within-effect variances themselves need not agree. Let $\mathcal{P}$ be the set of unordered pairs of distinct candidate indices and put $P = |\mathcal{P}| = \binom{M}{2}$. For $a = \{i, j\} \in \mathcal{P}$, define $I_a = \mathbf{1}\{Z_i = Z_j\}$ and $d_a = \|X_i - X_j\|^2$. The collision-count and pairwise-distance identities give

$$D = 2\sum_{a\in\mathcal{P}} I_a, \qquad S = \frac{1}{M}\sum_{a\in\mathcal{P}} d_a, \qquad \mathbb{E}[d_a \mid v] = 2\{A(v) + W(v)\}. \tag{83}$$

For identical pairs, conditional independence of the two contributions given their common effect implies

$$\begin{aligned}
\mathbb{E}[I_a d_a \mid v] &= \sum_{e=1}^{J} \frac{1}{J^2}\mathbb{E}[\|X_i - X_j\|^2 \mid v, Z_i = Z_j = e] \\
&= \frac{2}{J^2}\sum_{e=1}^{J} w_e(v) = \frac{2W(v)}{J}.
\end{aligned} \tag{84}$$

For distinct pairs $a \neq b$, at least one endpoint of a is not an endpoint of b. If the pairs share one endpoint, condition on the effect labels and contributions at the two endpoints of b; the unused endpoint of a still has an independent uniform effect label. The conditional collision probability is therefore $1/J$. If the pairs are disjoint, the collision event is independent of d_b and again has probability $1/J$. Both cases yield

$$\mathbb{E}[I_a d_b \mid v] = \frac{1}{J}\mathbb{E}[d_b \mid v] = \frac{2\{A(v) + W(v)\}}{J}, \qquad a \neq b. \tag{85}$$

There are P identical ordered pairs of pairs and $P(P-1)$ distinct ones. Summing both contributions gives

$$\begin{aligned}
\mathbb{E}[DS \mid v] &= \frac{2}{M}\sum_{a,b\in\mathcal{P}} \mathbb{E}[I_a d_b \mid v] \\
&= \frac{4P}{MJ}\left[W(v) + (P-1)\{A(v) + W(v)\}\right], \\
\frac{\mathbb{E}[DS \mid v]}{M^2(M-1)} &= \frac{M-1}{MJ}\{A(v) + W(v)\} - \frac{2A(v)}{M^2 J}.
\end{aligned} \tag{86}$$

Because the conditional mean is $\bar{X}$, orthogonality makes this quantity the extra variance above the full width-M direct estimator. Averaging over nodes and using $Q = W/J$ gives

$$\begin{aligned}
V_{\mathrm{R}}(M) &= T + \frac{A+W}{M} + \frac{M-1}{MJ}(A+W) - \frac{2A}{M^2 J}, \\
V_{\mathrm{E}}(M) &= T + \frac{A+W}{M} + \frac{M-1}{MJ}W, \\
V_{\mathrm{R}}(M) - V_{\mathrm{E}}(M) &= \left(\frac{M-1}{MJ} - \frac{2}{M^2 J}\right)A = \frac{(M-2)(M+1)}{M^2 J}A.
\end{aligned} \tag{87}$$

The gap is zero for $M = 2$ and is strictly positive for $M \geq 3$ whenever $A > 0$. For $M = 2$, the equality can also be seen pool by pool: the possible block-size lists are (2) and $(1, 1)$, neither of which leaves any grouping decision. Width one gives identical collectors trivially.

With a common deterministic continuation length L, both collectors incur the same realized pool cost $M + LK$, because the random control preserves K exactly. Hence their expected costs are identical, as are their deterministic complete-pool counts under an expected budget. For any common positive count n, their mean-squared-error difference is exactly the gap in Eq. (87) divided by n. The hard-budget asymptotic result multiplies that gap by their common expected cost. This cost equality uses the common-length assumption; variable continuation lengths can yield different costs despite matched block sizes. □

D.3.4 A STRICT IMPROVEMENT REGIME WITH POSITIVE NODE AND WITHIN-EFFECT VARIANCE

Proposition 3 (An attainable positive-variance improvement regime)**.** *Suppose there are exactly $J \geq 2$ equiprobable effects at each node, all continuations have deterministic length $L > 0$, and $M > 1$. Set $\kappa = J[1 - (1 - 1/J)^M]$ and define*

$$\begin{aligned} a_T &= \frac{M[M - 1 + L(\kappa - 1)]}{L(M - \kappa)}, \\ a_W &= \frac{(M + L\kappa)(M - 1)/J - L(M - \kappa)}{L(M - \kappa)}. \end{aligned} \tag{88}$$

Both coefficients are strictly positive. ECE has a strictly smaller asymptotic interaction-budget MSE constant than every fixed-width direct sampler if and only if

$$A > a_T T + a_W W. \tag{89}$$

For $M = 4$, $J = 2$, and $L = 8$, the exact condition becomes

$$17A > 40T + \frac{23}{2} W. \tag{90}$$

In particular, $T = W = 1$ and $A = 4$ give strictly positive variance components and a strict efficiency improvement. These moments are attainable by a bounded finite-horizon policy-gradient model.

Proof. Uniform effects imply $Q = W/J$ without a homoscedasticity assumption. The expected occupied-class count is the stated κ. Since $K \leq M$ always and all M candidates have the same effect with strictly positive probability, $\kappa < M$. Also $\kappa \geq 1$. Thus the denominator in Eq. (88) and the numerator of a_T are strictly positive.

To check a_W, let N_e denote the count of any one effect. Cauchy–Schwarz and the binomial second moment yield

$$\begin{aligned} (\mathbb{E}N_e)^2 &= \big(\mathbb{E}[N_e \mathbf{1}\{N_e > 0\}]\big)^2 \\ &\leq \mathbb{E}[N_e^2] \Pr(N_e > 0), \\ \Pr(N_e > 0) &\geq \frac{(M/J)^2}{M/J + M(M-1)/J^2} = \frac{M/J}{1 + (M-1)/J}, \\ \kappa \left(1 + \frac{M-1}{J}\right) &\geq M, \\ (M + L\kappa)\frac{M-1}{J} - L(M - \kappa) &= \frac{M(M-1)}{J} + L\left[\kappa\left(1 + \frac{M-1}{J}\right) - M\right] \\ &\geq \frac{M(M-1)}{J} > 0. \end{aligned} \tag{91}$$

Consequently $a_W > 0$.

