MEASURING ONE-STEP VISUAL ACQUISITION: PROBE SHIFTS AND OUTCOME-SUPERVISED VALUE

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual-acquisition decisions depend on the quantity used to value an observation. We study one-step forced-LOOK region selection from the same saved multimodal reasoning state, comparing two offline measurements over 35 candidate crops: RIG, a realized posterior-shift score given by the KL divergence of a fixed four-choice answer probe, and empirical Visual Intervention Effect (VIE), the change in expected continuation reward after inserting a crop. Discovery measurements select actions; independent fresh continuations evaluate the frozen choices. On a historical 120-question, 360-state Qwen3-VL-32B-Instruct cohort, VIE selections obtain 96.25% fresh stored-key agreement versus 89.31% for RIG, a +6.94 percentage-point difference (95% question-cluster CI [4.10,10.28]). A same-task Mistral-Small-3.2-24B-Instruct-2506 replication yields +10.56 points [6.39,14.86]. A separate 48-question, 30-cluster exploratory follow-up finds VIE–RIG gaps of +26.22, +23.44 and +25.52 points across three frozen interfaces, with all three Bonferroni-adjusted intervals excluding zero. Task-construction audits reveal answer-key shortcuts, and a later parent cohort did not pass the prespecified no-image admission criterion; we therefore restrict the main claim to the frozen measurement procedures and treat the interface follow-up as exploratory. Empirical VIE is an offline outcome-supervised reference rather than a deployable acquisition policy, and our critic experiments do not establish successful amortization. Within these boundaries, the supported result is a measurement-level distinction: posterior shift and downstream acquisition value are not interchangeable criteria, and under the evaluated states, actions, continuation procedures, and constructed rewards they induce region choices with substantially different fresh outcomes.

1 INTRODUCTION

Multimodal reasoning increasingly involves acquiring additional evidence, such as inspecting a chart label, a document field, or a localized geometric relation. Visual search and pixel-space operations provide mechanisms for such acquisition (Wu & Xie, 2024; Su et al., 2025), whereas interleaved reasoning introduces visual evidence into intermediate computation (Chen et al., 2025; Li & Ma, 2026). We investigate how the quantity used to score candidate observations shapes acquisition decisions.

Classical information gain and decision-theoretic value of information distinguish uncertainty reduction from downstream utility (Lindley, 1956; Howard, 1966). Applying this distinction to a model requires two operational choices: which answer readout represents the model's beliefs, and which reward measures task success. We therefore ask whether a fixed forced-choice posterior-shift score, termed realized information gain (RIG), and an outcome-supervised crop score select different regions from the same saved reasoning state.

Our primary comparison concerns *one-step forced-LOOK region selection*: conditional on taking one additional observation from a fixed candidate set, we ask how the ranking measure changes the selected region and its fresh downstream outcome. We fork the same saved state with and without a crop. Their expected continuation-reward difference defines VIE. Discovery selects actions, and independent fresh continuations evaluate them; NOLOOK and a gray placebo are separate historical controls.

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.

Empirical VIE provides an outcome-supervised offline reference for region selection. Predicting value before acquisition is a separate problem.

The historical measurement cohort contains 120 constructed questions. Qwen3-VL-32B-Instruct (Bai et al., 2025) and Mistral-Small-3.2-24B-Instruct-2506 (Mistral AI, 2025) each generate their own states on these tasks. A later 192-question parent cohort failed the pre-specified no-image admission rule for clean-task confirmation. After that decision, a result-independent rule selected 48 questions for an exploratory interface study, before policy-comparison outcomes were inspected. The evidence map summarizes the distinct roles of these studies. Their outcomes are analyzed separately, and the exploratory follow-up does not inherit confirmatory status.

Evidence map.

Evidence	Question addressed	Status and interpretation boundary
Historical Qwen	Do frozen selectors yield different fresh outcomes?	Original measurement; constructed reward
Same-task Mistral	Is the contrast Qwen-specific?	Second backbone; same task-validity limits
48-question follow-up	Does it persist across tested interfaces?	Exploratory; parent cohort did not meet clean-confirmation admission criterion
Critic studies	Can tested predictors amortize VIE?	Superiority not established; separate deployment question

Our contributions are: (i) a same-state protocol for comparing one-step region-value measurements with discovery separated from fresh evaluation; (ii) evidence of different fresh outcomes for posterior-shift and outcome-supervised selections, with same-task second-backbone replication and exploratory interface follow-up; and (iii) audits of task-reward validity, readout dependence, finite-rollout ties, stopping, cost, and amortization behavior. Section 5 reports the primary frozen-selection comparison, its robustness analyses, and the distinction between the procedural finding and its semantic interpretation.

2 Related work

Active and interleaved visual reasoning. Active perception frames sensing as part of decision-making (Bajcsy, 1988). Recent multimodal systems instantiate this principle through several visual-acquisition and evidence-integration mechanisms. V* uses guided visual search to answer questions that require fine-grained image evidence (Wu & Xie, 2024). Pixel Reasoner enables reasoning through pixel-space operations (Su et al., 2025), whereas MINT-CoT interleaves selected visual tokens with mathematical chain-of-thought reasoning (Chen et al., 2025). Learning When to Look examines strategic visual perception (Yang et al., 2026), and AIM-CoT combines active visual probing with dynamic evidence insertion (Li & Ma, 2026). ZoomEye, FOCUS, and FOVEA further investigate adaptive visual exploration (Shen et al., 2025; Zhong et al., 2025; Liu et al., 2026). Collectively, these approaches motivate visual acquisition as a component of multimodal reasoning. Our contribution is to compare acquisition-value targets under a fixed crop-acquisition protocol.

Information-based evidence selection. Metareasoning evaluates computations according to their contributions to subsequent decisions (Hay et al., 2012). Luo et al. (2026) formulate visual evidence utility for multimodal retrieval-augmented generation using output-distribution information gain and a tractable helpfulness surrogate. Recent analyses further examine context-dependent visual evidence utility and visual insensitivity (Fang & Wang, 2026; Zhang, 2026). These studies highlight the role of information-based criteria in selecting visual evidence. However, the retrieval setting of Luo et al. (2026) selects among retrieved evidence, whereas our setting generates candidate observations by cropping a fixed image. The evaluation target also differs: we compare a realized posterior-shift statistic with measured downstream success. We evaluate the implemented forced-choice RIG probe through its selected regions and their downstream continuations.

Belief readouts and outcome supervision. Prompt calibration, answer ordering, and multimodal prompting can affect elicited predictions (Zhao et al., 2021; Zheng et al., 2024; Ismithdeen et al.,

2025). Consequently, our RIG score is conditional on its fixed forced-choice readout. VisualPRM studies process reward modeling for multimodal reasoning (Wang et al., 2025). By contrast, VIE measures a same-state correctness contrast between acquisition and no acquisition.

Efficient visual computation. VisionZip reduces redundancy in visual token sequences (Yang et al., 2025). However, compressing already available tokens and selecting a new observation address different computational decisions and costs. In our setting, each LOOK contributes the same number of visual tokens, which controls the acquired token count and isolates region value.

3 TASK VALUE AND REALIZED INFORMATION GAIN

State and interventions. Let I be a static image, Q a question, and A^* its designated answer. A state $s_t = (Q, h_{\le t}, E_t)$ contains the reasoning prefix and the available visual evidence. In the present one-step experiments, E_t contains the initial global view. The early, middle, and late checkpoints therefore do not constitute a learned multi-acquisition rollout. The frozen region set $\mathcal{R}(I)$ contains $K = 35$ candidate crops. Executing $\mathrm{LOOK}(r)$ appends observation o_r and then continues generation to a final answer. STOP denotes making no further visual acquisition rather than terminating textual reasoning; the corresponding continuation is the NOLOOK control.

For a fixed backbone and decoding procedure, let $V(s)$ denote the expected final-answer reward under NOLOOK. Correspondingly, $V(s \oplus o_r)$ denotes the expected reward after inserting crop observation o_r. We define

$$\mathrm{VIE}(s, r) = V(s \oplus o_r) - V(s), \qquad V(s) = \mathbb{E}[R(A, A^*) \mid s, \mathrm{NOLOOK}], \tag{1}$$

where $R(A, A^*) = \mathbf{1}[A = A^*]$. The stored answer key defines this reward and represents correctness only when both the key and the answer options are valid. Historical accuracy figures therefore denote agreement with the stored keys. The expectation is taken over stochastic continuations. "Counterfactual" refers to computational forks originating from the same saved state and does not imply real-world causal identification. A negative VIE indicates that the intervention reduces expected reward under the specified continuation procedure. A zero VIE indicates no expected reward change, but not necessarily semantic redundancy.

Region selection versus stopping. The task-optimal region under fixed acquisition is $r^*(s) \in \arg\max_r \mathrm{VIE}(s, r)$. If STOP is available and the objective is correctness alone, the one-step action rule is

$$a^*(s) = \begin{cases} \mathrm{LOOK}(r^*(s)), & \max_r \mathrm{VIE}(s, r) > 0, \\ \mathrm{STOP}, & \max_r \mathrm{VIE}(s, r) \le 0. \end{cases} \tag{2}$$

When net utility is defined as $\mathrm{VIE}(s, r) - \lambda C(r)$, the stopping comparison must also account for acquisition cost. Equal LOOK costs preserve the ranking among regions but do *not* make cost irrelevant to the STOP decision. Our original frozen comparison evaluates region selection, whereas the learning pilot includes STOP.

Realized information gain. A read-only forced-choice probe returns normalized probabilities $p_s(a)$ over four candidate answer labels, where $a \in \{A, B, C, D\}$. After acquiring region r, the same probe returns $p_{s,r}(a)$. We define

$$\mathrm{RIG}(s, r) = D_{\mathrm{KL}}(p_{s,r} \| p_s) = \sum_a p_{s,r}(a) \log \frac{p_{s,r}(a)}{p_s(a)}. \tag{3}$$

RIG is our legacy name for a posterior-shift statistic computed under a fixed probe interface. It is a posterior-to-prior KL divergence for a *realized* observation, not an expectation over unobserved crops. Because its computation requires the crop, RIG is an offline diagnostic in the present formulation. It also differs from entropy reduction, $H(p_s) - H(p_{s,r})$, which we retain as an exploratory

baseline. We do not assume that the probe distribution equals the distribution of final answers produced by a full reasoning continuation.

KL divergence is non-negative and measures distributional change, whereas VIE is signed and measures reward change. Consequently, the two quantities need not rank candidate actions identically. Even under a deterministic answer rule, probabilities can change without changing the selected answer, producing positive KL divergence but no correctness gain. When a full reasoning continuation follows acquisition, subsequent computation can further affect the final outcome. This distinction motivates the empirical comparison between the specified RIG probe and empirical VIE selection.

4 MEASURING COUNTERFACTUAL VISUAL VALUE

Discovery estimator. For each saved state, we obtain $M_0 = 8$ NOLOOK continuations once and reuse their mean across all candidate regions. Each region receives $M = 4$ LOOK continuations:

$$\widehat{\mathrm{VIE}}(s, r) = \frac{1}{4} \sum_{m=1}^{4} R_{s,r,m}^{\mathrm{LOOK}} - \frac{1}{8} \sum_{j=1}^{8} R_{s,j}^{\mathrm{NOLOOK}}. \tag{4}$$

The LOOK and baseline continuations use independent seed namespaces. This design yields a shared-control estimator, *not* a paired-seed variance-reduction estimator. Reusing the same baseline ensures that all within-state comparisons use a common reference. It also requires $35 \times 4 + 8 = 148$ discovery continuations per state. The 360-state cohort therefore requires 53,280 discovery continuations. The resulting discrete, finite-sample labels are estimates rather than exact intervention values.

Because the baseline is shared, it cancels when computing within-state region rankings and regret. However, its estimation noise still affects the sign of VIE and therefore influences quadrant membership and STOP diagnostics. These quantities consequently have different measurement sensitivities.

