Measure What Changes the Decision: Bounded Lookahead for Queried-Only Deployment

Anonymous authors
Paper under double-blind review

Abstract

When an action must be measured before it can be deployed, information acquisition and resource allocation become coupled. A measurement budget Q controls which action outcomes are known, while a separate budget B couples the measured actions selected for deployment. Under this coupling, a query can have little immediate value yet make a later query useful. Exact one-step Expected Deployable-Utility Gain (EDUG) cannot price this delayed payoff. We therefore introduce A2-EDUG, a bounded two-step policy that ranks all queries by their immediate decision value, evaluates two-query plans for the four best candidates using 256 posterior draws, executes only the selected first query, and replans. With the belief model, query budget, and deployment solver held fixed, A2-EDUG reduces average regret along the query path by 11.99% on fresh DIV2K (95% group-bootstrap interval $[6.26\%, 17.55\%]$), by 16.16% on a speaker-disjoint FSDD confirmation, and by 7.72% ($[5.78\%, 9.57\%]$) in a 240-episode FSDD scale study. In both systems, the policy is about 3% worse after its first query and becomes better after the second, matching the predicted delayed-payoff mechanism. A control matched to the same number of exact solver calls and a monotone $k = 1, 2, 4, 8$ candidate sweep show that the gain comes from two-step planning, not simply from running more one-step evaluations.

1 Introduction

Many deployed systems must measure an action before deciding whether to use it. A simulator configuration must be run, a sensor must be read, a physical experiment must be performed, or a codec setting must produce a real bitstream. Computation can choose the next measurement, but it cannot replace the missing outcome.

Such systems spend two different resources. A measurement budget Q controls which cost–loss responses are observed; a deployment budget B couples the actions ultimately selected across items. An unmeasured action is unavailable for deployment. A query therefore reveals its outcome and expands the feasible action set at the same time. This setting connects costly information acquisition with constrained decision making (Singla, 2018; Goerigk et al., 2015). We ask how to spend a fixed budget of Q measurements so that the exact B-constrained decision is good after every query, not only at the end.

The difficulty is that the value of a measurement need not be immediate. A query can be useless by itself but valuable together with a later query, because the pair admits a different globally feasible allocation. A one-step policy will not pay this first-query cost. Conversely, unrestricted multistep planning is impractical because each simulated observation changes the set of deployable actions and requires another constrained solve. The central problem is therefore to recover delayed decision value without turning every query into a deep policy search.

Exact one-step Expected Deployable-Utility Gain (EDUG) provides the natural starting point: for each unseen cell, it averages the improvement in the exact queried-only allocation over posterior outcomes. We add a deliberately bounded continuation calculation. For an objective that averages improvement after the first and second queries, a candidate's value

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

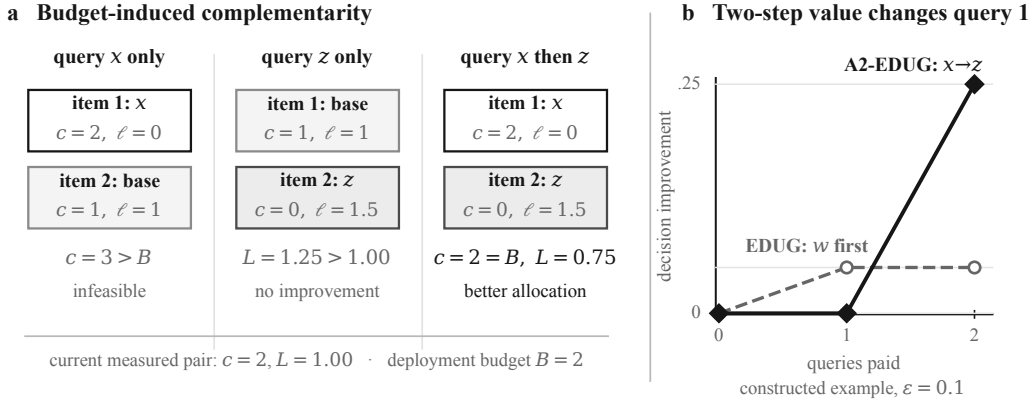


Figure 1: Budget-induced complementarity. Neither x nor z helps alone, but together they change the optimal deployment. A2-EDUG recovers the pair; EDUG selects w.

decomposes as $A_2 = D + \frac{1}{2}C$, where D is its immediate gain and C is the gain from the simulated second query. Lookahead changes the first query only when that future gain repays its immediate deficit. A2-EDUG evaluates the four highest-EDUG candidates with 256-draw two-step rollouts, executes only the selected first query, and replans from the observed outcome.

We evaluate the same acquisition rule in image compression and speech representation. On a previously unused DIV2K target split, A2-EDUG reduces AURC by 11.99% relative to exact one-step EDUG, with a 95% paired group-bootstrap interval of $[6.26\%, 17.55\%]$. We then instantiate the same controller for sample-rate–bit-depth allocation in real speech. A horizon selected on opened support data was frozen before any evaluation speaker was accessed. The resulting policy improved an untouched speaker-disjoint FSDD confirmation by 16.16%. Keeping the controller, prior, action menu, budgets, and horizon frozen, a 40× episode expansion over 960 previously unused utterances produced a tighter 7.72% reduction ($[5.78\%, 9.57\%]$). The raw losses differ by task, but the trajectory shape repeats: both systems are about 3% worse after the first query and turn positive after the second. In an additional analysis of fixed outcomes, top-four A2-EDUG also beats a control with the same number of exact solver calls by 9.97% on DIV2K and 16.72% on FSDD. The improvement therefore cannot be explained by extra one-step evaluations alone.

We make three contributions:

- We formulate queried-only deployment with separate measurement and deployment budgets, separate the value of new information from the value of making a new action deployable, and prove that exact one-step value can be arbitrarily myopic.
- We introduce A2-EDUG, which ranks all queries by immediate value, evaluates a two-query plan for the top four, executes one query, and replans. Its breadth k directly controls the compute–quality tradeoff.
- We demonstrate the same delayed-payoff pattern in learned image compression and speech representation. The gains persist when the number of exact solver calls is matched.