The direct cost–variance product at width w is $(1+L)(wT+A+W)$, which is minimized at $w=1$. Let $\Gamma_1 = c_{\mathrm{D}}(1)V_{\mathrm{D}}(1) - c_{\mathrm{E}}(M)V_{\mathrm{E}}(M)$. Using $Q = W/J$ in the exact moment formulas gives the identity

$$\begin{aligned} M\Gamma_1 &= M(1+L)(T+A+W) \\ &\quad - (M+L\kappa)\left[MT + A + W + \frac{M-1}{J}W\right] \\ &= L(M-\kappa)A - M[M-1+L(\kappa-1)]T \\ &\quad - \left[\frac{(M+L\kappa)(M-1)}{J} - L(M-\kappa)\right]W \\ &= L(M-\kappa)\{A - a_T T - a_W W\}. \end{aligned} \tag{92}$$

This identity gives a necessary-and-sufficient threshold for the leading-constant comparison. For $M=4, J=2, L=8$, direct substitution gives $\kappa = 15/8$ and Eq. (90). At $T=W=1, A=4$, the relevant quantities are

$$\begin{aligned} c_{\mathrm{D}}(1)V_{\mathrm{D}}(1) &= 9(1+4+1) = 54, \\ c_{\mathrm{E}}(4) &= 4 + 8(15/8) = 19, \\ V_{\mathrm{E}}(4) &= 1 + \frac{4+1}{4} + \frac{3}{4}\frac{1}{2} = \frac{21}{8}, \\ c_{\mathrm{E}}(4)V_{\mathrm{E}}(4) &= \frac{399}{8} < 54, \\ \frac{c_{\mathrm{E}}(4)V_{\mathrm{E}}(4)}{c_{\mathrm{D}}(1)V_{\mathrm{D}}(1)} &= \frac{133}{144} < 1. \end{aligned} \tag{93}$$

The finite policy-gradient construction in Eq. (78), with $J=2$, $L=8$, $m_0=1$, $\alpha=\sigma=1$, and $\beta=2$, realizes exactly $T=W=1$ and $A=4$; the centered unit-variance effect function is Rademacher under the two equiprobable effects. Its rewards and score contributions are bounded, although rewards may be signed. The strict inequality persists on an open neighborhood of these positive moments within the uniform-effect model. □

The propositions give a structural, measurable regime in which the grouping criterion matters: separating effect-dependent conditional gradient means preserves information that a matched-size random partition can discard. The conditions on A and W are additional requirements within the simplified uniform-effect model; empirical applications require measuring these quantities.

D.4 Asymptotic Efficiency under a Hard Budget

D.4.1 Exact asymptotic risk under a hard interaction budget

Theorem 3 (Exact asymptotic risk under a hard budget)**.** *Fix a collector and let (Y_j, C_j) be iid complete-pool estimate–cost pairs with $\mathbb{E}Y_j = g_\star$, $\mathbb{E}\|Y_j - g_\star\|^2 = V < \infty$, and $0 < b_0 \le C_j \le R < \infty$ almost surely. Dependence between a pool's gradient estimate and cost is unrestricted. For $B \ge R$, repeatedly start and complete a pool whenever at least R interactions remain, and stop when fewer than R remain. Let N_B be the number completed and let $c = \mathbb{E}C_j$. The actual random-count average satisfies $B\,\mathbb{E}\|N_B^{-1}\sum_{j=1}^{N_B} Y_j - g_\star\|^2 \to cV$. Therefore, a strict ordering of cost–variance products gives the same ordering of actual hard-budget MSE for all sufficiently large budgets.*

Random sample sizes have a classical sequential-estimation theory (Anscombe, 1952). We prove the particular L^2 limit required by Theorem 3 directly. Fix a collector and an infinite iid sequence of potential estimate–cost pairs $(Y_i, C_i)_{i\ge1}$. The policy is frozen throughout this sequence. Assume $\mathbb{E}Y_i = g_\star$, $V = \mathbb{E}\|Y_i - g_\star\|^2 < \infty$, and $0 < b_0 \le C_i \le R < \infty$ almost surely. No independence between Y_i and C_i is imposed. Write $c = \mathbb{E}C_i$, $s^2 = \mathrm{Var}(C_i)$, $\xi_i = Y_i - g_\star$, $S_m = \sum_{i=1}^m \xi_i$, and $T_m = \sum_{i=1}^m C_i$, with $S_0 = 0$ and $T_0 = 0$.

For $B \ge R$, the collector starts and completes the next pool whenever at least R interactions remain. Its completed-pool count is therefore $N_B = \min\{m \ge 1 : T_m > B - R\}$. This definition includes

the pool that crosses the reservation threshold $B - R$; its cost never exceeds the remaining budget. Our objective is the actual random-denominator risk

$$\lim_{B\to\infty} B\,\mathbb{E}\left\|\frac{1}{N_B}\sum_{i=1}^{N_B} Y_i - g_\star\right\|^2 = cV. \tag{94}$$

This equality identifies the leading constant of the stopped collector under finite gradient second moments and bounded costs, allowing cost–gradient dependence.