Frozen selection and fresh evaluation. The empirical VIE selector maximizes $\widehat{\mathrm{VIE}}$ over the 35 regions; the RIG selector maximizes Eq. (3). Ties are resolved using a fixed canonical region-ID rule. Importantly, both selectors always choose a LOOK in the original frozen experiment, even when all empirical gains are non-positive. After freezing the selected actions, we evaluate them, NOLOOK, and the placebo using four fresh seeds per unique state–treatment pair. Policies that select the same treatment share its evaluation. The evaluation seeds are separate from the discovery seeds, and no fresh outcome is used for action selection. This separation limits reuse-of-noise bias in downstream accuracy. However, discovery regret remains defined relative to a noisy empirical maximum.

Cohort and measurement controls. The historical cohort contains 120 questions, with 30 from each of four domains: chart reasoning, document OCR, synthetic geometry, and synthetic visual search. Records from ChartQA (Masry et al., 2022) and DocVQA (Mathew et al., 2021) are adapted into four-choice questions. The resulting tasks constitute a development study rather than an evaluation of the original benchmarks under their standard answer protocols. New images are selected independently of the earlier development cohort, and each question contributes three saved checkpoints. The frozen backbone is Qwen3-VL-32B-Instruct (Bai et al., 2025). Every candidate uses the same B35 grid, a 512×512 crop canvas, and 256 additional visual tokens. A uniform gray observation matches the crop canvas and token count, controlling for the insertion of visual input without scene evidence.

Construction validity. A subsequent audit identified 13/30 Geometry questions with integer-truncated keys and no exact correct option. Deterministic rules based only on the answer options and source type recover all 120 stored keys. This finding reveals shortcuts in the task construction but does not show that either model exploits them. We retain the original inputs, selections, and scores. Appendix O reports the fixed-input rescoring analysis and the remaining validity limitations.

Decoding uses temperature 0.7, top-p 0.9, and an initial limit of 2,048 new tokens. The frozen answer parser assigns a score of zero to invalid or missing final answers. Only genuinely truncated

Table 1: One-step fresh evaluation with frozen Qwen3-VL-32B-Instruct on 120 four-choice questions / 360 states. RIG-KL is the forced-choice posterior-shift probe. Each LOOK acquires one 256-token crop; controls are separate. Accuracy denotes stored-key agreement, including the defective Geometry keys, averaged over states and fresh continuations. Intervals resample questions, conditional on realized discovery selections.

Policy or paired contrast	Estimate	95% confidence interval
Empirical VIE LOOK	96.25%	[94.17, 97.99]%
RIG-KL LOOK	89.31%	[85.28, 92.92]%
Entropy reduction LOOK	88.54%	[83.26, 93.19]%
Attention LOOK	89.10%	[84.58, 93.20]%
Relevance LOOK	85.83%	[80.42, 90.90]%
Random LOOK	85.76%	[80.28, 90.63]%
NOLOOK	87.29%	[81.81, 92.36]%
Gray placebo	86.88%	[80.90, 92.01]%
VIE − RIG	+6.94 pp	[+4.10, +10.28] pp
VIE − NOLOOK	+8.96 pp	[+4.65, +13.68] pp
RIG − NOLOOK	+2.01 pp	[−1.88, +6.11] pp
VIE − placebo	+9.38 pp	[+4.79, +14.38] pp

generations are eligible for one replay using the same seed and a 4,096-token limit. Appendix A provides the complete source-construction procedure, grid geometry, and implementation details.

Decision metrics and uncertainty. For a region selector π, empirical region regret is

$$\widehat{\mathrm{Reg}}_{\mathcal{R}}(s,\pi) = \max_{r\in\mathcal{R}} \widehat{\mathrm{VIE}}(s,r) - \widehat{\mathrm{VIE}}(s,\pi(s)). \tag{5}$$

Material regret is strictly greater than $\epsilon = 0.25$, whereas an ϵ-optimal selection has regret at most ϵ. This threshold is a relatively coarse tolerance on the success-probability scale. Within each state, high and low RIG denote the top and bottom $\lceil 35/4 \rceil = 9$ regions, respectively. High VIE means $\widehat{\mathrm{VIE}} \geq 0.25$, whereas low VIE means $\widehat{\mathrm{VIE}} \leq 0$. "High" RIG therefore denotes a within-state rank rather than an absolute threshold for evidence strength. For the original cohort, reported 95% intervals use 10,000 question-cluster bootstrap replicates. The later 48-question study instead resamples image/question clusters for real-source tasks and four-item construction groups for synthetic tasks within domain (Section 5.4). The resampling preserves each question's checkpoints and paired policy outcomes. Regions, states, and seeds are not treated as independent questions.

5 Frozen-selection results

5.1 Primary frozen-procedure comparison

Table 1 reports 96.25% fresh stored-key agreement for empirical VIE and 89.31% for RIG, a paired difference of +6.94 percentage points [4.10, 10.28]. VIE also exceeds NOLOOK and the gray placebo under the historical constructed reward. The RIG–NOLOOK interval includes zero. Together, these comparisons establish a fresh-outcome difference between the frozen selection procedures under the specified reward and continuation protocol.

Entropy reduction on the same forced-choice readout reaches 88.54% agreement, close to RIG-KL at 89.31%. Replacing KL with entropy decrease therefore does not account for the observed VIE–RIG separation. Both scores remain conditional on the same forced-choice readout.

5.2 Task-reward validity and the procedural contrast

The historical task audit identified 13/30 Geometry questions without an exact correct option and source-conditioned option rules that recover all 120 stored keys without images (Appendix O). Accordingly, Table 1 measures differences between the frozen selection procedures under the stored-key reward.

Table 2: Exploratory 48-question (30-cluster) fresh multi-interface follow-up. Fresh stored-key agreement is in %, and paired VIE-minus-RIG differences are in percentage points. The three Bonferroni intervals are 98.33% per-contrast bootstrap percentile intervals forming a nominal 95% family. Intervals condition on realized discovery selections.

RIG interface	Agreement	VIE gap	95% CI	Bonferroni CI
P_0R_0: baseline	58.33	+26.22	[19.61,33.16]	[18.23,34.90]
P_1R_0: paraphrase	61.11	+23.44	[17.53,29.69]	[16.49,31.08]
P_0R_1: readout	59.03	+25.52	[19.44,31.77]	[18.40,33.16]

The *fixed-input reward-sensitivity diagnostic* scores every option as wrong on the 13 defective items while keeping inputs, states and selections fixed. The gaps remain +4.65 pp [2.43,7.22] for Qwen and +8.61 pp [4.93,12.43] for Mistral (Appendix O). Because the original options, states, and selections remain fixed, this diagnostic does not evaluate repaired questions.

5.3 Backbone, score and interface checks

On the same 120 tasks, Mistral independently generates its own states, discovery outcomes, selections, and fresh continuations. Empirical VIE reaches 83.68% stored-key agreement versus 73.13% for RIG, reproducing a positive VIE–RIG gap of +10.56 pp [6.39,14.86] (Appendix N). This same-task result replicates the positive procedural contrast on a second backbone.

The probe audit extends the score-functional checks beyond KL: Appendix M evaluates entropy reduction, Jensen–Shannon (JS) divergence, and total variation (TV) as alternative probe scores, alongside multiple tie rules. Full fresh-outcome comparisons are available for entropy reduction; archived fresh-action coverage is incomplete for JS/TV, so we report their selection diagnostics without imputing unobserved outcomes.

A frozen 12-question development study found that a prompt paraphrase preserved RIG selections in 10/36 states and unprefixed answer labels in 26/36. This establishes selection sensitivity, not downstream accuracy: that study collected no variant-selected fresh continuations. The separate follow-up below supplies fresh outcomes for the three frozen interfaces.

5.4 Data admission and exploratory interface evaluation

Evidence status. The 192-question parent cohort failed its pre-specified no-image admission rule: Document options-only stored-key agreement was 20/48 = 41.67% (Wilson 95% CI [28.85, 55.72]), above the 40% engineering threshold. After this decision, a result-independent hash rule selected 48 questions, with 12 per domain, independently of per-item diagnostic outcomes and before policy-comparison outcomes were inspected. We therefore report the follow-up as exploratory.

Matched comparisons. The follow-up includes 12 questions per domain, three checkpoints per question, the same 35-region action space, four LOOK and eight shared NOLOOK discovery continuations, and four fresh continuations per selected action. The four selectors are empirical VIE and three RIG interfaces: P_0R_0 (baseline prompt and space-prefixed labels), P_1R_0 (a prompt paraphrase), and P_0R_1 (baseline prompt with unprefixed labels). All select one crop; the continuation procedure is shared. Actions were frozen before fresh evaluation; Appendix P documents treatment coverage, integrity checks, and parser handling. NOLOOK is used for discovery, but no fresh NOLOOK or placebo arm is included in this follow-up.

Estimand and uncertainty. We average the 12 fresh outcomes for each question and then compute an equally weighted mean of paired VIE–RIG differences across domains. The 48 questions form 30 resampling clusters: 12 image/question clusters in each real-source domain and three four-question construction groups in each synthetic domain. A stratified cluster bootstrap with 10,000 draws preserves domain weights and paired outcomes. Table 2 reports marginal 95% intervals and three-comparison Bonferroni percentile intervals (98.33% per contrast). All intervals condition on the realized discovery selections. The three synthetic clusters per domain limit precision and bootstrap calibration.

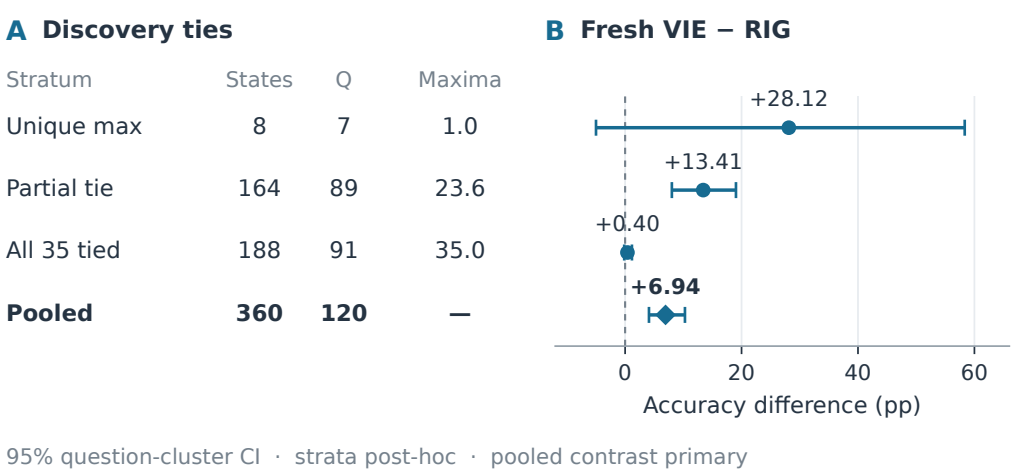

Figure 1: Discovery tie structure and frozen-selection fresh stored-key agreement on 360 states from 120 questions. **A:** State/question counts and mean number of exact empirical VIE maximizers; question counts overlap across strata. **B:** VIE–RIG gaps with 95% question-cluster intervals, conditional on realized discovery selections. Only the pooled row is the original primary comparison. The unique-max stratum has eight states and a wide interval. Partial ties contribute most of the pooled gap, but tie-breaking remains a possible contributor (Appendix K).

Fresh outcomes and interpretation. VIE achieves 84.55% fresh stored-key agreement, compared with 58.33%, 61.11%, and 59.03% for the three RIG interfaces. The respective paired differences are +26.22, +23.44, and +25.52 percentage points; all three Bonferroni-adjusted intervals exclude zero (Table 2). A positive VIE–RIG difference was therefore observed under each tested interface in this exploratory follow-up.

Domain contributions differ between the historical and follow-up cohorts. Geometry has the largest historical point difference, but its VIE–baseline RIG difference is zero in the follow-up; Document, Chart, and Search account for the new aggregate difference. Thus, the follow-up difference is not confined to Geometry. Appendix P reports the secondary interface comparisons and domain summaries.