2 Queried-Only Decisions under Dual Budgets

Unknown response table. There are N items and a finite action menu $\mathcal{A} = \{0, \dots, M-1\}$. Executing action a on item i reveals a bounded task-specific cost–loss pair

$$\theta_{i,a} = (c_i(a), \ell_i(a)) \in \mathbb{R}_+^2, \qquad \ell_i(a) \in [0, \ell_{\max}]. \tag{1}$$

After q optional measurements, history $\mathcal{H}_q$ contains mandatory anchors and all paid outcomes. Let $\mathcal{O}_q$ be the measured cells and $\mathcal{A}_i(\mathcal{O}_q)$ the actions measured for item i.

Queried-only deployment. The final action vector $\mathbf{a} = (a_1, \ldots, a_N)$ must come from the visible product menu

$$\mathcal{D}(\mathcal{O}_q) = \prod_{i=1}^{N} \mathcal{A}_i(\mathcal{O}_q) \tag{2}$$

and satisfy the global budget $C_\theta(\mathbf{a}) = \sum_i c_i(a_i) \le B$. Its loss is $L_\theta(\mathbf{a}) = N^{-1} \sum_i \ell_i(a_i)$. The deployed allocation is

$$\widehat{\mathbf{a}}_q \in \arg \min_{\mathbf{a} \in \mathcal{D}(\mathcal{O}_q) : C_\theta(\mathbf{a}) \le B} L_\theta(\mathbf{a}), \tag{3}$$

with deterministic tie-breaking. Every selected cell has therefore been executed and its realized cost and loss are known.

For evaluation only, the full response table defines

$$\mathbf{a}_B^\star \in \arg \min_{\mathbf{a} \in \mathcal{A}^N : C_\theta(\mathbf{a}) \le B} L_\theta(\mathbf{a}). \tag{4}$$

The regret after q optional measurements is $r_q = L_\theta(\widehat{\mathbf{a}}_q) - L_\theta(\mathbf{a}_B^\star)$. We evaluate the entire fixed-Q acquisition path rather than only its endpoint:

$$\mathrm{AURC} = \frac{1}{Q} \sum_{q=0}^{Q-1} \frac{r_q + r_{q+1}}{2}. \tag{5}$$

A useful measurement must therefore improve the constrained deployable decision immediately or create value later along the remaining query path.

3 Why Immediate Decision Value Can Be Myopic

Let $J_B(h)$ denote the exact loss of the queried-only optimum at history h. Because a query adds one measured cell without removing old cells, the feasible set expands pointwise:

$$\mathcal{D}(\mathcal{O}_h) \subseteq \mathcal{D}(\mathcal{O}_{h \oplus (x,y)}), \qquad J_B(h \oplus (x,y)) \le J_B(h). \tag{6}$$

The expected immediate decision gain from query x is consequently nonnegative,

$$D_h(x) = \mathbb{E}_{Y_x | h}[J_B(h) - J_B(h, x, Y_x)] \ge 0. \tag{7}$$

This monotonicity has a Bayesian interpretation. Let $\mathcal{R}(h; \mathcal{D})$ be the Bayes risk of a terminal action set, assigning a bounded rejection loss to infeasible allocations. If $\mathcal{D}^-$ is the old set and $\mathcal{D}_x^+$ the set after measuring x, then the one-step value decomposes as follows.

Proposition 1 (Two sources of one-step value). For every query x,

$$\begin{aligned} V_1(x \mid h) = & \underbrace{\mathcal{R}(h; \mathcal{D}^-) - \mathbb{E}_{Y_x} \mathcal{R}(h, x, Y_x; \mathcal{D}^-)}_{V_{\mathrm{belief}}(x) \ge 0} \\ & + \underbrace{\mathbb{E}_{Y_x}\left[\mathcal{R}(h, x, Y_x; \mathcal{D}^-) - \mathcal{R}(h, x, Y_x; \mathcal{D}_x^+)\right]}_{V_{\mathrm{avail}}(x) \ge 0}. \end{aligned} \tag{8}$$

If every currently deployable cell is measured exactly, then $V_{\mathrm{belief}}(x) = 0$ and immediate value comes entirely from making the queried action available.

The first term is the value of learning about actions already available for deployment; the second is the value of making the queried action available. An observation can therefore have high entropy but zero decision value if no possible outcome changes feasibility or the optimizer. Conversely, an almost certain response can matter because querying it makes that action eligible. Appendix B gives two related stability results.

The global budget also creates complementarity between measurements. A cell that cannot improve the allocation alone may become useful after another cell changes the feasible combination. The resulting queried-set utility is monotone but need not exhibit diminishing returns, which is precisely where a one-step policy can fail.

Two-step value. Let $\pi(Y_x)$ choose a second query after the first outcome. Define its incremental continuation value

$$C_h(x;\pi) = \mathbb{E}[J_B(h,x,Y_x) - J_B(h,x,Y_x,\pi(Y_x),Y_\pi)]. \tag{9}$$

The two-step path gain—the average improvement after the first and second queries—has an exact decomposition.

Proposition 2 (Two-step path value). For any non-anticipating second-query rule π,

$$A_h(x;\pi) = \tfrac{1}{2}\mathbb{E}[(J_0 - J_1) + (J_0 - J_2)] = D_h(x) + \tfrac{1}{2}C_h(x;\pi). \tag{10}$$

Let $x_D = \arg\max_x D_h(x)$ be the one-step leader and x_A the best bounded lookahead candidate. Define the direct deficit $\Delta_D = D_h(x_D) - D_h(x_A)$ and continuation advantage $\Delta_C = \frac{1}{2}[C_h(x_A;\pi) - C_h(x_D;\pi)]$. Then

$$A_h(x_A;\pi) - A_h(x_D;\pi) = \Delta_C - \Delta_D. \tag{11}$$

Under strict ranking, lookahead changes the first query exactly when the future advantage repays the immediate sacrifice. A useful second query can therefore change the first choice even when outcomes are deterministic.