D.4.2 Predictable isometry and the number of admitted pools

Lemma 3 (Predictable isometry)**.** *Let $\mathcal{F}_i = \sigma((Y_j, C_j) : 1 \le j \le i)$ and let $\mathcal{F}_0$ be trivial. Suppose Z_i is a square-integrable, mean-zero vector determined by (Y_i, C_i), with $\mathbb{E}\|Z_i\|^2 = v$, and the Z_i have the same distribution. For a deterministic integer K, let a_i be bounded real $\mathcal{F}_{i-1}$-measurable coefficients. Then*

$$\mathbb{E}\left\|\sum_{i=1}^{K} a_i Z_i\right\|^2 = v\sum_{i=1}^{K}\mathbb{E}a_i^2. \tag{95}$$

Consequently, for every deterministic $n \ge 1$,

$$\mathbb{E}\|S_{N_B} - S_n\|^2 = V\,\mathbb{E}|N_B - n|. \tag{96}$$

Proof. For $i < j$, both $a_i Z_i$ and a_j are $\mathcal{F}_{j-1}$-measurable. Independence and Cauchy–Schwarz give the integrable identities

$$\begin{aligned}
\mathbb{E}\langle a_i Z_i, a_j Z_j\rangle &= \mathbb{E}\left[a_j\,\langle a_i Z_i, \mathbb{E}[Z_j \mid \mathcal{F}_{j-1}]\rangle\right] = 0,\\
\mathbb{E}[a_i^2\|Z_i\|^2] &= \mathbb{E}\left[a_i^2\mathbb{E}[\|Z_i\|^2 \mid \mathcal{F}_{i-1}]\right] = v\mathbb{E}a_i^2,\\
\mathbb{E}\left\|\sum_{i=1}^{K} a_i Z_i\right\|^2 &= \sum_{i=1}^{K}\mathbb{E}[a_i^2\|Z_i\|^2] + 2\sum_{i<j}\mathbb{E}\langle a_i Z_i, a_j Z_j\rangle\\
&= v\sum_{i=1}^{K}\mathbb{E}a_i^2.
\end{aligned} \tag{97}$$

Let $K_B = 1 + \lfloor (B-R)/b_0 \rfloor$. Since costs are positive, $N_B \le K_B$, and its admission indicators are predictable:

$$\{i \le N_B\} = \{T_{i-1} \le B - R\} \in \mathcal{F}_{i-1}. \tag{98}$$

For $K = \max\{n, K_B\}$ set $a_i = \mathbf{1}\{i \le N_B\} - \mathbf{1}\{i \le n\}$. These signed coefficients are bounded and predictable. Their sum of squares counts exactly the indices between n and N_B, irrespective of which is larger. Therefore,

$$\begin{aligned}
S_{N_B} - S_n &= \sum_{i=1}^{K} a_i\xi_i,\\
\sum_{i=1}^{K} a_i^2 &= |N_B - n|,\\
\mathbb{E}\|S_{N_B} - S_n\|^2 &= V\sum_{i=1}^{K}\mathbb{E}a_i^2 = V\mathbb{E}|N_B - n|.
\end{aligned} \tag{99}$$

Conditioning throughout uses the history before each admission. The final count N_B can reveal information about earlier gradient errors. □

Lemma 4 (Count concentration in mean)**.** *Let $n_B = \lfloor B/c \rfloor$. Under the stated reservation rule,*

$$\begin{aligned}
&B - R < T_{N_B} \le B, \qquad \frac{B-R}{R} < N_B \le \frac{B}{b_0},\\
&\frac{B-R}{c} < \mathbb{E}N_B \le \frac{B}{c},\\
&\mathbb{E}|N_B - n_B| \le \frac{s}{c}\sqrt{\frac{B}{c}} + \frac{R}{c} + 1.
\end{aligned} \tag{100}$$

In particular, $N_B/B \to 1/c$ in L^1 and in probability. For $B \geq 2R$, one also has $B/N_B < 2R$ almost surely.

Proof. By definition, $T_{N_B-1} \leq B - R < T_{N_B}$ and $C_{N_B} \leq R$. Hence the realized cost and count satisfy

$$\begin{aligned} B - R < T_{N_B} = T_{N_B-1} + C_{N_B} \leq B - R + R = B, \\ b_0 N_B \leq T_{N_B} \leq R N_B, \\ \frac{B-R}{R} < N_B \leq \frac{B}{b_0}. \end{aligned} \tag{101}$$

The bounded number of predictable admissions gives the finite-sum identity

$$\begin{aligned} \mathbb{E}(T_{N_B} - cN_B) &= \sum_{i=1}^{K_B} \mathbb{E}[\mathbf{1}\{i \leq N_B\}(C_i - c)] \\ &= \sum_{i=1}^{K_B} \mathbb{E}\left[\mathbf{1}\{i \leq N_B\}\mathbb{E}[C_i - c \mid \mathcal{F}_{i-1}]\right] = 0, \\ \frac{B-R}{c} &< \frac{\mathbb{E}T_{N_B}}{c} = \mathbb{E}N_B \leq \frac{B}{c}. \end{aligned} \tag{102}$$

Apply Lemma 3 to $Z_i = C_i - c$ and $a_i = \mathbf{1}\{i \leq N_B\}$. Its exact identity, followed by Cauchy–Schwarz and the reserve sandwich, gives

$$\begin{aligned} \mathbb{E}|T_{N_B} - cN_B|^2 &= s^2 \mathbb{E}N_B \leq \frac{s^2 B}{c}, \\ \mathbb{E}|N_B - B/c| &\leq \frac{1}{c}\mathbb{E}|cN_B - T_{N_B}| + \frac{1}{c}\mathbb{E}|T_{N_B} - B| \\ &\leq \frac{1}{c}\left(\mathbb{E}|cN_B - T_{N_B}|^2\right)^{1/2} + \frac{R}{c} \\ &\leq \frac{s}{c}\sqrt{\frac{B}{c}} + \frac{R}{c}, \\ \mathbb{E}|N_B - n_B| &\leq \mathbb{E}|N_B - B/c| + |B/c - n_B| \\ &\leq \frac{s}{c}\sqrt{\frac{B}{c}} + \frac{R}{c} + 1. \end{aligned} \tag{103}$$

After division by B, the right-hand side tends to zero, proving the L^1 assertion. Markov's inequality proves convergence in probability. For $B \geq 2R$, Eq. (101) gives $N_B > (B-R)/R \geq B/(2R)$, which proves the reciprocal bound. □

D.4.3 Uniform integrability using only gradient second moments

To pass from convergence of N_B/B to expected squared error, we control dependence between the random denominator and numerator through uniform integrability. The next two lemmas establish it using only second moments of Y_i.