5.5 Tie structure and randomized-max robustness

Of 360 historical states, eight have a unique empirical maximum, 164 have partial ties, and 188 tie across all 35 actions. The all-action-tied stratum has a +0.40 pp fresh gap and contributes only 0.208 pp to the pooled 6.944 pp gap; partial ties contribute 6.111 pp (Figure 1, Appendix K). Thus, states with no empirical ranking information do not account for most of the observed difference. Partial ties average 23.57 maximizers, reflecting the coarse resolution of the finite-rollout estimator.

To test whether the historical gap is an artifact of the canonical tie rule, we additionally froze a randomized-max follow-up before collecting a new outcome bank (Appendix L). Sampling uniformly from the empirical VIE-max set yields a +6.81 pp gap over RIG, with a family-adjusted interval [3.45,10.71]. It also exceeds uniform sampling over all regions by +10.63 pp [5.76,16.44] and a scale/position-matched sampling control by +6.22 pp [3.04,10.05]. All three Bonferroni-adjusted intervals exclude zero. In this new outcome bank, sampling from the empirical VIE-max set outperforms the tested controls, so the original lexicographic tie-break alone does not explain the historical fresh-outcome difference. The follow-up was designed after the historical analysis but frozen before new outcomes were collected. It remains conditional on the original tasks and discovery measurements and does not establish stability across independent discovery banks.

Under the historical shared oracle-STOP gate the gap remains +4.38 pp [2.01,7.29]. Further diagnostics appear in Appendices B and G–J; Appendix K provides outcome-selected qualitative illustrations.

6 AMORTIZATION IS A SEPARATE PREDICTION PROBLEM

VIE is measured from post-acquisition outcomes; predicting it before acquisition is a separate statistical problem. Step 1A yields identical 90.625% fresh agreement for VIE and validation-VIE-calibrated RIG on 24 test questions (gap 0.00 pp $[-2.78, 3.47]$); Step 1B, a matched VIE–RIG critic comparison, did not meet the pre-specified matched-comparison criterion.

Target precision was directly investigated in Step 1C: training labels increase from 4/8 to 8/16 LOOK/NOLOOK rollouts, and all three critic arms are evaluated against the same 8/16 reference on 48 new Chart questions. The HP-minus-original VIE regret point estimate favors the higher-precision target (-0.01563, 95% CI $[-0.03762, 0.00405]$), but the confidence interval crosses zero and therefore does not meet the pre-specified primary confidence-bound criterion. The experiment provides directional rather than conclusive evidence of a precision benefit. Appendices D–E retain the matched recipe, three seeds, 42,624 evaluation continuations and secondary RIG comparison.

7 DISCUSSION AND CONCLUSION

This study compares acquisition-value measurements under specified one-step protocols, rather than ranking information gain and value of information in general.

The historical reward measures stored-key agreement on constructed four-choice tasks, whose validity is limited by the Geometry-option defects and recoverable key structure. The interface follow-up remains exploratory because its parent cohort did not pass the prespecified no-image admission criterion. In addition, the experiments consider a fixed region set and one additional crop, the follow-up lacks a fresh NOLOOK control, and empirical VIE requires repeated outcome supervision. These conditions limit claims about semantic visual evidence, absolute acquisition benefit, repeated-discovery stability, and deployable efficiency.

Although empirical VIE is too expensive to use directly as an inference-time acquisition rule, it remains useful as an offline outcome-supervised reference. Under a fixed task reward and continuation protocol, it provides a criterion against which cheaper acquisition scores can be diagnosed: a proxy such as RIG can be evaluated by whether the regions it favors lead to improved downstream outcomes, rather than merely large shifts in a particular belief readout. In this sense, the role of empirical VIE is analogous to an expensive reference measurement rather than a deployable policy. Amortizing this reference into an inexpensive pre-acquisition critic is a separate learning problem and remains open; our pilot critic experiments did not establish such a deployable advantage.

Within these boundaries, the historical comparison, same-task backbone replication, and exploratory interface follow-up support a consistent measurement-level finding: the evaluated posterior-shift and outcome-supervised scores lead to different one-step region choices and fresh stored-key outcomes.

AI USE STATEMENT

Generative AI tools assisted with organizing the manuscript, editing English prose, locating and formatting references, and preparing LaTeX tables, figures, and reproducibility notes. They also provided code and analysis support for selected post-hoc diagnostics and the frozen Mistral replication. Experimental values were taken from author-provided reports and recorded model outputs rather than generated by the language model. The authors will verify all AI-assisted text, code, analyses, citations, tables, and figures before submission and take responsibility for the final content.

ETHICS STATEMENT

This work studies model behavior in controlled visual-acquisition tasks using adapted source data and programmatically generated images. The experiments do not recruit or interact with human participants and do not evaluate deployment decisions. Some document images may contain identifiable or sensitive information, and source datasets may restrict redistribution. Dataset licenses, provenance, and potentially identifiable natural-image content will be reviewed and documented before release. Sensitive or restricted examples will not be redistributed. The reported results do not support deployment in safety-critical settings.

REPRODUCIBILITY STATEMENT

Sections 4, 5.4, and 6 specify the estimator, the separation between discovery and evaluation, and the critic inputs. Appendices A–E document the experimental protocols and training procedures. An anonymous numerical-reproduction and experimental-source package has been prepared and validated locally but has not been publicly released. The validation covers numerical reconstruction, code dry runs, and the Mistral recorded-output and native-input audits (Appendix N), but not a complete GPU rerun. For the 48-question follow-up, the reported point estimates, marginal and family-adjusted intervals, and paired interface contrasts were reconstructed from archived per-question summaries. The action-freeze hashes and recorded discovery and fresh audit statuses were also checked. This validation did not independently reconstruct every raw continuation (Appendix P). Subject to data permissions and author verification, the final release will include or document the training manifests, prompt templates, data generators, decoding settings, optimizer hyperparameters, evaluation scripts, and per-example predictions required to reproduce the reported comparisons.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025. doi: 10.48550/arXiv.2511.21631. URL https://arxiv.org/abs/2511.21631.

Ruzena Bajcsy. Active perception. *Proceedings of the IEEE*, 76(8):966–1005, 1988.

Xinyan Chen, Renrui Zhang, Dongzhi Jiang, Aojun Zhou, Shilin Yan, Weifeng Lin, and Hongsheng Li. MINT-CoT: Enabling interleaved visual tokens in mathematical chain-of-thought reasoning. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 69110–69139. Curran Associates, Inc., 2025. doi: 10.52202/085713-2324. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/6403d7e0824e7be2e699f47d9703ed80-Abstract-Conference.html.

Yunxuan Fang and Xinhe Wang. Conditional visual evidence utility: State-dependent rank reversals in frozen vision-language encoders, 2026. URL https://arxiv.org/abs/2608.28316.

Nicholas Hay, Stuart Russell, David Tolpin, and Solomon Eyal Shimony. Selecting computations: Theory and applications. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, pp. 346–355. AUAI Press, 2012.

Ronald A. Howard. Information value theory. *IEEE Transactions on Systems Science and Cybernetics*, 2(1):22–26, Aug 1966. ISSN 2168-2887. doi: 10.1109/TSSC.1966.300074.

Mohamed Insaf Ismithdeen, Muhammad Uzair Khattak, and Salman Khan. Promptception: How sensitive are large multimodal models to prompts? In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 23950–23985, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.1302. URL https://aclanthology.org/2025.findings-emnlp.1302/.

Xiping Li and Jianghong Ma. AIM-CoT: Active information-driven multimodal chain-of-thought for vision-language reasoning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 26656–26681, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.1227. URL https://aclanthology.org/2026.acl-long.1227/.

D. V. Lindley. On a Measure of the Information Provided by an Experiment. *The Annals of Mathematical Statistics*, 27(4):986 – 1005, 1956. doi: 10.1214/aoms/1177728069. URL https://doi.org/10.1214/aoms/1177728069.

Anjie Liu, Ziqin Gong, Yan Song, Yuxiang Chen, Xiaolong Liu, Hengtong Lu, Kaike Zhang, Chen Wei, and Jun Wang. The perceptual bandwidth bottleneck in vision-language models: Active visual reasoning via sequential experimental design, 2026. URL https://arxiv.org/abs/2605.01345.

Weiqing Luo, Zongye Hu, Xiao Wang, Zhiyuan Yu, Haofeng Zhang, and Ziyi Huang. Utility-oriented visual evidence selection for multimodal retrieval-augmented generation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 35091–35124, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.1620. URL https://aclanthology.org/2026.acl-long.1620/.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.177. URL https://aclanthology.org/2022.findings-acl.177/.

Minesh Mathew, Dimosthenis Karatzas, and C.V. Jawahar. DocVQA: A dataset for VQA on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2200–2209, January 2021. URL https://openaccess.thecvf.com/content/WACV2021/html/Mathew_DocVQA_A_Dataset_for_VQA_on_Document_Images_WACV_2021_paper.html.

Mistral AI. Mistral Small 3.2 24B Instruct 2506: Model card. https://huggingface.co/mistralai/Mistral-Small-3.2-24B-Instruct-2506, 2025. Revision 95a6d26c4bfb886c58daf9d3f7332c857cb27b43.

Haozhan Shen, Kangjia Zhao, Tiancheng Zhao, Ruochen Xu, Zilun Zhang, Mingwei Zhu, and Jianwei Yin. ZoomEye: Enhancing multimodal LLMs with human-like zooming capabilities through tree-based image exploration. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 6602–6618, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.335. URL https://aclanthology.org/2025.emnlp-main.335/.

Alex Su, Haozhe Wang, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel Reasoner: Incentivizing pixel space reasoning via curiosity-driven reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 8222–8251. Curran Associates, Inc., 2025. doi: 10.52202/085713-0280. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/0c38f54740062529aa4117a04b583f3c-Abstract-Conference.html.

Weiyun Wang, Zhangwei Gao, Lianjie Chen, Zhe Chen, Jinguo Zhu, Xiangyu Zhao, Yangzhou Liu, Yue Cao, Shenglong Ye, Xizhou Zhu, Lewei Lu, Haodong Duan, Yu Qiao, Jifeng Dai, and Wenhai Wang. Visualprm: An effective process reward model for multimodal reasoning, 2025. URL https://arxiv.org/abs/2503.10291.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13084–13094, June 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Wu_V_Guided_Visual_Search_as_a_Core_Mechanism_in_Multimodal_CVPR_2024_paper.html.

Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. VisionZip: Longer is better but not necessary in vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19792–19802, June 2025. URL https://openaccess.thecvf.com/content/

CVPR2025/html/Yang_VisionZip_Longer_is_Better_but_Not_Necessary_in_Vision_Language_CVPR_2025_paper.html.

Siqi Yang, Zilve Gao, Haibo Qiu, Fanfan Liu, Peng Shi, Zhixiong Zeng, Qingmin Liao, and Lin Ma. Learning when to look: A disentangled curriculum for strategic perception in multimodal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 9738–9748, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026F/html/Yang_Learning_When_to_Look_A_Disentangled_Curriculum_for_Strategic_Perception_CVPRF_2026_paper.html.

Genpei Zhang. The visual insensitivity gap: Diagnosing when vision-language models fail to use visual evidence, 2026. URL https://arxiv.org/abs/2609.00868.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12697–12706. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/zhao21c.html.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 19426–19454, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/54dd9e0cff6d9214e20d97eb2a3bae49-Paper-Conference.pdf.

Liangyu Zhong, Fabio Philipp Rosenthal, Joachim Sicking, Fabian Hüger, Thorsten Bagdonat, Hanno Gottschalk, and Leo Schwinn. Focus: Internal mllm representations for efficient fine-grained visual question answering. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 96498–96535. Curran Associates, Inc., 2025. doi: 10.52202/085713-3227. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/8b02cdb04bc6351d36565b5c874bb804-Paper-Conference.pdf.