Proposition 3 (Unbounded one-step myopia gap). For every finite $c \geq 1$, there is a deterministic two-item queried-only allocation with three unseen actions and a two-measurement horizon for which exact one-step EDUG obtains less than $1/c$ of the decision improvement of the best two-query path. On this instance, A2-EDUG's top-four shortlist contains that path and recovers its improvement.

The construction uses two individually worthless measurements that jointly cross a budget boundary and a distractor with arbitrarily small positive immediate value (Appendix A). Because it has only three unseen actions, the top-four shortlist is exhaustive; deterministic outcomes also make the simulated second-query proxy exact.

From value to a practical policy. These results show why one-step scoring can be insufficient. A2-EDUG uses EDUG to shortlist four first queries and evaluates a two-query plan for each. The experiments keep the posterior and deployment solver fixed and ask whether this bounded lookahead chooses better measurements.

4 A2-EDUG: Bounded Planning for Delayed Value

4.1 Posterior model and shared setup

Every testbed uses the same structure: 13 possible cost–loss outcomes per item, two mandatory endpoint anchors, a queried-only exact multiple-choice knapsack, and a static support-fitted posterior. Within a testbed, all methods share the same prior, conditioning rule, random draws, anchors, and task-specific loss. Different testbeds fit independent priors; no posterior is transferred from images to audio.

For each item, the implementation interleaves $[\log(1 + c_a), \log(1 + \ell_a)]_{a=0}^{12}$ into a 26-dimensional vector. It fits an empirical mean and a covariance $0.9\widehat{\Sigma} + 0.1\,\mathrm{diag}(\widehat{\Sigma}) + 10^{-6}I$ to complete support curves (24 DIV2K images per codec; 12 FSDD clips), projects that covariance onto the positive-semidefinite cone, and uses analytic Gaussian conditioning with a pseudoinverse on the paid dimensions. At every matched state, 256 standard normal base draws are generated from a shared seed determined by episode, budget, query depth, and item identity. After inverse transformation, costs are rounded to positive integers, losses are clipped nonnegative, and paid cells retain their exact observed outcomes. Appendix C gives the complete numerical specification. Fitting each testbed independently keeps posterior construction fixed while isolating the acquisition policy.

4.2 Exact one-step EDUG

Let J_q be the loss of the current exact queried-only optimum and $J_{q+1}^{s,x}$ the optimum after inserting only candidate x with its outcome from posterior draw s. The empirical one-step

Algorithm 1 A2-EDUG: bounded planning for delayed decision value

Require: paid history $\mathcal{H}_q$, remaining measurement budget, deployment budget B
while an optional measurement remains do
  condition the frozen posterior on paid outcomes
  if at least two measurements remain then
    score all unseen cells by exact empirical EDUG; retain top k
    evaluate k two-query plans with 256 posterior draws
    select the first cell with the largest two-query score
  else
    select exact one-step EDUG
  end if
  execute only the selected first cell; observe, condition, and replan
end while
deploy the exact B-feasible allocation using measured cells only

score is

$$\widehat{V}_{\mathrm{EDUG}}(x \mid \mathcal{H}_q) = \frac{1}{S}\sum_{s=1}^{S}[J_q - J_{q+1}^{s,x}]_+, \qquad S = 256. \tag{12}$$

All unseen candidates are enumerated, and every per-draw deployment problem is solved exactly. "Exact one-step" refers to candidate enumeration, unbinned draw evaluation, and the terminal solver; the posterior expectation remains a 256-draw empirical approximation. Exact one-step EDUG executes the largest score, observes its real outcome, conditions the posterior, and repeats.

4.3 A2-EDUG: bounded continuation

A2-EDUG scores every unseen cell with Equation (12) and retains the top k possible first queries

$$\mathcal{C}_k(\mathcal{H}_q) = \mathrm{TopK}_{x \in \mathcal{U}(\mathcal{H}_q)}\, \widehat{V}_{\mathrm{EDUG}}(x \mid \mathcal{H}_q). \tag{13}$$

The primary experiments freeze $k = 4$. For each $x \in \mathcal{C}_k$ and draw s, the posterior is conditioned on the sampled first outcome. A simulated second cell $z_s(x)$ is chosen from all remaining unseen cells by evaluating the one-step score at the conditional posterior mean. The terminal solver inserts the same joint draw's outcome for that second cell, giving $J_{q+2}^{s,x,z_s(x)}$. The bounded score is

$$\widehat{V}_{\mathrm{A2}}(x \mid \mathcal{H}_q) = \frac{1}{2S}\sum_{s=1}^{S}\left([J_q - J_{q+1}^{s,x}]_+ + [J_q - J_{q+2}^{s,x,z_s(x)}]_+\right). \tag{14}$$

Only the maximizing first cell is paid for. Its real outcome is appended to history; all simulated second actions are discarded; the policy then replans. With one optional query remaining, it reverts to exact one-step EDUG.

A2-EDUG uses additional planning to choose more valuable measurements. We therefore evaluate average regret along the fixed-Q query path, not wall-clock planning speed.

5 Experiments: Image Compression and Speech Representation

Controlled comparison. Within each testbed, both methods share the support-fitted posterior, conditioning rule, anchors, task-specific loss, deployment budgets, exact queried-only solver, and shared random draws; each conditions only on its own measured outcomes. We compare A2-EDUG with one-step EDUG on mean AURC. This design isolates the paid-query policy: A2-EDUG spends additional planning work on bounded continuation while Q and B remain fixed. Bootstrap resampling uses episode groups without pooling tasks or treating repeated method, budget, codec, or query-state views as independent.