Lemma 5 (Uniform integrability of squared normalized sums)**.** *If ξ_i are iid mean-zero vectors in $\mathbb{R}^d$ and $\mathbb{E}\|\xi_i\|^2 = V < \infty$, then $\{\|m^{-1/2}\sum_{i=1}^m \xi_i\|^2 : m \geq 1\}$ is uniformly integrable. More explicitly, define $r_a = \mathbb{E}[\|\xi_1\|^2 \mathbf{1}\{\|\xi_1\| > a\}]$ for $a > 0$. For every $t > 0$,*

$$\sup_{m \geq 1} \mathbb{E}\left[\left\|\frac{S_m}{\sqrt{m}}\right\|^2 \mathbf{1}\left\{\left\|\frac{S_m}{\sqrt{m}}\right\|^2 > t\right\}\right] \leq 2\sqrt{\frac{(4a^2V + 3V^2)V}{t}} + 2r_a. \tag{104}$$

Proof. For $a > 0$, truncate and recenter each increment by setting $\eta_i = \xi_i \mathbf{1}\{\|\xi_i\| \leq a\} - \mathbb{E}[\xi_i \mathbf{1}\{\|\xi_i\| \leq a\}]$ and $\zeta_i = \xi_i - \eta_i$. Both sequences are iid and centered, although η_i and ζ_i

need not be independent within a pair. Subtracting the mean can only decrease a second moment, and $\mathbb{E}\xi_i = 0$ gives

$$\begin{aligned} \|\eta_i\| &\le 2a, \qquad v_a := \mathbb{E}\|\eta_i\|^2 \le V, \\ \mathbb{E}\|\eta_i\|^4 &\le 4a^2\mathbb{E}\|\eta_i\|^2 \le 4a^2V, \\ \zeta_i &= \xi_i \mathbf{1}\{\|\xi_i\| > a\} - \mathbb{E}[\xi_i \mathbf{1}\{\|\xi_i\| > a\}], \\ \mathbb{E}\|\zeta_i\|^2 &\le \mathbb{E}[\|\xi_i\|^2 \mathbf{1}\{\|\xi_i\| > a\}] = r_a \longrightarrow 0 \quad (a \to \infty). \end{aligned} \tag{105}$$

The last convergence follows directly from integrability of $\|\xi_i\|^2$ and dominated convergence.

To bound the fourth moment of the truncated sums, put $Q_j = \sum_{i=1}^{j} \eta_i$. Independence and centering yield $\mathbb{E}\|Q_{j-1}\|^2 = (j-1)v_a$. When expanding $\|Q_{j-1} + \eta_j\|^4$, the two terms linear in $\langle Q_{j-1}, \eta_j \rangle$ have zero expectation: one by $\mathbb{E}\eta_j = 0$, the other by $\mathbb{E}Q_{j-1} = 0$. Thus the pointwise Cauchy–Schwarz inequality and independence give

$$\begin{aligned} \mathbb{E}\|Q_j\|^4 &= \mathbb{E}\|Q_{j-1}\|^4 + \mathbb{E}\|\eta_j\|^4 + 2\mathbb{E}\|Q_{j-1}\|^2\mathbb{E}\|\eta_j\|^2 \\ &\quad + 4\mathbb{E}\langle Q_{j-1}, \eta_j \rangle^2 \\ &\le \mathbb{E}\|Q_{j-1}\|^4 + \mathbb{E}\|\eta_1\|^4 + 6\mathbb{E}[\|Q_{j-1}\|^2\|\eta_j\|^2] \\ &= \mathbb{E}\|Q_{j-1}\|^4 + \mathbb{E}\|\eta_1\|^4 + 6(j-1)v_a^2, \\ \mathbb{E}\left\|\frac{Q_m}{\sqrt{m}}\right\|^4 &\le \frac{m\mathbb{E}\|\eta_1\|^4 + 6v_a^2\sum_{j=1}^{m}(j-1)}{m^2} \\ &= \frac{\mathbb{E}\|\eta_1\|^4}{m} + 3\frac{m-1}{m}v_a^2 \le 4a^2V + 3V^2. \end{aligned} \tag{106}$$

Only truncated variables were required to have fourth moments.

Write $U_m = S_m/\sqrt{m}$, $A_{m,a} = Q_m/\sqrt{m}$, and $D_{m,a} = m^{-1/2}\sum_{i=1}^{m}\zeta_i$. Independent cross-term cancellation gives $\mathbb{E}\|U_m\|^2 = V$ and $\mathbb{E}\|D_{m,a}\|^2 = \mathbb{E}\|\zeta_1\|^2 \le r_a$. For $E_{m,t} = \{\|U_m\|^2 > t\}$, the norm inequality, Cauchy–Schwarz, and Markov's inequality now imply

$$\begin{aligned} \mathbb{E}[\|U_m\|^2 \mathbf{1}_{E_{m,t}}] &\le 2\mathbb{E}[\|A_{m,a}\|^2 \mathbf{1}_{E_{m,t}}] + 2\mathbb{E}[\|D_{m,a}\|^2 \mathbf{1}_{E_{m,t}}] \\ &\le 2\big(\mathbb{E}\|A_{m,a}\|^4\big)^{1/2}\Pr(E_{m,t})^{1/2} + 2r_a \\ &\le 2\sqrt{(4a^2V + 3V^2)\frac{\mathbb{E}\|U_m\|^2}{t}} + 2r_a \\ &= 2\sqrt{\frac{(4a^2V + 3V^2)V}{t}} + 2r_a. \end{aligned} \tag{107}$$

This is uniform in m. First choose a so that r_a is arbitrarily small, and then let t tend to infinity. The definition of uniform integrability follows, including the trivial case $V = 0$. □