A INTERVENTION AND DATA PROTOCOL

Evidence scope. The manuscript reports the historical 120-question oracle comparison, the initial critic pilot, the later matched and independent Chart studies, and the 48-question exploratory multi-interface follow-up. We refer to the original frozen comparison as the historical measurement cohort; this designation does not imply that it passed the later task-validity audit. The earlier 50-question diagnostic study is reported separately. Full-scale, sequential, and multi-budget experiments remain proposed rather than completed. For the oracle and initial critic studies, per-file SHA256 checks verified the synchronized experiment directories and source code. A local audit reconstructed parser outcomes, rewards, labels, selected-action outcomes, and fresh question-cluster bootstrap intervals. Later studies rely on final reports and server measurement audits. This writing audit did not independently reconstruct every Step 1C raw continuation or rerun model inference or training. The anonymous numerical package has been prepared locally but is not yet public.

Source construction. The historical cohort comprises 120 questions, with 30 from each of four domains: chart reasoning, document OCR, synthetic geometry, and synthetic visual search. Source records were grouped by image and ordered by a stable hash. Previously used image SHA256 values were excluded, and one question per selected image was chosen deterministically. The frozen candidate generator assembled each correct answer and its distractors into four labeled options. Consequently, accuracy on these adapted tasks is not standard ChartQA or DocVQA benchmark accuracy. The remaining 60 images comprise 30 geometry and 30 visual-search items generated from a new index range. This range ensures distinct instances but not necessarily distinct template families. The frozen manifest records 120 distinct image SHA256 values.

Checkpoints. Each question contributes early, middle, and late saved reasoning prefixes produced with the initial 224×224 global image. The preparation code uses the first three decoded `</step>` boundaries; if fewer are available, its fallback uses 30%, 60%, and 85% of the trajectory tokens. The frozen manifest contains 345 step-boundary checkpoints and 15 fallback checkpoints. Each treatment preserves the question, reasoning prefix, and low-resolution context; only the added observation changes. Acquisition-value estimates are therefore conditional on these saved prefixes. Exact prefix text, token IDs, and checkpoint indices are archived because early, middle, and late labels alone are insufficient for replication.

The B35 grid. Coordinates are normalized image-space (x_1, y_1, x_2, y_2). The grid contains 16 non-overlapping quarter-width tiles, four half-width quadrants, nine quarter-width tiles shifted by one eighth of image width/height, five additional half-width overlapping tiles, and one central crop $[0.125, 0.125, 0.875, 0.875]$. The five additional half-width boxes are $[0.25, 0, 0.75, 0.5]$, $[0, 0.25, 0.5, 0.75]$, $[0.25, 0.25, 0.75, 0.75]$, $[0.5, 0.25, 1, 0.75]$, and $[0.25, 0.5, 0.75, 1]$. All methods share the same canonical region IDs. Ties in maximization choose the lexicographically greatest region ID. RIG quantiles use ascending $(\mathrm{RIG}, \mathrm{ID})$ sorting and take the first/last nine entries.

Read-only belief probe. The prompt is: "Based on the evidence available so far, if you had to answer now, choose one option: A / B / C / D. Answer:" The configured candidate strings are space-prefixed A, B, C, and D, with Qwen2TokenizerFast token IDs 362, 425, 356, and 422. Candidate conditional log likelihoods are normalized over the four labels, without length normalization. This auxiliary branch does not append its forced choice to the reasoning continuation. The offline RIG selector evaluates every candidate crop; the learned RIG critic does not receive these crops.

Seeds and measurement repair. Seeds are derived by SHA256 from question ID, checkpoint ID, namespace, and rollout index; the first eight bytes are reduced modulo 2,147,483,647. Discovery LOOK, shared NOLOOK, and fresh evaluation use separate namespaces. Fresh treatments at a state use common seed indices, and coincident policy selections are evaluated once. This sharing is distinct from claiming matched LOOK–NOLOOK random-number variance reduction in Eq. (4). The parser taxonomy was fixed before scientific analysis. Missing or invalid final answers retain zero reward. Truncation repair changes only the token cap and preserves the original raw output. A same-cost gray placebo uses RGB $(127, 127, 127)$ and the same 512×512 canvas.

The historical fresh-evaluation bank contains 2,656 distinct state–treatment pairs. After the permitted repair, 18 discovery continuations and one fresh continuation still reached the 4,096-token cap.

Table 3: Reconciled measurement counts from the final raw outputs. Missing-final counts include terminal repair-cap exhaustion, as specified by the frozen repair rule. The pilot row uses only the final three-policy subset.

Stage	Continuations	Valid final	Invalid final	Missing final
Historical discovery bank	53,280	52,866	6	408
Historical fresh bank	10,624	10,541	0	83
Critic pilot fresh	504	500	0	4

Table 4: Post-hoc descriptive fresh breakdown of the original measurement cohort. Q/S: questions/states; accuracies in %, paired VIE−RIG gaps and question-cluster intervals in pp. All four domains and three checkpoints are shown; intervals condition on realized selections (Appendix B.2).

Group	Q/S	VIE	RIG	NOLOOK	Δ [95% CI]
Chart	30/90	88.89	81.94	81.67	6.94 [3.33, 11.11]
Doc	30/90	98.33	92.78	95.00	5.56 [0.00, 13.06]
Geometry	30/90	98.33	83.33	72.50	15.00 [6.94, 24.17]
Search	30/90	99.44	99.17	100.00	0.28 [-1.11, 1.94]
Early	120/120	95.83	87.50	85.62	8.33 [4.58, 12.50]
Middle	120/120	95.42	89.17	89.17	6.25 [2.29, 10.62]
Late	120/120	97.50	91.25	87.08	6.25 [2.70, 10.62]

The frozen procedure records these terminal responses as missing-final failures with zero reward, and they remain in the accuracy denominators. Thus, "zero unresolved truncations" indicates completion of the repair protocol, not the absence of cap-exhausted responses. Excluding failures after observing their outcomes would change the estimand.

B Diagnostic definitions and supplementary results

B.1 Post-hoc diagnostics of domains, ties, and stopping

The domain and checkpoint decompositions and the shared-STOP analysis are post-hoc descriptive diagnostics. The randomized-max follow-up was designed after inspection of the original results, but its protocol was frozen before collection of the new outcome bank. All analyses remain conditional on the original questions and discovery measurements.

Domain and checkpoint decomposition. Geometry showed the largest point difference (+15.00 pp). Search showed a +0.28 pp difference with an interval crossing zero and 100% NOLOOK agreement (Table 4). Excluding Geometry yielded a +4.26 pp difference [1.94,7.13]. This descriptive deletion analysis is distinct from the full-cohort rescoring reported above. Appendix B.2 reports all domain deletions and the 12 domain–checkpoint cells. Neither the positive checkpoint differences nor these post-hoc summaries establish benefits in every domain.

Empirical regret and probe associations. Relative to the noisy empirical reference, mean RIG regret was 0.0965 [0.0639,0.1340]. Material regret occurred in 10.28% [6.39,14.44] of states, and the ϵ-optimal rate was 89.72%. Finite-rollout maximization can inflate the empirical reference. The fresh contrast instead compares frozen selectors on independent stored-key outcomes. It does not estimate regret relative to a population-optimal region.

Nonpositive VIE occurred in 92.93% of high-RIG candidates, compared with 96.66% of other candidates and 95.70% overall. The post-hoc difference was -3.72 $[-6.57, -1.25]$ pp. Thus, Table 5 does not show enrichment of nonpositive values among high-RIG crops. RIG-selected crops had nonpositive/negative discovery VIE in 91.67/9.17% of states. Supervised diagnostics found no increase in stored-correct-option probability for 77.93% of high-RIG crops. Redistribution among the other options accounted for 36.40% of the summed KL. These associations neither establish semantic irrelevance nor identify a mechanism for the fresh difference (Appendix G). Mean within-

Table 5: Frozen off-diagonal diagnostics. Region mass uses all 12,600 state–region pairs as its denominator. State presence is the fraction of 360 states containing at least one qualifying region. Both intervals cluster by question.

Region category	Region mass (95% CI)	State presence (95% CI)
High-RIG / low-VIE	23.90 [23.05, 24.61]%	99.72 [99.17, 100]%
Low-RIG / high-VIE	0.56 [0.29, 0.89]%	7.22 [4.17, 10.83]%

state RIG–VIE Spearman correlation was -0.018 and was defined in only 172/360 states; undefined correlations were not replaced by zero.

Finite-rollout ties and conditional stability. Split-half ϵ-set Jaccard was 0.9163, hypothetical STOP agreement was 97.22%, and symmetric region regret was 0.0319 [0.0194,0.0465]. Split-ranking Spearman correlation was 0.5958 for the 131/360 states in which it was defined. Full-discovery maxima were tied in 352/360 states, including 188 states with all-action ties. In the outcome-defined positive subset of 57 states from 27 questions, Jaccard decreased to 0.6538 and regret increased to 0.1447. Empirical resampling retained the selected actions in 97.14% of draws overall and 87.78% of draws in this subset. These diagnostics describe the recorded discovery bank, not stability under independent discovery runs or fresh stopping performance (Appendix I).

Shared-gate decomposition of forced LOOK. The post-hoc common gate allowed LOOK only when the original maximum empirical VIE was positive, which occurred in 57/360 states (15.83%). Reusing the original fresh outcomes yielded a +4.38 pp VIE–RIG difference [2.01,7.29]. This difference was 2.57 pp [0.97,4.58] smaller than under forced LOOK. The analysis diagnoses contributions under a common outcome-supervised gate. It is neither a replacement for the primary comparison nor a comparison of independently calibrated deployable stopping policies (Appendix J).

Randomized tie and geometry diagnostic. Four frozen draws per state from the empirical maxima, evaluated using four new continuation seeds per action, achieved 96.88% stored-key agreement. Differences relative to RIG, uniform-all sampling, and scale/position-matched sampling were +6.81/+10.63/+6.22 pp, respectively. The corresponding family-adjusted intervals excluded zero. The original lex-max minus sampled-max difference was -0.42 pp $[-1.32, 0.36]$. These finite sampled policies compare specified alternatives conditional on the original discovery measurements and stored labels. They do not independently confirm the discovery of semantically useful evidence (Appendix L).

State conditioning and illustration. Freezing the early-checkpoint region produced a later-checkpoint empirical regret of 0.0500 [0.0208,0.0865]. The selected reference region changed in 22.08% [15.00,29.58] of middle and late states, sometimes among tied regions. The change in selected regions is consistent with state dependence, but this one-step analysis does not test sequential acquisition. Figure 4 illustrates an outcome-selected case in which a blank VIE crop and NOLOOK both matched the stored key on all four fresh seeds, whereas RIG did not. Agreement in this case does not show that the crop supplied useful visual evidence. Appendix K provides the selection rule and additional cases.

Reliability. For each region, split A uses the first two LOOK rewards and first four baseline rewards; split B uses the remaining rewards. Let $v_A(r)$ and $v_B(r)$ be these half-sample gains, and r_A, r_B their deterministic maximizers. The symmetric cross-split regret is

$$\frac{1}{2}\left[\max_r v_B(r) - v_B(r_A) + \max_r v_A(r) - v_A(r_B)\right]. \tag{6}$$

For $j \in \{A, B\}$, the ϵ-optimal set is $S_j = \{r : v_j(r) \geq \max_{r'} v_j(r') - 0.25\}$. Jaccard is $|S_A \cap S_B|/|S_A \cup S_B|$. STOP agreement compares $\max_r v_A(r) \leq 0$ with $\max_r v_B(r) \leq 0$. These checks measure finite-sample decision consistency; a low cross-split regret is not a bound on population-oracle regret.