An earlier CLIC experiment established EDUG as the strongest of five one-step acquisition rules under a common posterior. It reduced AURC by 8.00% relative to the per-setting envelope of four baselines that do not use downstream decision value (Appendix H). The final experiments ask a narrower question: does two-step planning improve this decision-aware one-step comparator?

Fresh DIV2K image compression. Before processing any target image, we fixed the official DIV2K validation split as 24 support, 36 target, and 40 unused images and fitted both codec priors on support. A mean-scale and a scale-hyperprior learned codec (Minnen et al., 2018; Ballé et al., 2018; Choi et al., 2019; Bae et al., 2023) produce 13 real entropy-coded outcomes per image; cost is serialized bitstream length and loss is RGB squared error. The target forms nine four-image episodes, evaluated at three bits-per-pixel budgets. The complete design has 54 matched method settings, 108 trajectories, and 1,560 real support-plus-target codec outcomes. The nine content groups are synchronously resampled in 50,000 paired bootstrap replicates.

FSDD speech signal representation. The second system uses real FSDD waveforms (Jackson et al., 2018). Thirteen independently generated sample-rate–bit-depth representations reveal actual packed payload bits and normalized waveform MSE; each representation is encoded directly from the original waveform rather than derived from another action. Four clips form a globally budgeted episode. The same top-four, 256-draw A2-EDUG controller and exact EDUG comparator are used, with an audio-specific prior fitted to 12 clips from two support speakers.

A 12-query horizon was selected on support outcomes and then frozen together with the method, prior, action menu, budgets, metric, and evaluation protocol before any evaluation-speaker waveform was processed. We then evaluated the fixed policy in a two-speaker pilot and a held-out confirmation on two further speakers. Appendix G reports the complete support-selection and evaluation sequence.

Forty-fold same-task scale study. Finally, we keep the FSDD controller, prior, action menu, three budgets, loss, and $Q = 12$ horizon fixed while increasing the study from six confirmation episodes to 240 four-clip episodes: 960 previously unused utterances and 12,480 real representation outcomes. This larger study estimates the same-task effect more precisely within the same six-speaker population. Its 50,000-replicate paired bootstrap resamples the 240 episode groups within 16 recording-index strata; the three budgets remain repeated views within each group.

6 Results: The First Query Pays Off after the Second

Table 1: Primary results for A2-EDUG. Each testbed is analyzed separately.

Evaluation	Matched units	Reduction	Evidence
Fresh DIV2K	9 groups × 2 codecs × 3 B	**11.99**%	95% CI [6.26, 17.55]%
FSDD confirmation	6 groups × 3 B	**16.16**%	5/6 groups; 3/3 B
FSDD scale	240 groups × 3 B	**7.72**%	95% CI [5.78, 9.57]%

On fresh DIV2K, A2-EDUG lowers AURC from (113.364 to 99.776) $\times 10^{-6}$, an 11.99% reduction. The paired nine-group interval is [6.26%, 17.55%]; both codecs, all three budgets, and all nine content-group means are positive, although not every paired cell improves.

Both the FSDD pilot and held-out confirmation improve over EDUG; the six-group confirmation reduces AURC by 16.16%. In the 240-episode study, the reduction is 7.72%, with a stratified paired interval of [5.78%, 9.57%]. Mean effects are positive for all three budgets and all recording-index strata. A2-EDUG wins or ties in 176 of 240 episode groups; 64 favor EDUG.

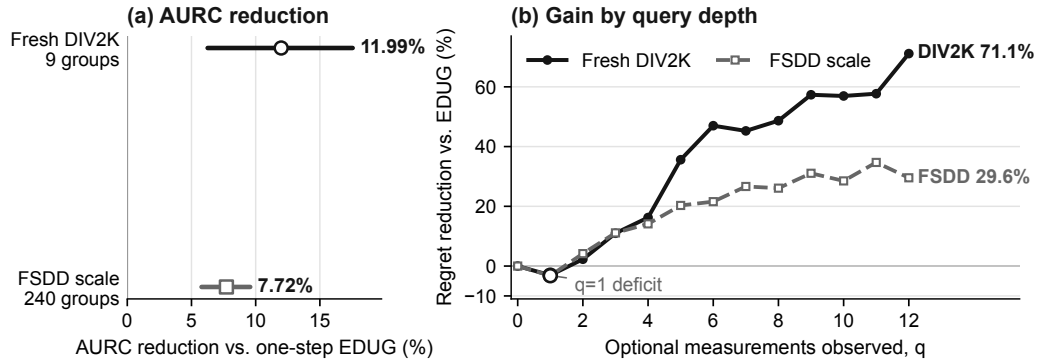


Figure 2: Delayed payoff across both systems. (a) AURC reduction with system-specific 95% intervals. (b) Mean regret reduction by query depth; positive values favor A2-EDUG.

The gain appears after the second query. Figure 2 resolves the AURC gain by query depth. On DIV2K, A2-EDUG is 3.07% worse after the first query, 2.28% better after the second, and 71.12% better at $q = 12$. In the 240-group FSDD study it is 3.23% worse at $q = 1$, 4.17% better at $q = 2$, and 29.55% better at $q = 12$. In both systems, the first-query deficit is repaid at $q = 2$, and the advantage is larger by $q = 12$, matching the behavior predicted by two-step planning.

Number of candidates and matched compute. An additional analysis evaluates $k \in \{1, 2, 4, 8\}$ without changing the posterior, paid history, rollout depth, or deployment solver; $k = 1$ exactly reproduces EDUG and $k = 4$ exactly reproduces every primary trajectory. AURC improves monotonically with k in both systems. With the number of solver calls matched, a repeated-EDUG control uses the same 568,844 solver calls per trajectory as top-four A2-EDUG but remains 9.97% worse on DIV2K (95% interval $[5.18, 14.55]$%) and 16.72% worse on FSDD ($[9.86, 22.57]$%). The gain is therefore not explained by additional one-step estimates. Top-eight yields another 4.07% and 2.99% reduction at about 1.84× the planning time, making k a direct control over computation and accuracy. Appendix E compares the shortlist with exhaustive search, checks posterior calibration, and reports a fixed-second-query control.