Lemma 6 (Vanishing bounded multipliers)**.** *Suppose $W_B \ge 0$ is a uniformly integrable family and $0 \le H_B \le K$ almost surely for a fixed finite constant K. If $H_B \to 0$ in probability, then $\mathbb{E}[H_B W_B] \to 0$. No independence assumption is needed.*

Proof. For any $\varepsilon > 0$, boundedness gives

$$\begin{aligned} \mathbb{E}H_B &\le \varepsilon + K\Pr(H_B > \varepsilon), \\ \limsup_{B\to\infty} \mathbb{E}H_B &\le \varepsilon, \qquad \mathbb{E}H_B \longrightarrow 0. \end{aligned} \tag{108}$$

For any truncation threshold $t > 0$, split the product according to whether W_B exceeds t:

$$\begin{aligned} 0 \le \mathbb{E}[H_B W_B] &\le t\mathbb{E}H_B + K\mathbb{E}[W_B \mathbf{1}\{W_B > t\}], \\ \limsup_{B\to\infty} \mathbb{E}[H_B W_B] &\le K \sup_B \mathbb{E}[W_B \mathbf{1}\{W_B > t\}] \longrightarrow 0 \quad (t \to \infty). \end{aligned} \tag{109}$$

This proves the assertion. □

D.4.4 PROOF OF THE EXACT HARD-BUDGET MSE LIMIT

We now combine the preceding lemmas. It suffices to take $B \geq 2R$, so $n_B = \lfloor B/c \rfloor \geq 1$. Define $A_B = S_{N_B}/\sqrt{B}$, $Z_B = S_{n_B}/\sqrt{B}$, and $\rho_B = B/N_B$. Lemmas 3 and 4 first give the quantitative comparison with a deterministic sample count:

$$\begin{aligned} \mathbb{E}\|A_B - Z_B\|^2 &= \frac{V}{B}\mathbb{E}|N_B - n_B| \\ &\leq \frac{V}{B}\left(\frac{s}{c}\sqrt{\frac{B}{c}} + \frac{R}{c} + 1\right) \longrightarrow 0, \\ \mathbb{E}\|Z_B\|^2 &= \frac{n_B}{B}V \longrightarrow \frac{V}{c}. \end{aligned} \tag{110}$$

Furthermore, $n_B/B \leq 1/c$, so the family $\{\|Z_B\|^2 : B \geq 2R\}$ is uniformly integrable by Lemma 5. To make this scaling step explicit, if $F_m = \|S_m/\sqrt{m}\|^2$, then

$$\begin{aligned} \|Z_B\|^2 &= \frac{n_B}{B}F_{n_B} \leq \frac{1}{c}F_{n_B}, \\ \sup_{B\geq 2R}\mathbb{E}[\|Z_B\|^2\mathbf{1}\{\|Z_B\|^2 > t\}] &\leq \frac{1}{c}\sup_{m\geq 1}\mathbb{E}[F_m\mathbf{1}\{F_m > ct\}] \longrightarrow 0. \end{aligned} \tag{111}$$

The count lemma gives $N_B/B \to 1/c$ in probability. Its lower bound also controls the reciprocal:

$$\begin{aligned} &0 < \rho_B < 2R, \qquad c \leq R, \\ &|\rho_B - c| = \frac{c|N_B/B - 1/c|}{N_B/B} \leq 2Rc|N_B/B - 1/c| \longrightarrow 0 \quad \text{in probability}, \\ &0 \leq |\rho_B - c|^2 \leq 9R^2. \end{aligned} \tag{112}$$

Applying Lemma 6 with $H_B = |\rho_B - c|^2$ and $W_B = \|Z_B\|^2$ now shows $\mathbb{E}[|\rho_B - c|^2\|Z_B\|^2] \to 0$, despite possible cost–gradient dependence. Thus the random-denominator normalized estimator converges in L^2 to the deterministic-count comparison:

$$\begin{aligned} \mathbb{E}\|\rho_B A_B - cZ_B\|^2 &\leq 2\mathbb{E}[\rho_B^2\|A_B - Z_B\|^2] + 2\mathbb{E}[|\rho_B - c|^2\|Z_B\|^2] \\ &\leq 8R^2\mathbb{E}\|A_B - Z_B\|^2 + 2\mathbb{E}[|\rho_B - c|^2\|Z_B\|^2] \longrightarrow 0. \end{aligned} \tag{113}$$

Finally put $D_B = \rho_B A_B - cZ_B$. Expanding the squared norm and using Cauchy–Schwarz yields the expectation comparison

$$\begin{aligned} \left|\mathbb{E}\|\rho_B A_B\|^2 - c^2\mathbb{E}\|Z_B\|^2\right| &\leq 2c\,|\mathbb{E}\langle Z_B, D_B\rangle| + \mathbb{E}\|D_B\|^2 \\ &\leq 2c\big(\mathbb{E}\|Z_B\|^2\big)^{1/2}\big(\mathbb{E}\|D_B\|^2\big)^{1/2} + \mathbb{E}\|D_B\|^2 \longrightarrow 0, \\ B\,\mathbb{E}\left\|\frac{S_{N_B}}{N_B}\right\|^2 &= \mathbb{E}\|\rho_B A_B\|^2 \longrightarrow c^2\frac{V}{c} = cV. \end{aligned} \tag{114}$$

This proves Theorem 3 and Eq. (94), identifying the leading MSE constant for the random-count estimator.

Finite-budget bias and the role of the reservation rule. Although predictable isometry gives $\mathbb{E}S_{N_B} = 0$, $\mathbb{E}[S_{N_B}/N_B]$ need not vanish. For example, let iid costs be uniform on $\{1, 2\}$, set $Y_i = C_i$, and take $B = 3$, $R = 2$. Then $N_B = 1$ when $C_1 = 2$ and $N_B = 2$ when $C_1 = 1$, whence

$$\mathbb{E}\left[\frac{1}{N_B}\sum_{i=1}^{N_B} Y_i\right] = \frac{1}{2}\cdot 2 + \frac{1}{2}\cdot\frac{1 + 3/2}{2} = \frac{13}{8} \neq \frac{3}{2} = \mathbb{E}Y_1. \tag{115}$$

The limiting equality remains valid. The proof uses complete pools admitted on the basis of past costs and the fixed reserve R. Rejecting an already observed pool on the basis of its realized cost, retaining an incomplete pool, or updating the policy between the pools changes the probability model and requires a separate analysis.