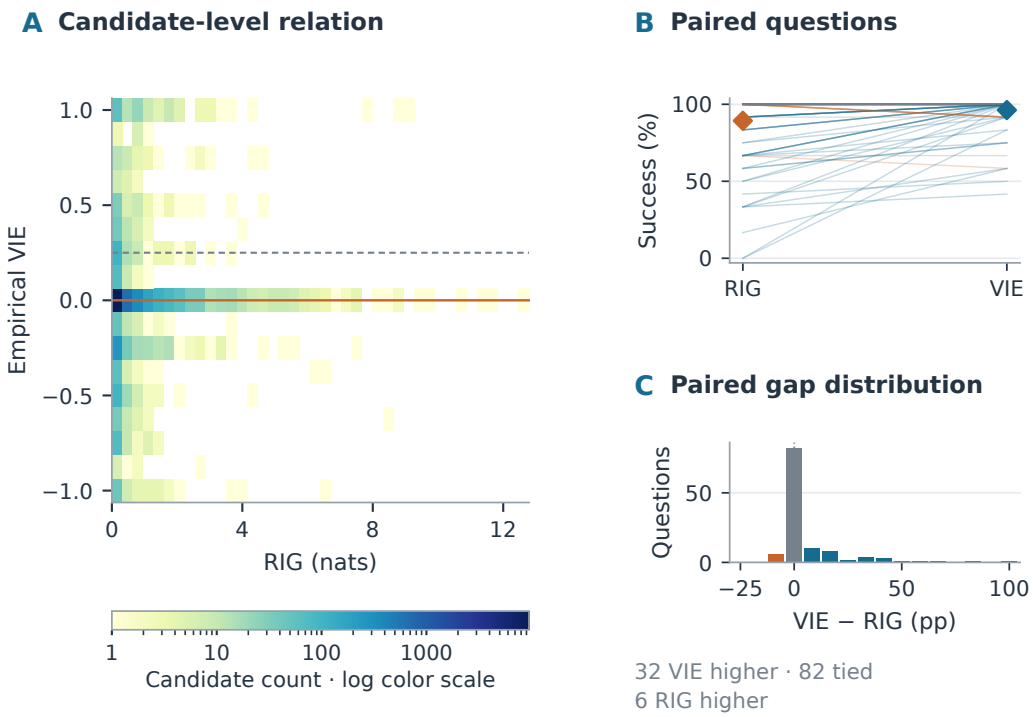

Figure 2: Post-hoc probe/value and fresh-selection views. **A:** All 12,600 candidate pairs, with logarithmic count color and exact 0.125 VIE bins; lines mark VIE 0 and 0.25, not a global high-RIG cutoff. **B:** All 120 question pairs, averaging three checkpoints per question; diamonds are cohort means. **C:** The same paired gaps, including ties and counterexamples. Within-state correlations exclude constant vectors; their question-cluster intervals include zero (Appendix K).

State conditioning. For each question, the static region is selected using only its early-state empirical values. At middle and late states, regret compares that same region to the corresponding state-specific empirical maximum. The all-checkpoint mean is 0.0333, with interval $[0.0139, 0.0569]$, while the middle/late mean is 0.0500, with interval $[0.0208, 0.0865]$. The early state's regret is zero by construction. No policy is retrospectively assembled from different checkpoints to define the static baseline.

Confidence intervals. The analysis resamples question IDs with replacement and retains all their state-level records. It then recomputes the statistic and reports the 2.5th and 97.5th bootstrap percentiles. All questions have three checkpoints, so averaging state means weights questions equally in the main cohort. Paired accuracy contrasts are bootstrapped directly; their uncertainty is not derived by subtracting independently estimated interval endpoints. The reported intervals quantify question-cluster uncertainty under the recorded observations, not all uncertainty from changing backbone, data source, or discovery labels. Secondary diagnostics are descriptive and are not presented as a multiplicity-adjusted family of hypothesis tests.

B.2 Post-hoc fresh domain and checkpoint breakdown

This post-hoc breakdown uses the original 120-question cohort, its frozen selections, and its fresh rewards. The four domains, three checkpoints, 12 cross-cells, and four leave-one-domain-out summaries were specified in a new analysis configuration before subgroup values were computed. This is a descriptive analysis, not preregistered subgroup confirmation. Source-hash checks verify the selection manifest, question–state mapping, analysis and parser settings, and selected fresh treatments. The pooled accuracies and original $+6.94$ pp contrast $[4.10, 10.28]$ were reproduced before decomposition.

Table 6: All 12 post-hoc domain–checkpoint cells. Each row contains 30 questions and 30 states; Q/S, accuracy (%), and paired gap/CI (pp) follow Table 4. All intervals cluster by question within domain.

Group	Q/S	VIE	RIG	NOLOOK	Δ [95% CI]
Chart / Early	30/30	87.50	75.00	74.17	12.50 [5.83, 19.17]
Chart / Middle	30/30	89.17	85.83	86.67	3.33 [-2.50, 10.00]
Chart / Late	30/30	90.00	85.00	84.17	5.00 [-0.83, 11.67]
Doc / Early	30/30	98.33	93.33	95.83	5.00 [-0.83, 11.67]
Doc / Middle	30/30	96.67	94.17	96.67	2.50 [-3.33, 10.00]
Doc / Late	30/30	100.00	90.83	92.50	9.17 [0.00, 20.00]
Geometry / Early	30/30	98.33	82.50	72.50	15.83 [5.00, 28.33]
Geometry / Middle	30/30	96.67	78.33	73.33	18.33 [5.83, 31.67]
Geometry / Late	30/30	100.00	89.17	71.67	10.83 [1.67, 21.67]
Search / Early	30/30	99.17	99.17	100.00	0.00 [-2.50, 2.50]
Search / Middle	30/30	99.17	98.33	100.00	0.83 [-2.50, 5.00]
Search / Late	30/30	100.00	100.00	100.00	0.00 [0.00, 0.00]

Table 7: All four post-hoc leave-one-domain-out results. Group names denote the omitted domain; each row retains 90 questions / 270 states. Accuracy is in %; paired gaps, intervals, and deletion influence I_d are in pp. Intervals preserve the three remaining domains.

Group	Q/S	VIE	RIG	NOLOOK	Δ [95% CI]	I_d
Chart	90/270	98.70	91.76	89.17	6.94 [3.61, 10.83]	0.00
Doc	90/270	95.56	88.15	84.72	7.41 [4.26, 10.83]	-0.46
Geometry	90/270	95.56	91.30	92.22	4.26 [1.94, 7.13]	2.69
Search	90/270	95.19	86.02	83.06	9.17 [5.56, 13.33]	-2.22

For each state–treatment, accuracy averages all four frozen fresh rewards, retaining invalid/missing answers as zero. Coincident policy selections share one treatment and its seeds. For question q, let d_q average its three paired VIE–RIG state differences. Domain gaps are $\Delta_d = \mathrm{mean}_{q \in d}\, d_q$; all domains contain 30 questions and 90 states. Table 4 and the 12 cells in Table 6 each recover the pooled result with their original state weights.

The analysis uses 10,000 percentile bootstrap replicates (NumPy default RNG seed 20260911). Domain and cross-cell intervals resample questions within each domain, retaining their checkpoints and paired outcomes. Checkpoint intervals use the same pooled 120-question draws. Neither seeds nor states are independent sampling units. We do not test differences between domains or checkpoints; whether individual intervals include zero does not establish a between-group difference. Search/Late has identical observed outcomes and therefore a degenerate empirical interval, which does not imply zero population uncertainty.

For each omitted domain d, $\Delta_{-d} = \mathrm{mean}_{q \notin d}\, d_q$ retains the original weights of the other 90 questions. Its interval resamples 30 questions within each of the three remaining domains, preserving that fixed domain composition. Table 7 reports all deletions and influences $I_d = \Delta_{\mathrm{all}} - \Delta_{-d}$. All four deletion point estimates are positive and their descriptive intervals exclude zero; this is conditional sensitivity evidence, not a new confirmatory superiority test.

The weakest-domain point estimate is Search's $\Delta_{\min} = +0.28$ pp. The most adverse deletion is Geometry, with $L_{\min} = +4.26$ pp and $I_{\mathrm{Geometry}} = +2.69$ pp. Geometry is also the stronger of the two fixed synthetic domains (Geometry and Search), so deleting the strongest synthetic domain gives that same row. These extrema are chosen after observing all domain results; ordinary row intervals are not selection-adjusted intervals for the extrema. The analysis does not establish gains in every domain, identify a causal mechanism, or integrate over new discovery selections.

C Earlier exploratory study

An earlier exploratory cohort, collected before the frozen 120-question study, comprised 50 questions and 150 states. Its prespecified Random Core subgroup comprised 30 questions and 90 states;

the remaining items were diagnostic cases. We report Random Core and the full cohort separately and do not pool either with the historical measurement cohort. All intervals cluster by question.

Table 8: Fresh results in the earlier 30-question Random Core. This is a separate exploratory cohort; its estimates are not pooled with the historical 120-question results in Table 1.

Region selector or control	Accuracy	95% confidence interval
Empirical VIE	95.0%	[89.2, 99.2]%
RIG-KL	86.1%	[75.8, 94.7]%
Entropy reduction	88.1%	[75.8, 98.1]%
Attention	88.9%	[78.3, 97.2]%
Relevance	84.2%	[71.9, 94.4]%
Random region	84.7%	[73.6, 94.4]%
NOLOOK	83.9%	[70.8, 95.0]%
Gray placebo	83.3%	[70.6, 94.4]%

The paired VIE–RIG difference was $+8.89$ pp, with interval $[+2.50, +17.22]$ pp. Mean RIG regret was 0.1139, with interval $[0.0472, 0.1944]$; material regret occurred in 11.1% of states, with interval $[3.3, 20.0]\%$. High-RIG/low-VIE mass was 23.49%, while low-RIG/high-VIE mass was 0.73%. In the full 50-question diagnostic cohort, the fresh VIE–RIG difference was $+6.33$ pp, with interval $[+2.33, +11.50]$ pp. This full set contains the Random Core and is not an independent replication.

In Random Core, middle/late static-query regret was 0.0708 $[0, 0.1583]$, despite a 21.7% action-migration rate. Because the interval includes zero, state conditioning remains a secondary claim even though the larger cohort yields a positive regret interval.

The study produced 22,200 discovery continuations: 22,191 valid finals and nine invalid explicit finals. Fresh evaluation produced 4,464 continuations: 4,462 valid finals and two invalid explicit finals. Invalid outcomes retained zero reward. The archived audit reports no unresolved truncations after the predefined repair procedure. These counts apply only to the 50-question exploratory study and are not substituted for the historical measurement audit.

D Critic implementation and pilot audit

D.1 Overview of amortization evidence

Predicting VIE before acquisition is a separate question from measuring it after acquisition. The initial pilot used a question/image-grouped 72/24/24 split of the historical cohort and features available before crop acquisition: saved reasoning, initial low-resolution visual features, geometry, and coverage. High-resolution crops, labels, and measured outcomes did not enter these features. The RIG comparator used validation VIE to calibrate checkpoints and STOP thresholds, so its comparison with VIE-balanced is not a target-only ablation. Later matched forced-LOOK studies removed this stopping-calibration asymmetry. Across the studies, Step 1A tied, Step 1B did not favor VIE, and both higher-precision studies favored HP-VIE in point estimates; only the within-corpus crossfit interval excluded zero. Table 9 separates these evidence levels, and Appendices D and E provide implementation details.

Step 1C's secondary HP-minus-RIG regret contrast is -0.03356 $[-0.07552, +0.00637]$ and remains descriptive under the fixed testing sequence. Improvement across three training seeds does not constitute three independent cohort replications, and the independent confidence interval includes zero. Different pilot LOOK rates also do not establish an accuracy–compute advantage because feature overhead, calibration, and matched-budget accuracy were not evaluated jointly. No further learned-policy validation is reported here.

D.2 Implementation and audit

Features and architecture. We reuse the historical 120-question corpus for critic development, with a subsequently frozen question/image-grouped split of 72 training, 24 validation, and 24 test questions. Each domain contributes 18/6/6 questions. Split assignment uses a deterministic hash

Table 9: Amortization evidence status. Intervals are 95% question-cluster intervals. Regret contrasts use different cohorts and references and must not be pooled.