Execution and mechanism checks. Every reported allocation uses measured cells only, respects the realized budget, and issues distinct optional queries; no simulated second action is executed. In the large FSDD study, A2-EDUG changes the query path in 639 of 720 paired budget–group cells.

7 Related Work

Probing before combinatorial choice. Price-of-information and stochastic optimization study decisions whose parameters become known only after costly probes (Singla, 2018; Goerigk et al., 2015). Structured variants place linear, knapsack, or other combinatorial constraints on the downstream decision or on the probes themselves (Ryzhov & Powell, 2012; Chen et al., 2014). We study a fixed measurement budget and evaluate every point along the query path. Each query reveals an action's cost and loss, and the deployment decision couples all items.

Decision-aware and nonmyopic acquisition. Bayesian decision theory, decision-region active learning, loss-driven active learning, and targeted active learning value observations through downstream utility rather than uncertainty alone (Russell & Wefald, 1991; Javdani et al., 2014; Huang et al., 2026; Neiswanger et al., 2022; Filstroff et al., 2024). Knowledge gradient provides a canonical one-step decision-value rule (Frazier et al., 2008), while nongreedy

feature acquisition and multi-step Bayesian optimization use future evaluations to avoid myopic choices (Valancius et al., 2024; Wu & Frazier, 2019; Zhang et al., 2021; Cheon et al., 2025). A2-EDUG applies bounded two-step planning to queried-only, globally coupled deployment. Each simulated query updates the posterior and the available action set before re-solving the deployment problem. We evaluate regret over the full query path at a fixed measurement budget.

Test-time resources. Test-time scaling allocates additional search, sampling, verification, or routing computation (Snell et al., 2025; Ong et al., 2025). Our setting is complementary: planning computation chooses which external outcomes to buy, but only paid measurements make actions deployable. We therefore keep planning compute, measurement count, and deployment budget as three distinct quantities.

8 Scope and Discussion

Under queried-only deployment, bounded continuation spends a fixed measurement budget more effectively than exact one-step decision value. Matching 568,844 solver calls per trajectory separates two-step planning from simply repeating one-step calculations: repeated EDUG remains worse, while evaluating more first-query candidates improves AURC in both systems. The fact that A2-EDUG is initially worse but becomes better after the second query is the behavior predicted by its two-step objective.

The fixed-second-query control performs similarly to outcome-adaptive continuation. The evidence therefore supports the value of a two-step horizon, not a separate benefit from branching on the simulated first outcome. The top-four shortlist contains the best two-query plan in 91.36% of DIV2K states and 83.33% of FSDD states; a larger shortlist offers further gains at higher planning cost.

Learned compression and speech representation instantiate the same decision structure with real measured bit costs, but use different losses and independently fitted priors. Effects are therefore estimated separately rather than pooled. The audio horizon was selected on support before the speaker-disjoint evaluation, and the 240-episode study improves same-task precision within the same six-speaker universe. Within this evidence boundary, the two systems support the same conclusion: paying a small immediate cost can improve the entire subsequent decision path.

9 Conclusion

When measurements determine which actions may be deployed, their value is not always visible one step at a time. EDUG prices immediate improvement; A2-EDUG adds a bounded continuation calculation and replans after each paid outcome. Across image compression and speech representation, the policy accepts a small first-query cost, becomes better after the second query, and lowers average query-path regret under the same measurement and deployment budgets. Across these two applications, the result shows that planning for the decision ahead can be more valuable than refining the next decision alone.

AI Use Statement

In this work, we used generative AI tools to help implement parts of our baseline methods. We did not use generative AI tools to formulate hypotheses, design experiments, develop theoretical results, write proofs, or interpret results; generating synthetic data, translation, and qualitative data analysis are not applicable to this work. Additionally, we used generative AI tools to edit the manuscript for readability and to help format references. All AI-generated code was reviewed, tested, and verified against reference implementations by two of the authors, and all edited text was checked by the authors for accuracy. The authors take full responsibility for the content of this paper, including any text, code, or claims produced with AI assistance.

References

Juhan Bae, Michael R. Zhang, Michael Ruan, Eric Wang, So Hasegawa, Jimmy Ba, and Roger Grosse. Multi-rate VAE: Train once, get the full rate-distortion curve. In International Conference on Learning Representations, 2023. URL https://openreview.net/forum?id=OJ8aSjCaMNK.

Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. In International Conference on Learning Representations, 2018. URL https://openreview.net/forum?id=rkcQFMZRb.

Shouyuan Chen, Tian Lin, Irwin King, Michael R. Lyu, and Wei Chen. Combinatorial pure exploration of multi-armed bandits. In Advances in Neural Information Processing Systems, volume 27, 2014.

Mujin Cheon, Jay H. Lee, Dong-Yeun Koh, and Calvin Tsay. EARL-BO: Reinforcement learning for multi-step lookahead, high-dimensional Bayesian optimization. In Proceedings of the 42nd International Conference on Machine Learning, volume 267 of Proceedings of Machine Learning Research, pp. 10182–10198. PMLR, 2025. URL https://proceedings.mlr.press/v267/cheon25a.html.

Yoojin Choi, Mostafa El-Khamy, and Jungwon Lee. Variable rate deep image compression with a conditional autoencoder. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3146–3154, 2019.

Louis Filstroff, Iiris Sundin, Petrus Mikkola, Aleksei Tiulpin, Juuso Kylmäoja, and Samuel Kaski. Targeted active learning for Bayesian decision-making. Transactions on Machine Learning Research, 2024. URL https://openreview.net/forum?id=KxPjuiMgmm.