D.5 FINITE-BUDGET CONFIDENCE BOUNDS

D.5.1 A FINITE-BUDGET CONFIDENCE CERTIFICATE

Theorem 3 identifies the actual asymptotic MSE. The following proposition gives a finite-budget certificate under the same second-moment assumption, using conservative Bernstein and maximal-inequality bounds.

Proposition 4 (Hard-budget accuracy with cost-dependent stopping)**.** *Fix either collector. Let $(Y_j, C_j)_{j\ge 1}$ be iid pool estimate–cost pairs with $\mathbb{E}Y_j = g_\star$, $\mathbb{E}\|Y_j - g_\star\|^2 = V < \infty$, and $b_0 \le C_j \le R$ almost surely, where $b_0 > 0$. Dependence between Y_j and C_j is unrestricted. For $B \ge R$, repeatedly start and complete a pool whenever at least R interactions remain; stop as soon as fewer than R remain. Let N_B be the completed-pool count. Write $c = \mathbb{E}C_j$, $s^2 = \mathrm{Var}(C_j)$, $a = (R-c)/3$, and $x = \log(1/\delta_c)$. For $\delta_c, \delta_g \in (0,1)$ with $\delta_c + \delta_g < 1$, choose $n = 1$ or any integer $n \ge 2$ satisfying*

$$(n-1)c + \sqrt{2(n-1)s^2 x} + ax \le B - R. \tag{116}$$

Then the realized total cost is at most B, and

$$\Pr\left(N_B \ge n, \quad \left\| \frac{1}{N_B} \sum_{j=1}^{N_B} Y_j - g_\star \right\|^2 \le \frac{V}{n\delta_g} \right) \ge 1 - \delta_c - \delta_g. \tag{117}$$

For fixed collector parameters and fixed confidence levels, the largest certified n satisfies $n = B/c + O(\sqrt{B} + 1)$ as $B \to \infty$. Consequently, the squared-error radius in Eq. (117) *has leading constant cV/δ_g.*

D.5.2 ONE-SIDED CONCENTRATION FOR RANDOM COLLECTION COSTS

Lemma 7 (Bernstein bound with a one-sided cost range)**.** *Let $C_1, \dots, C_k$ be iid with $C_i \le R$, $\mathbb{E}C_i = c$, and $\mathrm{Var}(C_i) = s^2 < \infty$. Set $a = (R-c)/3$. For every $x > 0$,*

$$\Pr\left(\sum_{i=1}^{k} (C_i - c) > \sqrt{2ks^2 x} + ax \right) \le e^{-x}. \tag{118}$$

Proof. If $s^2 = 0$, the sum is zero almost surely and the assertion is immediate. Otherwise $b = R - c > 0$. Let $Z = C_i - c$, so $Z \le b$, $\mathbb{E}Z = 0$, and $\mathbb{E}Z^2 = s^2$. For $\lambda \ge 0$, Taylor's integral remainder and monotonicity of the exponential imply the pointwise chain

$$\begin{aligned} e^{\lambda Z} - 1 - \lambda Z &= \lambda^2 Z^2 \int_0^1 (1-t) e^{t\lambda Z}\, dt \\ &\le \lambda^2 Z^2 \int_0^1 (1-t) e^{t\lambda b}\, dt \\ &= \frac{Z^2}{b^2}(e^{\lambda b} - 1 - \lambda b). \end{aligned} \tag{119}$$

This argument also covers negative Z, so only an upper bound on Z is needed. For each integer $r \ge 2$, induction gives $r! \ge 2 \cdot 3^{r-2}$. Taking expectations and using $1 + t \le e^t$ therefore yields, for $0 \le \lambda < 3/b$,

$$\begin{aligned} \mathbb{E}e^{\lambda Z} &\le 1 + \frac{s^2}{b^2}(e^{\lambda b} - 1 - \lambda b) \\ &\le \exp\left\{ \frac{s^2}{b^2} \sum_{r=2}^{\infty} \frac{(\lambda b)^r}{r!} \right\} \\ &\le \exp\left\{ \frac{s^2 \lambda^2}{2} \sum_{r=2}^{\infty} (\lambda b/3)^{r-2} \right\} \\ &= \exp\left\{ \frac{s^2 \lambda^2}{2(1 - a\lambda)} \right\}. \end{aligned} \tag{120}$$

Let $v_k = ks^2$, $q = \sqrt{2x/v_k}$, and choose $\lambda = q/(1+aq)$. Then $1 - a\lambda = 1/(1+aq) > 0$, and the threshold is $t = \sqrt{2v_k x} + ax = v_k q + a v_k q^2/2$. Markov's inequality, independence, and Eq. (120) give the explicitly evaluated Chernoff chain

$$\begin{aligned}
\Pr\left(\sum_{i=1}^{k} Z_i > t\right) &\le e^{-\lambda t}\mathbb{E}e^{\lambda \sum_{i=1}^{k} Z_i} \\
&= e^{-\lambda t}\prod_{i=1}^{k}\mathbb{E}e^{\lambda Z_i} \\
&\le \exp\left\{-\lambda t + \frac{v_k\lambda^2}{2(1-a\lambda)}\right\} \\
&= \exp\left\{-\frac{q(v_k q + a v_k q^2/2)}{1+aq} + \frac{v_k q^2}{2(1+aq)}\right\} \\
&= \exp\left\{-\frac{v_k q^2(1+aq)}{2(1+aq)}\right\} = e^{-x}.
\end{aligned} \tag{121}$$

This choice of λ retains the one-sided coefficient $a = (R-c)/3$ in the linear term. □