Study	Observed comparison	Evidence status
Step 1A: 24 test questions	Fresh VIE / calibrated RIG: 90.625% / 90.625%; gap 0.00 pp $[-2.78, +3.47]$	Fresh tie; comparative superiority unresolved
Step 1B: 24 validation questions	Matched VIE / RIG region regret: 0.09491 / 0.06944; one of three seeds favors VIE	Pre-specified matched criterion not met
Chart crossfit: 24 diagnosed questions	HP minus original VIE regret: -0.04340 $[-0.07986, -0.00926]$	Within-corpus diagnostic support; no RIG arm
Step 1C: 48 new Chart questions	Primary HP minus original VIE: -0.01563 $[-0.03762, +0.00405]$	Primary criterion not met; directional improvement

Table 10: Critic pilot on 24 held-out questions / 72 states. Regret includes STOP with value zero, unlike region-only regret in Eq. (5). "Opt." is ϵ-optimal recall; "Pos." is LOOK-positive recall (10 states). The final fresh evaluation includes the two deployment policies and STOP; dashes denote no finalized estimate.

Policy	Regret ↓	Opt. ↑	Pos. ↑	LOOK rate	Fresh acc.
VIE-balanced	0.0816	91.67%	80.00%	29.17%	90.625%
VIE-natural	0.0747	90.28%	80.00%	50.00%	—
RIG-native	0.0938	90.28%	100.00%	100.00%	—
RIG-VIE-calibrated	0.0833	90.28%	80.00%	55.56%	90.625%
Attention	0.1007	90.28%	100.00%	100.00%	—
Relevance	0.1181	87.50%	100.00%	100.00%	—
Always STOP	0.1111	87.50%	0.00%	0.00%	88.194%

and domain information, never VIE/RIG labels. Consequently, the pilot test is held out from critic fitting, but is not a wholly new external cohort relative to the oracle analysis.

For each region, the critic receives the final-layer last-token reasoning representation h_s, bbox-pooled low-resolution visual features z_r, ten geometry values g_r, and four coverage values $c_{s,r}$. The visual representation is an 8×8 fused token grid from the initial view. No candidate high-resolution crop, answer/evidence label, RIG/VIE value, fresh outcome, or attention/relevance scalar enters the features. Coverage is zero throughout this one-step pilot and cannot establish history-aware behavior.

Separate towers project h_s and z_r to 256 dimensions. A small MLP predicts a scalar from

$$\widehat{q}_\phi(s, r) = \mathrm{MLP}\big([h', z', h' \odot z', |h' - z'|, G(g_r), C(c_{s,r})]\big). \tag{7}$$

STOP has fixed score zero. Training combines scaled Smooth-L1 regression and within-state pairwise logistic ranking, including LOOK–STOP pairs. Pairs are included only when target separation reaches a fixed margin: 0.25 for VIE and $0.25(P_{95} - P_5)$ of training labels for RIG. Architecture, optimizer, and three training seeds are matched. The exact objective and settings follow below.

Separating target, sampling, and calibration. VIE-natural and RIG-native use natural state sampling and select models with their respective target-space validation criteria. They are the closest matched comparison of target objectives, although their checkpoint objectives also differ by target. VIE-balanced samples LOOK-positive and non-positive training states 1:1. The strong RIG-VIE-calibrated baseline trains on RIG but receives *validation VIE for checkpoint and threshold selection*. Therefore VIE-balanced versus calibrated RIG is a deployment-oriented comparison, not a target-only ablation. Scores are averaged over three models, and a frozen validation threshold τ determines whether the best predicted LOOK is chosen or the policy stops.

Empirical regret trends and unresolved fresh comparison. Table 10 shows directionally lower empirical regret for VIE-natural than RIG-native (difference -0.0191). VIE-balanced also improves

the point estimate over simple baselines, but its advantage over calibrated RIG is only 0.0017. These are point-estimate comparisons, not confirmed superiority tests. Moreover, STOP alone achieves 87.50% ϵ-optimal recall: the natural distribution and generous tolerance make this metric insufficient by itself. Only 10 test states have positive empirical LOOK value.

Fresh evaluation freezes weights, thresholds, and selections before running 126 deduplicated treatments with four seeds each, totaling 504 continuations. VIE-balanced and calibrated RIG both reach 90.625%; their paired gap is 0.00 pp with interval $[-2.78, +3.47]$ pp. STOP reaches 88.194%. The two critics' different LOOK rates do not establish an accuracy–compute frontier: feature-extraction overhead, threshold sweeps, and matched-budget accuracy were not evaluated. The pilot supports further investigation of pre-acquisition utility prediction, but does not show that learned VIE outperforms learned RIG on fresh accuracy.

Split and effective sample size. The pilot contains 216/72/72 train/validation/test states and 7,560/2,520/2,520 LOOK rows, plus one virtual STOP row per state. After the label-blind split was frozen, the counts of states with $\max_r \widehat{\mathrm{VIE}} > 0$ were 36/11/10. Validation and test distributions were not rebalanced. A large number of correlated region rows must not be mistaken for a large number of independent questions. The test set has six questions per domain; its positive LOOK states occur only in chart and geometry domains.

Architecture. Each hidden/visual tower uses LayerNorm, a linear projection to 256, GELU, and dropout 0.1. Geometry $(x_1, y_1, x_2, y_2, w, h, wh, w/h, c_x, c_y)$ is projected from ten dimensions to 32 with GELU; coverage is projected from four dimensions to 16 with GELU. The interaction vector has $4 \times 256 + 32 + 16 = 1072$ dimensions. The output head is $1072 \to 256 \to 64 \to 1$, with GELU between layers and dropout 0.1 after the first hidden layer. Coverage records prior LOOK count, covered area, candidate overlap, and round index, but all four are zero here.

Objective. Let $y_{s,r}$ be the chosen training target (empirical VIE or RIG), and $\sigma = \max(\mathrm{std}(y_{\mathrm{train}}), 10^{-3})$. Regression is applied only to LOOK values; STOP is fixed at target and prediction zero and is included in the pairwise term. For unordered action pairs $\mathcal{P}_s = \{(i,j) : i < j, |y_i - y_j| \geq \delta\}$, the implemented loss is

$$\mathcal{L}_{\mathrm{reg}} = \mathrm{mean}_{s,r}\ \mathrm{SmoothL1}\left((\widehat{q}_{s,r} - y_{s,r})/\sigma, 0\right), \tag{8}$$

$$\mathcal{L}_{\mathrm{rank}} = \mathrm{mean}_{s,(i,j)\in\mathcal{P}_s} \log\left(1 + \exp\left[-\,\mathrm{sgn}(y_i - y_j)(\widehat{q}_i - \widehat{q}_j)/\sigma\right]\right), \tag{9}$$

$$\mathcal{L} = \mathcal{L}_{\mathrm{reg}} + 0.5\mathcal{L}_{\mathrm{rank}}. \tag{10}$$

The ranking loss is zero for a batch with no eligible pair. The VIE margin is 0.25. For RIG, $\delta = \max[0.25(P_{95} - P_5), 10^{-8}]$, computed only from training labels. RIG targets are not assigned artificial negative utility to force STOP superiority.

Training and model selection. The implementation uses AdamW with learning rate 3×10^{-4}, weight decay 10^{-4}, batches of eight states, and gradient-norm clipping at 1.0. The entry point defaults to at most 100 epochs and patience 10. The three seeds are 20260825, 20260826, and 20260827. Balanced training samples positive and non-positive states with replacement in equal numbers; natural training visits the original state distribution. The frozen selections record the checkpoints and ensemble thresholds in Table 11; all 12 checkpoint hashes were verified against the synchronized files.

Table 11: Frozen critic checkpoints and ensemble thresholds. Epochs follow the three seeds in ascending order. Thresholds belong to different target scales and are not directly comparable as utilities.

Variant	Selected epochs	Ensemble threshold τ
VIE-balanced	20, 10, 17	0.159315
VIE-natural	21, 8, 24	0.048630
RIG-native	5, 30, 1	0.038179
RIG-VIE-calibrated	4, 5, 4	0.471965

Validation scans score-derived stopping thresholds. VIE variants minimize VIE regret; RIG-native minimizes RIG regret; calibrated RIG minimizes VIE regret. Tie-breaking uses optimal recall and STOP agreement, with validation loss additionally used in checkpoint selection. The two RIG selectors share a training trajectory. Its patience counter resets when either selection objective improves, so training duration can depend indirectly on validation VIE. Under RIG-only patience, the three runs would have ended at epochs 15/40/11 rather than 15/40/14; the selected native checkpoints (5/30/1) remain unchanged. This coupling should be removed in a stricter replication.

Table 12: Pilot uncertainty for empirical regret. Intervals are question-cluster intervals, not training-seed standard deviations. They do not establish a significant difference between the two deployment policies.

Policy	Mean regret	95% confidence interval	STOP agreement
VIE-balanced	0.0816	[0.0069, 0.1858]	79.17%
VIE-natural	0.0747	[0.0122, 0.1632]	58.33%
RIG-native	0.0938	[0.0243, 0.1806]	13.89%
RIG-VIE-calibrated	0.0833	[0.0174, 0.1701]	52.78%
Attention	0.1007	[0.0208, 0.2049]	—
Always STOP	0.1111	[0.0174, 0.2361]	—
Relevance	0.1181	[0.0243, 0.2396]	—
Empirical RIG	0.1250	[0.0382, 0.2361]	—

Metric interpretation. Pilot regret uses $\max(0, \max_r \widehat{\mathrm{VIE}}(s, r))$ as its reference. A STOP selection has chosen value zero. LOOK-positive recall measures choosing LOOK among states whose empirical best LOOK value is positive; it does not by itself establish that the selected region was useful. The archived rank diagnostic assigns distinct ordinal ranks to tied values rather than conventional average ranks. We therefore omit it as a Spearman estimate pending tie-aware recomputation. The separately defined regret and fresh-accuracy calculations are unaffected.

Fresh outcomes and audit. VIE-balanced fresh accuracy is 90.625% (interval $[79.86, 98.26]\%$); calibrated RIG is 90.625% ($[80.90, 97.92]\%$); STOP is 88.194% ($[75.69, 98.26]\%$). After a single allowed truncation replay, the 504-continuation audit records 500 parser successes, four failures, and no final output reaching its token cap. The remaining failures come from one NOLOOK treatment and are retained as zero reward. The directory also retains 68 treatment files outside the final three-policy subset; these are preserved for provenance and excluded from the reported 126-treatment evaluation. No finalized fresh comparison for VIE-natural, RIG-native, attention, or relevance is claimed here. The intended fresh superiority criterion was not met, so the comparative deployment result is inconclusive.

E Independent Chart confirmation protocol

Matched pilot and precision diagnostics. A subsequent forced-LOOK pilot (Step 1B) used 96 training and 24 validation questions across four domains, with identical architecture, features, natural batches, optimizer, 2,000 steps, and three seeds for VIE and RIG. There was no STOP, early stopping, or threshold calibration. Mean validation region regret was 0.09491 for VIE and 0.06944 for RIG; only one seed favored VIE, and the pre-specified matched-comparison criterion was not met. Train-only follow-up measurements suggested domain-dependent decision identifiability rather than an established global representation ceiling. A Chart precision intervention improved training fit but did not establish stable gains on six validation questions. Four-fold cross-fitting on the 24 already-diagnosed Chart training questions subsequently favored higher-precision (HP) labels: mean regret difference -0.04340, 95% question-cluster interval $[-0.07986, -0.00926]$, with improvement in all folds and seeds. This was diagnostic evidence: it used a previously examined corpus and lacked a RIG arm.

Independent Chart confirmation (Step 1C). We therefore froze 48 new Chart questions (144 states), used only for evaluation. Question IDs and image hashes did not overlap the exclusion in-

Table 13: Independent Chart confirmation: paired HP-minus-reference regret, averaged over 48 question clusters and three seeds. Both intervals cross zero. The secondary contrast remains descriptive because the primary criterion was not met.