Peter I. Frazier, Warren B. Powell, and Savas Dayanik. A knowledge-gradient policy for sequential information collection. SIAM Journal on Control and Optimization, 47(5): 2410–2439, 2008. doi: 10.1137/070693424.

Marc Goerigk, Manoj Gupta, Jonas Ide, Anita Schöbel, and Sandeep Sen. The robust knapsack problem with queries. Computers & Operations Research, 55:12–22, 2015. doi: 10.1016/j.cor.2014.09.010.

Zhuoyue Huang, Freddie Bickford Smith, and Tom Rainforth. Loss-driven Bayesian active learning. In Proceedings of the 29th International Conference on Artificial Intelligence and Statistics, 2026. URL https://openreview.net/forum?id=0B1RuEcUph.

Zohar Jackson, César Souza, Jason Flaks, Yuxin Pan, Hereman Nicolas, and Adhish Thite. Jakobovski/free-spoken-digit-dataset: v1.0.8, 2018. URL https://doi.org/10.5281/zenodo.1342401.

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

David Minnen, Johannes Ballé, and George D. Toderici. Joint autoregressive and hierarchical priors for learned image compression. In Advances in Neural Information Processing Systems, volume 31, 2018.

Willie Neiswanger, Lantao Yu, Shengjia Zhao, Chenlin Meng, and Stefano Ermon. Generalizing Bayesian optimization with decision-theoretic entropies. In Advances in Neural Information Processing Systems, volume 35, pp. 21016–21029. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper/2022/hash/8451a20c5a7e0ee5671dda28f7daf7f3-Abstract-Conference.html.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M. Waleed Kadous, and Ion Stoica. RouteLLM: Learning to route LLMs with preference data. In International Conference on Learning Representations, 2025. URL https://openreview.net/forum?id=8sSqNntaMr.

Stuart J. Russell and Eric H. Wefald. Do the Right Thing: Studies in Limited Rationality. MIT Press, Cambridge, Massachusetts, 1991. ISBN 9780262181440.

Ilya O. Ryzhov and Warren B. Powell. Information collection for linear programs with uncertain objective coefficients. SIAM Journal on Optimization, 22(4):1344–1368, 2012. doi: 10.1137/12086279X.

Sahil Singla. The price of information in combinatorial optimization. In Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 2523–2532. SIAM, 2018. doi: 10.1137/1.9781611975031.161. URL https://doi.org/10.1137/1.9781611975031.161.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling model parameters for reasoning. In International Conference on Learning Representations, 2025. URL https://openreview.net/forum?id=4FWAwZtd2n.

Michael Valancius, Maxwell Lennon, and Junier Oliva. Acquisition conditioned oracle for nongreedy active feature acquisition. In Proceedings of the 41st International Conference on Machine Learning, volume 235 of Proceedings of Machine Learning Research, pp. 48957–48975. PMLR, 2024. URL https://proceedings.mlr.press/v235/valancius24a.html.

Jian Wu and Peter I. Frazier. Practical two-step lookahead Bayesian optimization. In Advances in Neural Information Processing Systems, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/hash/2e6d9c6052e99fcdfa61d9b9da273ca2-Abstract.html.

Yunxiang Zhang, Xiangyu Zhang, and Peter Frazier. Constrained two-step look-ahead Bayesian optimization. In Advances in Neural Information Processing Systems, volume 34, pp. 12563–12575, 2021.

A Proofs

Proposition 1. Add and subtract $\mathbb{E}_{Y_x}\mathcal{R}(h, x, Y_x; \mathcal{D}^-)$ in the one-step risk reduction. The belief term is nonnegative because a Bayes action may ignore the new observation:

$$\mathbb{E}_{Y_x} \min_{\mathbf{a}\in\mathcal{D}^-} \mathbb{E}[\overline{L}_B(\mathbf{a}, \theta) \mid h, Y_x] \le \min_{\mathbf{a}\in\mathcal{D}^-} \mathbb{E}_{Y_x}\mathbb{E}[\overline{L}_B(\mathbf{a}, \theta) \mid h, Y_x] \tag{15}$$

$$= \mathcal{R}(h; \mathcal{D}^-). \tag{16}$$

The availability term is nonnegative because $\mathcal{D}^- \subseteq \mathcal{D}_x^+$. If every old cell outcome is observed exactly, all old terminal losses are constants; conditioning on Y_x cannot change them, so the belief term is zero.

Proposition 2. Write $J_0 = J_B(h)$, $J_1 = J_B(h, x, Y_x)$, and $J_2 = J_B(h, x, Y_x, \pi(Y_x), Y_\pi)$. Then

$$\tfrac{1}{2}\mathbb{E}[(J_0 - J_1) + (J_0 - J_2)] = \tfrac{1}{2}\mathbb{E}[(J_0 - J_1) + (J_0 - J_1) + (J_1 - J_2)] \tag{17}$$

$$= D_h(x) + \tfrac{1}{2}C_h(x; \pi). \tag{18}$$

Subtracting this identity at x_D from the identity at x_A gives Equation (11).

Proposition 3. Fix $c \ge 1$ and choose $0 < \epsilon < \min\{1/4, 1/(2c)\}$. There are two items, budget $B = 2$, and one measured action per item, each with cost 1 and loss 1. Three actions are unseen: x for item one has cost 2 and loss 0; z for item two has cost 0 and loss 1.5; and distractor w for item one has cost 1 and loss $1 - \epsilon$. Querying x alone cannot replace the measured pair because its cost with item two is 3; querying z alone increases loss. Their one-step values are zero. Querying w immediately reduces average loss by $\epsilon/2$, so exact one-step EDUG selects it. With one query left, neither x nor z can improve that allocation, and the greedy path gains only $\epsilon/2$. All three unseen actions enter A2-EDUG's top-four shortlist. The two-step scores of x and z are both $\frac{1}{2}(0 + 1/4) = 1/8$, while that of w is $\epsilon/2 < 1/8$, so A2-EDUG selects one member of the complementary pair. Its deterministic second-query proxy selects the other, unlocking a feasible allocation of cost 2 and average loss 0.75, for decision improvement 1/4. The achieved ratio is $2\epsilon < 1/c$.