D.5.3 A REVERSE-MARTINGALE INEQUALITY FOR ALL SAMPLE MEANS

Lemma 8 (Dimension-free maximal inequality)**.** *Let $\xi_1, \dots, \xi_K$ be iid mean-zero vectors in $\mathbb{R}^d$ with $\mathbb{E}\|\xi_i\|^2 = V < \infty$. For $1 \le n \le K$ and $u > 0$,*

$$\Pr\left(\max_{n \le m \le K}\left\|\frac{1}{m}\sum_{i=1}^{m}\xi_i\right\|^2 \ge u\right) \le \frac{V}{nu}. \tag{122}$$

Proof. Set $S_m = \sum_{i=1}^{m}\xi_i$ and $\mathcal{F}_m = \sigma(S_m, \xi_{m+1}, \dots, \xi_K)$. These σ-fields decrease with m, because $S_{m+1} = S_m + \xi_{m+1}$ is $\mathcal{F}_m$-measurable. Exchangeability of the first $m+1$ vectors, together with independence from the remaining vectors, implies that their conditional expectations given $\mathcal{F}_{m+1}$ all coincide. Summing these expectations gives

$$\begin{aligned}
\mathbb{E}[\xi_i \mid \mathcal{F}_{m+1}] &= \frac{S_{m+1}}{m+1}, \qquad i \le m+1, \\
\mathbb{E}\left[\frac{S_m}{m}\middle|\mathcal{F}_{m+1}\right] &= \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}[\xi_i \mid \mathcal{F}_{m+1}] = \frac{S_{m+1}}{m+1}.
\end{aligned} \tag{123}$$

For precision, the first equality follows by testing against bounded $\mathcal{F}_{m+1}$-measurable functions: permuting the first $m+1$ coordinates leaves both their joint law and the test function unchanged.

Reverse the index by setting $J = K - n$, $M_j = S_{K-j}/(K-j)$, and $\mathcal{G}_j = \mathcal{F}_{K-j}$, for $0 \le j \le J$. Equation (123) makes $(M_j, \mathcal{G}_j)$ a square-integrable martingale. Conditional Jensen gives

$$\|M_j\|^2 = \|\mathbb{E}[M_J \mid \mathcal{G}_j]\|^2 \le \mathbb{E}[\|M_J\|^2 \mid \mathcal{G}_j]. \tag{124}$$

Let τ be the first index $j \in \{0, \dots, J\}$ for which $\|M_j\|^2 \geq u$, with $\tau = \infty$ if no crossing occurs. Since $\{\tau = j\} \in \mathcal{G}_j$, the disjoint first-crossing events yield the following maximal-inequality proof:

$$\begin{aligned} u \Pr(\tau \leq J) &= \sum_{j=0}^{J} u \Pr(\tau = j) \\ &\leq \sum_{j=0}^{J} \mathbb{E}[\|M_j\|^2 \mathbf{1}\{\tau = j\}] \\ &\leq \sum_{j=0}^{J} \mathbb{E}[\mathbb{E}[\|M_J\|^2 \mid \mathcal{G}_j] \mathbf{1}\{\tau = j\}] \\ &= \mathbb{E}[\|M_J\|^2 \mathbf{1}\{\tau \leq J\}] \\ &\leq \mathbb{E}\|M_J\|^2 = \mathbb{E}\left\| \frac{1}{n} \sum_{i=1}^{n} \xi_i \right\|^2 = \frac{V}{n}. \end{aligned} \tag{125}$$

The last equality follows by the independent cross-term cancellation in Eq. (67). The crossing event is exactly the maximum event in Eq. (122); dividing by u finishes the proof. □

D.5.4 Proof of the finite-budget confidence certificate

We couple the actual collector to an infinite iid sequence of potential pairs (Y_j, C_j). Unused pairs are not executed. Set $D = B - R$ and $K_B = 1 + \lfloor D/b_0 \rfloor$. Because costs are positive, the reservation rule gives the exact identities

$$\begin{aligned} N_B &= \max \left\{ m \in \{1, \dots, K_B\} : \sum_{j=1}^{m-1} C_j \leq D \right\}, \\ \{N_B \geq n\} &= \left\{ \sum_{j=1}^{n-1} C_j \leq D \right\}, \\ \sum_{j=1}^{N_B} C_j &\leq D + C_{N_B} \leq D + R = B. \end{aligned} \tag{126}$$

Equation (116) implies $(n-1)b_0 \leq (n-1)c \leq D$, so $n \leq K_B$. For $n \geq 2$, Lemma 7 yields

$$\begin{aligned} \Pr(N_B < n) &= \Pr\left(\sum_{j=1}^{n-1} C_j > D \right) \\ &\leq \Pr\left(\sum_{j=1}^{n-1} (C_j - c) > \sqrt{2(n-1)s^2 x} + ax \right) \\ &\leq e^{-x} = \delta_c. \end{aligned} \tag{127}$$

For $n = 1$ admission is certain. If $V = 0$, every $Y_j = g_\star$ almost surely, so only admission needs to be controlled. Otherwise apply Lemma 8 to $\xi_j = Y_j - g_\star$, with $u = V/(n\delta_g)$. Even though N_B can depend on these vectors through the costs, the event $\{N_B \geq n\}$ places its value among the

deterministic indices already covered by the maximum. Thus

$$\begin{aligned}
&\Pr\left(N_B < n \text{ or } \left\|\frac{1}{N_B}\sum_{j=1}^{N_B} Y_j - g_\star\right\|^2 > \frac{V}{n\delta_g}\right) \\
&\le \Pr(N_B < n) + \Pr\left(\max_{n\le m\le K_B}\left\|\frac{1}{m}\sum_{j=1}^{m}\xi_j\right\|^2 > \frac{V}{n\delta_g}\right) \\
&\le \delta_c + \frac{V}{n[V/(n\delta_g)]} = \delta_c + \delta_g.
\end{aligned} \tag{128}$$

Taking complements proves Eq. (117). The argument requires neither independence of costs and gradients nor a stopping-time property relative to the reverse filtration.