Contrast	Reference	Δ regret	95% cluster CI
Primary	Original VIE	-0.01563	$[-0.03762, +0.00405]$
Secondary	RIG	-0.03356	$[-0.07552, +0.00637]$

ventory of 660 prior questions/images, which included the entire paused development cohort. Sampling used source-identity hashes without difficulty or outcome selection. The canonical Step 1B test remained sealed; only frozen split metadata was used for exclusion. Three critics trained on the same 24 Chart training questions: original VIE labels (4 LOOK/8 NOLOOK), combined HP VIE labels (8/16), and frozen RIG labels. The three arms shared the fixed training recipe, initialization, batches, and seeds. All nine final checkpoints and action selections were frozen before evaluation measurement. All arms were scored using the same B35, 8/16 evaluation labels, totaling 42,624 continuations. This estimates empirical region regret, not an additional fresh-accuracy contrast.

Consistent direction, but confirmation criteria not met. The primary contrast tested HP versus original VIE; HP versus RIG was secondary under a fixed testing sequence. Passing required negative aggregate regret difference, improvement in at least two seeds, an upper 95% cluster confidence bound below zero, non-decreasing ϵ-optimal recall, and a negative difference after removing any single question. Both contrasts improved in all three seeds and met the recall and leave-one-question-out conditions, but neither met the confidence-bound condition (Table 13). Thus the independent study did not meet the pre-specified primary confidence-bound criterion. Seed agreement is not three independent cohort replications. The point estimates provide directional rather than conclusive evidence of a precision benefit. We retain the earlier cross-fitting result as diagnostic support and do not pool it with this confirmation cohort. The frozen analysis and measurement audit follow below.

Step 1C trained on 24 frozen Chart questions (72 states) and evaluated on 48 new questions (144 states). The six previously inspected Chart validation questions were not used as the evaluation cohort. All evaluation images were unique and disjoint from the prior exclusion inventory; source-question identity combined image hash and normalized question. The cohort was frozen before rollout, and no question was replaced after an outcome was observed.

The three arms reused the matched pilot's dual-projection MLP, pre-acquisition hidden state, low-resolution region pooling, 10-dimensional geometry, and zero coverage features. AdamW used learning rate 3×10^{-4} and weight decay 10^{-4}, with gradient norm clipping at 1, batch size 8, and exactly 2,000 steps. Seeds were 20260825, 20260826, and 20260827. Each seed shared initialization, random-number state, and all batch IDs across arms. VIE-Original and VIE-HP shared the original precision-control loss scale 0.200350821018219; the RIG scale followed the original recipe, $\max(\mathrm{std}(y_{\mathrm{RIG,train}}), 10^{-3}) = 0.8156421184539795$. Thus target-unit normalization differs, while the optimizer recipe is matched. No evaluation label selected a checkpoint, scale, or threshold. Unlike the initial STOP-inclusive pilot, these fits used only region candidates and the matched pilot's loss.

Table 14: Step 1C absolute metrics under common 8/16 evaluation labels. These region-only regrets have a different evaluation cohort and reference set from the STOP-inclusive initial critic pilot.

Arm	Mean regret	ϵ-optimal recall
VIE-Original	0.201100	76.8519%
VIE-HP	0.185475	77.5463%
RIG	0.219039	73.8426%

The matched training loss was scaled Smooth-L1 regression plus 0.5 times pairwise logistic ranking over within-state pairs with absolute target difference greater than 10^{-12}. It did not inherit the initial pilot's 0.25 training margin.

The analysis averaged the three states and paired seeds within each question, then weighted all 48 questions equally. It used 10,000 paired question-cluster percentile bootstrap replicates with seed 20260917; all arms, states, and seeds remained together under resampling. The tolerance was $\epsilon = 0.25$, and action ties used the first maximum after lexicographic action-ID sorting. HP-minus-original per-seed differences were -0.013021, -0.026910, and -0.006944; HP-minus-RIG differences were -0.037326, -0.040799, and -0.022569. Recall differences were $+0.6944$ and $+3.7037$ percentage points, respectively. These quantities do not override the failed primary confidence-bound criterion.

Measurement completed all 5,184 treatment files and 42,624 continuations. LOOK and shared NOLOOK used distinct frozen namespaces. Of 85 genuine truncations replayed once at a 4,096-token cap, 76 became valid and nine remained terminal missing-final outcomes with zero reward. The final parser audit reported zero unresolved truncations under its recorded taxonomy; the nine terminal failures remained in the evaluation denominators.

F OFFLINE POWER AND COST SCREEN

After Step 1C, this offline screen assessed resource requirements under a planning improvement of 0.010 regret, an evaluation-only cap of 100,000 continuations, and an 80% full-gate pass-probability target. These are offline planning assumptions, not a future experiment preregistration. We screened new cohort sizes $N \in \{24, 36, 48, 60, 72, 84, 96, 108, 112\}$ and improvements $\delta \in \{0, 0.005, 0.010, 0.015625, 0.020\}$. The observed effect is only a sensitivity scenario. Each new question costs 888 evaluation continuations; preparation, features, probes, repairs, and training remain unpriced.

All 48 historical questions were retained. The paired regret and recall vectors include both contrasts and all three training seeds, averaging states within each question first. We jointly resampled question vectors, centered regret at each hypothetical effect, and varied between-question standard deviation by factors 0.75, 1, and 1.5 while preserving seed offsets. Recall means were either historical or zero. Each sample size used 400 simulated cohorts and 10,000 question-cluster percentile bootstrap replicates per cohort, preserving the full gate and secondary fixed testing sequence. Location–scale equivariance reused bootstrap calculations across effects. The original two intervals and all question summaries were reproduced from frozen evaluation details before simulation.

Table 15: Offline design screen under historical question variability and recall means. Entries are full primary gate pass probabilities. Null denotes the model-conditional false-pass rate. All sample sizes denote new question clusters, not extensions pooled with Step 1C.

N	Continuations	Null	.005	.010	.015625	.020
24	21,312	3.0%	6.8%	13.5%	22.5%	35.8%
36	31,968	2.8%	5.0%	11.8%	29.0%	43.8%
48	42,624	2.2%	6.2%	15.8%	35.8%	52.8%
60	53,280	2.8%	8.5%	20.8%	43.8%	57.5%
72	63,936	1.8%	7.2%	20.8%	47.2%	65.2%
84	74,592	1.0%	6.0%	23.0%	52.2%	69.8%
96	85,248	2.2%	10.8%	29.0%	58.5%	73.2%
108	95,904	2.0%	11.5%	32.0%	63.0%	75.8%
112	99,456	2.5%	11.2%	33.8%	66.8%	78.5%

At $N = 112$ and $\delta = 0.01$, mean primary interval width was 0.0265 and Monte Carlo standard error of the pass probability was 2.4 percentage points; its Wilson interval was $[29.3, 38.5]\%$. With standard deviation multipliers 0.75 and 1.5, pass probabilities were 53.0% and 17.5%, respectively. The 0.75 multiplier applies to standard deviation, corresponding to 0.5625 times variance. The historical primary question standard deviation was 0.07274, with a 48-cluster bootstrap range $[0.04408, 0.09500]$. Under a zero recall-difference mean, the baseline pass probability at 112 questions was 33.2%. No screened design reached the 80% target at the planning effect.

These conditional summary simulations do not model new training or guarantee realizable discrete rewards. Unseen question types and effect-dependent correlations remain unidentified. Monte Carlo

intervals cover simulation error only; null rates do not establish general type-I control. The original testing gate and primary contrast were retained, and the canonical Step 1B test set remained unopened. No full Step 1B evaluation is reported.

G POST-HOC RIG MARGINAL RATES AND SUPERVISED BELIEF DIAGNOSTICS

This post-hoc analysis reuses the original 120 questions, 360 states, and 12,600 candidate crops without new model generation. Its definitions were fixed in a separate configuration before calculation, although the analysis remains exploratory. Source-file checks match the recorded inventory. Saved four-option probabilities come from `belief_probe.probabilities` in each discovery LOOK and shared-NOLOOK treatment, not the state-preparation distribution. The state manifest supplies `target_letter`; frozen selections supply the original RIG action and high-RIG ranking. Saved parsed answers reproduce all discovery rewards (four per LOOK, eight per shared NOLOOK) and selected fresh rewards (four per treatment). No new selector is fitted or ranked.

Marginal rates and uncertainty. High-RIG is exactly the last nine entries in ascending $(\mathrm{RIG}, \text{region ID})$ order per state, including the original tie rule. Non-high-RIG contains the other 26; nonpositive and negative value mean $\widehat{\mathrm{VIE}} \leq 0$ and $\widehat{\mathrm{VIE}} < 0$. The original joint frequency is 3,011/12,600, not the conditional high-RIG rate. All intervals use 10,000 NumPy default-RNG replicates (seed 20260912), resampling 120 questions with replacement and retaining their three checkpoints and every candidate. Common draws preserve paired comparisons. Conditional rates are ratios of resampled numerator and denominator sums; seeds, states and candidates are not independent resampling units. These are descriptive intervals conditional on realized discovery, with no multiplicity adjustment or new gate.

High-minus-non-high nonpositive-rate difference is -3.72 [-6.57, -1.25] pp; high-minus-all is -2.77 [-4.88, -0.93] pp. The high-RIG point rate is lower, so these data do not support enrichment of nonpositive value in the high-RIG group.

KL direction and numerical rules. The frozen selector uses $\sum_i a_i \log(\tilde{a}_i/\tilde{b}_i)$, where $\tilde{a}_i = \max(a_i, 10^{-12})$ and likewise for b_i. It retains the saved floating-point values: after-LOOK relative to before-LOOK, in nats. We reproduce that score and retain its selections. For diagnostics only, each saved entry is floored at 10^{-12} and each four-entry vector is renormalized in float64. This makes a probability partition exact and defines zero handling without retuning a selector. Entropy change is $H(a) - H(b)$; negative values indicate entropy reduction. Correct-option probability change is $a_c - b_c$; an increase means $a_c - b_c > 10^{-12}$, with the tolerance used only for numerical equality.

Supervised diagnostic, not an inference-time baseline. Let c be the ground-truth option, $a_W = \sum_{i \neq c} a_i$ and $b_W = \sum_{i \neq c} b_i$. The chain rule gives

$$D_{\mathrm{KL}}(a \| b) = D_{\mathrm{KL}}\big((a_c, a_W) \| (b_c, b_W)\big) \tag{11}$$

$$+ a_W D_{\mathrm{KL}}\big(a(\cdot \mid W) \| b(\cdot \mid W)\big). \tag{12}$$

Both terms are nonnegative as complete KL quantities; an individual option's signed $a_i \log(a_i/b_i)$ is not a nonnegative contribution. Correct-option change and this correct/wrong partition use the answer and are supervised diagnostics only. Wrong-option share divides summed second terms by summed total KL, avoiding unstable per-candidate ratios at near-zero KL. A high share does not establish that a crop is semantically irrelevant.

The saved vectors have 0 zero entries and 0 entries below the floor (counting repeated state baselines per candidate). Maximum raw sum-to-one error is 1.4e-07; maximum diagnostic-minus-frozen KL discrepancy is 1.2e-06 nats. The chain identity holds to 3.6e-15 nats. The portable numeric release includes all vectors, reward arrays, target indices, cross-tabs and unrounded results.

Probe versus continuation disagreement. Among the 360 RIG-selected states, 44/69 probability-increasing probes have nonpositive fresh RIG-minus-NOLOOK differences (63.77 [50.00, 78.79]%). The complementary conditional rate is 2.06 [0.35, 4.15]%. Only the already

Table 16: Post-hoc discovery VIE marginal rates (percent, 95% question-cluster CI). Every row spans the same 120 questions and 360 states; N counts state–region pairs. High-RIG retains the frozen nine-of-35 rank definition. RIG-selected contains one frozen action per state.