Deterministic complementarity without branching. Let queried-set utility satisfy $F(\varnothing) = 0$, $F(a) = 4$, $F(b) = 3$, $F(c) = 0$, $F(a, b) = F(a, c) = 4$, and $F(b, c) = 7$. One-step value chooses a and obtains path $(4, 4)$, with path score 4. Two-step lookahead chooses b then c and obtains $(3, 7)$, with score 5. Outcomes are deterministic, so branch option value is zero; the advantage is fixed continuation complementarity.

B Additional Structural Results

Lemma 1 (Allocation stability). Assume a unique oracle optimizer. Let $\Delta_B > 0$ be its minimum average-loss gap to another feasible allocation and $\gamma_B = \min_{\mathbf{a}\in\mathcal{A}^N} |B - C_\theta(\mathbf{a})| > 0$ the minimum budget slack. If every cell loss and cost estimate is within η_ℓ and η_c of truth, with $2\eta_\ell < \Delta_B$ and $N\eta_c < \gamma_B$, then the estimated feasible set and optimizer equal the true ones.

The cost condition prevents any allocation from crossing the budget boundary. The loss-gap error between two allocations is at most $2\eta_\ell$, so no challenger overtakes the unique optimizer.

Let Π be a finite or countable positive-mass partition of possible query outcomes. If Π' refines Π, the expected terminal decision value under Π' is at least that under Π. The refined controller may always ignore the extra distinction. Strict improvement occurs when the refinement separates positive-mass sub-bins with different unique terminal optimizers. This follows from the standard "expectation of a minimum versus minimum of an expectation" argument.

C Static Gaussian Belief Specification

For one complete 13-action support curve, define the interleaved latent vector

$$u = (\log(1 + c_0), \log(1 + \ell_0), \ldots, \log(1 + c_{12}), \log(1 + \ell_{12})) \in \mathbb{R}^{26}. \tag{19}$$

The prior mean is the sample mean. With the unbiased sample covariance S, the frozen covariance is

$$\Sigma = 0.9S + 0.1 \operatorname{diag}(S) + 10^{-6} I, \tag{20}$$

followed by symmetrization and clipping negative eigenvalues to zero. Each DIV2K codec fits its own prior from 24 complete support-image curves; FSDD uses the 12 prespecified support-fit clips. The scale study reuses this fitted audio prior rather than refitting it on the 960 evaluation utterances. Neither evaluation outcomes nor task identifiers, budgets, or policy labels enter a fit, and the four items in an episode are conditioned independently without a cross-item covariance model.

For the two latent dimensions of every paid action, collect observed indices O and transformed values u_O; let U be the unseen dimensions. Per item, conditioning is

$$\mu_{U|O} = \mu_U + \Sigma_{UO} \Sigma_{OO}^{+} (u_O - \mu_O), \tag{21}$$

$$\Sigma_{UU|O} = \Sigma_{UU} - \Sigma_{UO} \Sigma_{OO}^{+} \Sigma_{OU}, \tag{22}$$

where $+$ is the Moore–Penrose pseudoinverse. Observed means are set to their exact transformed values, their covariance rows and columns are zeroed, and the conditional covariance is projected onto the positive-semidefinite cone.

For each matched episode–budget–depth–item state, the implementation hashes candidate-id|episode-id|B{budget}|Q{depth}|item{item} with SHA-256, interprets the first 16 bytes as a PCG64 seed, and generates $S = 256$ standard-normal vectors. Both methods therefore receive the same random vectors at a matched state. These vectors are multiplied by a matrix square root of the conditional covariance. Latent values are clipped to $[-20, 40]$ before expm1; cost is rounded to the nearest integer and lower-bounded by one, while loss is lower-bounded by zero. Conditional mean scores use $\operatorname{expm1}(\mu + \frac{1}{2} \operatorname{diag} \Sigma)$ with the same clipping. In the planner adapter, the mean and all 256 draws of every paid cell are replaced by its exact outcome; simulated conditional means rank unseen continuation cells only. We use no monotonicity projection, posterior-family switching, or target calibration; only the acquisition rule changes between methods.

D Acquisition and Compute Details

At every state with at least two optional measurements remaining, A2-EDUG scores all unseen candidates, retains four possible first queries, and evaluates a 256-draw two-query plan for each. For every simulated first outcome, it ranks all remaining second queries by their one-step score at the conditional posterior mean. It executes only the selected first query; the simulated second query is neither executed nor carried into the next decision. The final optional step uses exact one-step EDUG. Score and cell ties are lexicographic.

On fresh DIV2K, all 594 eligible states follow this procedure and all 2,376 two-query plans use 256 draws. The selected first query differs from the one-step choice in 168 states. In the large FSDD study, all 1,440 trajectories contain 12 distinct optional queries; every deployed action was measured and respects the budget. Planning compute is not matched in the primary comparison: the separate analysis below controls for the number of exact solver calls.

E Approximation Quality, Compute, and Calibration

After fixing the primary analysis, we used existing outcomes to compare planning variants on all 54 DIV2K settings and a deterministic 48-setting FSDD subset. Here, $k = 1$ reproduces EDUG and $k = 4$ reproduces A2-EDUG in every setting. Across all 630 trajectories,

Table 2: Planning comparison. Positive entries are AURC reductions relative to one-step EDUG on the same settings. Solver calls and seconds are per trajectory; time is reported as DIV2K/FSDD on the same RTX 5090 run.

Policy	Solver calls	Time (s)	DIV2K	FSDD
$k = 1$ (EDUG)	230.9k	10.10/9.47	0.00%	0.00%
$k = 2$	343.6k	14.77/14.56	9.37%	10.60%
$k = 4$ (primary)	568.8k	24.72/24.79	11.99%	16.25%
$k = 8$	1019.4k	45.57/45.57	15.56%	18.75%
Compute-matched EDUG	568.8k	20.86/19.80	2.24%	−0.57%
Fixed second, $k = 4$	568.8k	24.99/24.77	13.33%	15.44%

deployment uses measured actions only, respects the budget, and never carries a simulated second query forward.