For an explicit largest certified count, suppose $s^2 > 0$ and $D \ge ax$. Solving the quadratic inequality in $\sqrt{n-1}$ gives

$$n = 1 + \left\lfloor \left(\frac{\sqrt{2s^2x + 4c(D-ax)} - \sqrt{2s^2x}}{2c} \right)^2 \right\rfloor. \tag{129}$$

If $D < ax$, take $n = 1$. If $s^2 = 0$, costs are deterministic and the sharper exact count $n = 1 + \lfloor D/c \rfloor$ is available directly from Eq. (126), without using the conservative Bernstein buffer.

Finally, to verify the stated asymptotic rate without discarding the reservation cost, set $d = D - ax$ and $q_0 = \sqrt{2s^2x}$. For fixed confidence levels and sufficiently large B, the unrounded quantity k_* in Eq. (129) satisfies $ck_* + q_0\sqrt{k_*} = d$. Hence

$$\begin{aligned}
k_* &\le \frac{d}{c}, \\
k_* &= \frac{d}{c} - \frac{q_0}{c}\sqrt{k_*} \ge \frac{d}{c} - \frac{q_0}{c}\sqrt{\frac{d}{c}}, \\
\frac{d}{c} - \frac{q_0}{c}\sqrt{\frac{d}{c}} &\le n \le 1 + \frac{d}{c}, \\
\frac{n}{B} &\longrightarrow \frac{1}{c}, \qquad B\frac{V}{n\delta_g} \longrightarrow \frac{cV}{\delta_g}.
\end{aligned} \tag{130}$$

The deterministic-cost case follows from the same floor argument. This completes the proof of Proposition 4.

Batching and interpretation. For a fixed collection batch of b independent nodes, let $Y_j = b^{-1}\sum_{r=1}^{b}\widehat{g}_{j,r}$. Its variance is V_a/b and its expected cost is bc_a. A reserve large enough to complete all b pools makes both hard-budget results apply at the batch level, with the same leading product $(bc_a)(V_a/b) = c_aV_a$. The finite-budget bias example in Eq. (115) shows why the usual deterministic-count formula $\mathbb{E}\|\overline{Y}_{N_B} - g_\star\|^2 = V/\mathbb{E}N_B$ can fail. Predictable admission and the random-denominator analysis account for this dependence.

D.6 Proof of the Main Budget-Efficiency Criterion

Proof of Theorem 1. For a collector a, let $Y_{a,j}$ be its complete-pool estimate and $C_{a,j}$ its interaction cost. The fixed-policy sampling model makes these pairs iid across pools, with mean gradient $g_\star$. For ECE, $M \le C_{\mathrm{E},j} \le MH$; for a fixed direct width w, $w \le C_{\mathrm{D},w,j} \le wH$. Thus positive lower bounds and finite reserves independent of B exist. Theorem 2 supplies finite second moments and the exact risks; Theorem 3 therefore applies to the arithmetic means of completed pools. Writing

$\mathcal{R}_a^{\mathrm{hard}}(B)$ for their actual MSE, we obtain

$$\begin{aligned}\lim_{B\to\infty} B\mathcal{R}_{\mathrm{E}}^{\mathrm{hard}}(B) &= c_{\mathrm{E}}(M)V_{\mathrm{E}}(M)\\ &= c_{\mathrm{E}}(M)\left[T+\frac{A+W}{M}+\frac{M-1}{M}Q\right] =: \mathcal{C}_{\mathrm{E}},\\ \lim_{B\to\infty} B\mathcal{R}_{\mathrm{D},w}^{\mathrm{hard}}(B) &= c_{\mathrm{D}}(w)V_{\mathrm{D}}(w)\\ &= w(1+\bar{\ell})\left[T+\frac{A+W}{w}\right] = (1+\bar{\ell})(wT+A+W) =: \mathcal{C}_{\mathrm{D}}(w).\end{aligned} \tag{131}$$

Because $T \geq 0$ and every direct width is a positive integer,

$$\begin{aligned}\mathcal{C}_{\mathrm{D}}(w) &= (1+\bar{\ell})(wT+A+W)\\ &\geq (1+\bar{\ell})(T+A+W) = \mathcal{C}_{\mathrm{D}}(1),\\ \min_{w\geq 1}\mathcal{C}_{\mathrm{D}}(w) &= \mathcal{C}_{\mathrm{D}}(1) =: \mathcal{C}_{\mathrm{D}}^{\star}.\end{aligned} \tag{132}$$

Consequently, ECE's leading constant is strictly smaller than the minimum of this family if and only if $\mathcal{C}_{\mathrm{E}} < \mathcal{C}_{\mathrm{D}}^{\star}$.

Suppose the strict gap holds and fix a direct width w. Let $\Delta_w = \mathcal{C}_{\mathrm{D}}(w) - \mathcal{C}_{\mathrm{E}}$. Equation (132) gives $\Delta_w \geq \mathcal{C}_{\mathrm{D}}^{\star} - \mathcal{C}_{\mathrm{E}} > 0$. The two limits in Eq. (131) imply that there exists $B_0(w) < \infty$ such that, for every $B \geq B_0(w)$,

$$\begin{aligned}B\mathcal{R}_{\mathrm{E}}^{\mathrm{hard}}(B) &< \mathcal{C}_{\mathrm{E}} + \Delta_w/3\\ &< \mathcal{C}_{\mathrm{D}}(w) - \Delta_w/3\\ &< B\mathcal{R}_{\mathrm{D},w}^{\mathrm{hard}}(B),\\ \mathcal{R}_{\mathrm{E}}^{\mathrm{hard}}(B) &< \mathcal{R}_{\mathrm{D},w}^{\mathrm{hard}}(B).\end{aligned} \tag{133}$$

The threshold may depend on w, which remains fixed as B grows. Tied leading constants leave finite-budget ordering undetermined. Fixed independent-node batching preserves these constants, as shown above. □