At query depths $q \in \{0, 4, 8\}$, we compare the top-four shortlist with every possible first query. The shortlist contains the best two-query plan in 91.36% of DIV2K states and 83.33% of FSDD states; its median retained value equals the exhaustive best in both systems. Top-eight raises recall to 95.68% and 87.50%. On 6 DIV2K and 12 FSDD cells where the entire acquisition path is evaluated exhaustively, the $k = 4$ AURC gap to exhaustive is 13.14% and 10.85%; $k = 8$ reduces these gaps to 0.13% and 7.65%. Increasing k therefore improves accuracy at higher computational cost.

We also check the support-fitted posterior at $q = 0$ on every held-out interior cell. Central 90% loss-interval coverage is 98.74% on DIV2K and 64.82% on FSDD, with median empirical percentiles 0.504 and 0.609. Cost coverage is 95.45% and 100.00%, with median percentiles 0.531 and 0.473. The beliefs are centered in both systems, although the audio loss posterior is less well calibrated. Finally, top-four A2-EDUG differs from the fixed-second control by −1.55% on DIV2K and +0.96% on FSDD; both paired intervals cross zero. This supports the two-step policy as a whole, but not a separate benefit from branching on the simulated first outcome.

F Fresh DIV2K Protocol and Slices

Official DIV2K validation identifiers 0801–0824 are support, 0825–0860 are target, and 0861–0900 remain unused. We generate 624 support outcomes and fit both codec priors before processing any target image. Each codec evaluates 60 used images at 13 actions, for 1,560 real serialized-bit outcomes. The 36 target images form nine four-image groups. With two codecs, three budgets, and two methods, the design contains 54 paired cells and 108 trajectories. Bootstrap replicates resample content groups synchronously and retain codec, budget, and method views.

Table 3: Fresh DIV2K AURC by codec and deployment budget, in units of 10^{-6}.

Slice	A2-EDUG	One-step EDUG	Reduction
Mean-scale codec	106.150	120.420	11.85%
Scale-hyperprior codec	93.402	106.309	12.14%
0.5 bpp	125.095	150.509	16.89%
0.8 bpp	110.448	119.690	7.72%
1.1 bpp	63.786	69.894	8.74%

G FSDD Development, Confirmation, and Scale Study

Action construction. The dataset has six speakers, ten digits, and 50 recordings per speaker–digit pair (Jackson et al., 2018). Each action independently resamples the original waveform to a fixed sample rate, quantizes it to a fixed bit depth, packs the payload,

reconstructs to the original sample rate, and returns actual payload bits plus normalized waveform MSE. No higher-quality action supplies a lower-quality action for free. The two endpoint anchors are mandatory; deployment chooses one measured representation per clip under a global payload budget.

Data roles and evaluation sequence. Before processing any waveform, speaker and filename identities fixed 24 support clips, 8 pilot clips, 24 confirmation clips, and the unused remainder. The static prior uses only 12 support clips; the remaining support outcomes were reserved for method and horizon selection. This support stage selected $Q = 12$. Before any pilot or confirmation waveform was opened, we fixed $Q = 12$, A2-EDUG, exact EDUG, the prior, action menu, budgets, loss, and evaluation procedures.

The prospective two-speaker pilot improved AURC by 22.66%, with 2/2 groups, 3/3 budgets, and 6/6 cells positive. The subsequent two-speaker confirmation improved AURC by 16.16%, with 5/6 groups, 3/3 budgets, and 14/18 cells positive. Together, these evaluations test transfer to speakers excluded from prior fitting, using the horizon selected on support data.

Scale replication. The scale study reuses the exact frozen prior and controller and selects recording indices 2–17, excluding every clip used in support, pilot, or confirmation. For each index, 15 balanced four-clip groups cover the six speakers and ten digits exactly once, yielding 240 groups and 960 unique clips. The 50,000-replicate bootstrap samples groups within each recording-index stratum. It estimates same-task episode performance over this fixed speaker universe. Six per-speaker summaries collapse into three fixed co-occurrence pairs and are therefore not interpreted as six independent effects.

Table 4: FSDD scale AURC by payload budget.

Budget	A2-EDUG	One-step EDUG	Reduction
Low	0.078784	0.086624	9.05%
Medium	0.031013	0.032537	4.68%
High	0.009958	0.010613	6.17%

H Earlier One-Step Foundation Comparison

Before A2-EDUG was designed, an earlier CLIC experiment compared exact one-step EDUG with posterior total variance, visible-decision EVSI, decision-targeted QMI, and predictive log-volume under the same static Gaussian posterior. The design has nine content groups, two codecs, three budgets, and five methods, for 270 trajectories and 1,560 real codec outcomes. EDUG reduces AURC by 8.00% relative to the per-setting envelope of the four baselines that do not use downstream decision value, with an absolute group-bootstrap interval of $[1.902, 7.216] \times 10^{-6}$. We therefore use CLIC to choose a strong one-step comparator, not as evidence for A2-EDUG, which was designed later.

Table 5: CLIC one-step foundation comparison, AURC in units of 10^{-6}.

	EDUG	PTV	EVSI	DT-QMI	Log-volume
Overall	50.224	61.043	54.594	115.437	184.715

I Reproducibility and Data Roles

All reported gains use complete matched trajectories or compact trajectory summaries. DIV2K and FSDD use different task-specific losses, independent priors, independent group definitions, and independent bootstrap procedures. No cross-task AURC average or confidence interval is computed. We distinguish support-based selection, pilot, confirmation, and scale studies throughout. Code, protocol manifests, aggregation scripts, and figure source files accompany the submission.