Group	N	$\Pr(\widehat{\text{VIE}} \leq 0)$	$\Pr(\widehat{\text{VIE}} < 0)$
All candidates	12,600	95.70 [94.02, 97.25]	6.96 [5.01, 9.09]
High-RIG	3,240	92.93 [89.78, 95.74]	9.38 [6.94, 12.10]
Non-high-RIG	9,360	96.66 [95.10, 98.06]	6.12 [4.20, 8.22]
RIG-selected	360	91.67 [87.22, 95.56]	9.17 [5.83, 12.78]

Table 17: Supervised diagnostic: correct-option probability change (pp) and no-increase rate (percent); entropy change is $H(a) - H(b)$ in nats and does not use the answer. Brackets are 95% question-cluster intervals; N follows Table 16. Correct-option quantities are not inference-time selector baselines.

Group	Mean Δp_c (pp)	$\Pr(\Delta p_c \leq 10^{-12})$	Mean ΔH
All candidates	-1.84 [-3.33, -0.40]	71.13 [66.12, 76.06]	0.08 [0.06, 0.10]
High-RIG	-4.63 [-7.61, -1.75]	77.93 [72.44, 83.40]	0.16 [0.13, 0.20]
Non-high-RIG	-0.88 [-1.99, 0.25]	68.77 [63.62, 73.78]	0.05 [0.03, 0.07]
RIG-selected	-7.61 [-12.18, -3.06]	80.83 [75.28, 86.11]	0.23 [0.18, 0.28]

evaluated RIG action is used here; no fresh outcome is imputed for other candidates. Complete two-by-two counts for discovery and selected fresh comparisons are in the numeric release. These comparisons connect a read-only forced-choice probe to observed full-continuation rewards, but noisy reward signs are not evidence of individual causal effects. Neither these associations nor the KL partition explain the entire $+6.94$ pp fresh selector gap, establish mediation, or replace blinded semantic annotation.

H Recorded measurement cost and accounting limits

Measurement cost. Constructing the offline measurement reference costs $35 \times 4 + 8 = 148$ discovery continuations per state, or 53,280 for 360 states. This is measurement cost, not inference cost of a proposed policy. Fresh comparison adds 10,624 continuations (2,656 distinct state–treatment pairs, four seeds each), with shared selections evaluated once. State preparation, probes, superseded attempts, repairs and engineering remain additional costs. The ledger below separates recorded call timings from costs that were not measured. Equal crop tokens match acquired input budget, not supervision or total computation.

This ledger concerns the original 120-question, 360-state oracle cohort. Counts are reconciled against frozen selections and retained treatment records; matching selections are deduplicated within each state. A continuation count is a measurement count, not a measure of FLOPs or GPU-hours. Retained outputs may replace earlier attempts, so their call times are not a complete attempted-compute ledger.

The discovery probe count is $360(35 + 1)$: one per candidate and one shared NOLOOK distribution per state, not one per continuation seed. The fresh count includes all eight reported policies/controls, not only VIE and RIG; the original frozen treatment map defines its scope. Additional files retained in the fresh directory are outside that final selected-treatment set. The preparation and checkpoint-signal timings are medians from recorded summaries and cannot be multiplied into valid total wall times. Source hashes, per-treatment numerical timing inputs and aggregate results are included in the revision ledger.

The replay counts come from completed shard reports, not from the number of invalid or missing final answers. They cover documented one-shot truncation repairs only; a complete total over all superseded measurement attempts is unknown. Full preparation/feature extraction, training, threshold searches, failed attempts, scheduling, storage, human effort and monetary cost are not fully priced. For comparison of evidence scope, the initial critic fresh subset contains 504 continuations and Step 1C's common 8/16 evaluation labels contain 42,624; neither is part of the oracle discovery count.

Table 18: Supervised diagnostic: mean KL and its nonnegative chain-rule terms (nats). Wrong-option share is the ratio of summed weighted wrong-option KL to summed total KL, in percent with question-cluster CI; it is not a mean of per-candidate ratios. The partition uses ground truth; it does not identify semantic irrelevance or a causal mediator.

Group	Total KL	Binary KL	Wrong KL	Wrong share (%)
All candidates	0.244	0.150	0.093	38.29 [31.56, 45.40]
High-RIG	0.616	0.392	0.224	36.40 [29.99, 42.91]
Non-high-RIG	0.115	0.067	0.048	41.81 [32.20, 53.31]
RIG-selected	1.328	0.846	0.483	36.35 [29.57, 43.08]

Table 19: Supervised diagnostic: disagreement with observed discovery continuation value. Left: nonpositive VIE among probability-increasing crops. Right: positive VIE among crops without a probability increase. Fractions show exact conditional denominators; rates and intervals are percent. Signs refer to finite-rollout estimates, not population effects.

Group	$V \leq 0 \mid \Delta p_c > 0$	Rate [CI]	$V > 0 \mid \Delta p_c \leq 0$	Rate [CI]
All candidates	3,296/3,638	90.60 [87.26, 93.92]	200/8,962	2.23 [1.07, 3.72]
High-RIG	539/715	75.38 [66.63, 84.64]	53/2,525	2.10 [0.97, 3.48]
Non-high-RIG	2,757/2,923	94.32 [91.39, 96.91]	147/6,437	2.28 [1.02, 3.90]
RIG-selected	45/69	65.22 [51.90, 79.31]	6/291	2.06 [0.35, 4.17]

No total-compute advantage is inferred from differing LOOK rates or these unpaired accounting components.

I FINITE-SAMPLE STABILITY, TIES, AND EXACT EMPIRICAL RESAMPLING

This post-hoc analysis reuses the original 120 questions, 360 saved states, and 53,280 discovery rewards without new model generation. The resampling rules were fixed before these diagnostic calculations. Source-hash and ordered-reward checks match the recorded inputs. The original primary comparison, decision gate, and selections remain unchanged.

Reconciled split-half denominators. The original first-two/last-two LOOK split and first-four/last-four shared-NOLOOK split reproduce all archived state metrics. Jaccard, STOP agreement and cross-split regret are defined for all 360 states. Split-ranking Spearman is defined for 131 states (both rank vectors nonconstant), not 360; undefined values are not set to zero. Full-discovery exact maxima are tied in 352 states, with 188 states tying all 35 regions. The mean number of exact maximizers is 29.04 (median 35; range 1–35); the mean full-discovery ϵ-set size is 30.40. Half A/B have tied maxima in 354/355 states. Half-sample region means lie on a 0.5 grid, so at $\epsilon = 0.25$ each half's ϵ-set equals its exact best set. Thus the high Jaccard is not evidence of uniquely identified best crops.

The two halves both indicate STOP in 303 states and both indicate LOOK in 47; A-STOP/B-LOOK occurs in four and A-LOOK/B-STOP in six. These hypothetical zero-cost decisions are label diagnostics; the primary oracle comparison always selects a region. Table 21 includes the original full-discovery LOOK-positive subset (57 states from 27 questions), defined once from the unchanged full labels. It is a selected descriptive subset, not a separate cohort or confirmation.

What the shared baseline cancels. Write $\widehat{v}_r = \bar{R}_r - \bar{R}_0$. Then $\widehat{v}_r - \widehat{v}_t = \bar{R}_r - \bar{R}_t$ and $\max_r \widehat{v}_r - \widehat{v}_t = \max_r \bar{R}_r - \bar{R}_t$: the shared NOLOOK cancels from within-state region rankings, region regret, and region-only best/ϵ sets. It does not cancel from $\widehat{v}_r \leq 0$, the hypothetical STOP test, or quadrant membership defined by absolute VIE thresholds. Resampling therefore uses one baseline value shared across all candidates, never a different independent baseline per region.

Exact empirical seed resampling. Saved seeds verify that all 35 regions within a state share the same four LOOK indices; the eight NOLOOK seeds use a separate namespace. We enumerate all $4^4 = 256$ ordered samples of four LOOK indices with replacement and apply each sampled

Table 20: Observed oracle measurement ledger. Call-hours sum the saved per-call latency fields; they are serialized call-time equivalents, not elapsed pipeline time or GPU-hours. Repair outputs are already included in final retained outcomes; replays are listed as additional attempts, not added again as observations.

Component	Verified count	Recorded timing / scope
Initial low-resolution preparation	120 completed trajectories; 360 checkpoints	9.07 s median per trajectory
Checkpoint signal measurement	360 states	2.39 s median per state
Discovery continuations	$360(35 \times 4 + 8) = 53{,}280$	176.03 call-hours, final retained outputs
Discovery belief probes	12,960 saved distributions	28.28 call-hours, saved probe calls
Fresh evaluation, all reported policies	$2{,}656 \times 4 = 10{,}624$	33.58 call-hours, final retained outputs
Documented truncation replays	53 discovery; 6 fresh	One shot per targeted rollout
Total GPU-hours / monetary cost	Unknown	Preparation, retries, probes and engineering need separate accounting

Table 21: Post-hoc split-half accounting. Entries are estimates with newly computed 95% question-cluster intervals; they do not replace the original reported interval. Each column retains fixed original subgroup membership. Exact action agreement uses the frozen lexicographically greatest region-ID tie rule.

Metric	All: 360 states / 120 questions	Positive: 57 states / 27 questions
Best / ϵ-set Jaccard	0.9163 [0.8881, 0.9412]	0.6538 [0.5458, 0.7489]
Exact action agreement (%)	91.67 [88.06, 94.72]	64.91 [50.00, 78.26]
STOP agreement (%)	97.22 [95.00, 99.17]	82.46 [68.75, 93.75]
Cross-split regret	0.0319 [0.0187, 0.0472]	0.1447 [0.0820, 0.2171]

column vector jointly to all 35 regions. For the independent baseline bootstrap, eight binary observed rewards with mean $\bar{R}_0$ imply a resampled success count $K \sim \mathrm{Binomial}(8, \bar{R}_0)$. We integrate its nine possible values exactly, using the same $K/8$ for every region in that state. There is no within-state Monte Carlo error. The comparison is always to the original full-discovery region action and sets, not to another independent model run. RIG's saved probe scores are not regenerated or assigned fictitious seed variability.

Table 22: Exact finite-bank bootstrap sensitivity (post-hoc). Each state contributes its mean across 256 joint LOOK draws; STOP agreement also integrates the independent shared-baseline bootstrap. Brackets are question-cluster intervals conditional on the observed bank, not intervals over independent repeated discovery. Fresh coverage is an availability diagnostic, not accuracy.

Metric	All: 360 states / 120 questions	Positive: 57 states / 27 questions
Original action retained (%)	97.14 [95.74, 98.32]	87.78 [81.46, 93.01]
Best-set Jaccard	0.9693 [0.9585, 0.9788]	0.8700 [0.8244, 0.9089]
ϵ-set Jaccard	0.9645 [0.9525, 0.9752]	0.8451 [0.7982, 0.8881]
STOP decision retained (%)	99.33 [98.86, 99.73]	95.78 [92.90, 98.11]
Fresh action coverage (%)	97.43 [96.16, 98.51]	88.90 [83.37, 93.61]

Aggregation and limits. For each metric we sum state contributions within each question, resample the 120 question IDs 10,000 times with NumPy seed 20260914, and divide sampled sums by sampled valid-state counts. This retains checkpoint dependence and fixed subgroup membership. All 10,000 draws have nonzero denominators. For Spearman only states with two nonconstant rank vectors contribute. The original cross-split regret interval $[0.01944, 0.04653]$ used its original seed/implementation; Table 21 labels new descriptive intervals explicitly. The independent numer-

ical audit also reproduces the 256-draw results by enumerating 35 seed-multiplicity vectors with multinomial weights.

High action retention partly reflects deterministic tie-breaking and constant observed reward vectors. Empirical resampling cannot reveal an unseen success or failure from a constant vector, so it may overstate stability under independent discovery. In particular, 2.57% of pooled resampled selections (11.10% within the original positive subset) lack outcomes in the original frozen fresh treatment set. We do not compute a resampled fresh accuracy, substitute discovery rewards, impute missing outcomes, or claim a new confidence interval covering the whole discovery procedure. The original $+6.94$ pp fresh contrast remains conditional on the realized discovery selections. Independent repeated discovery would require new seed banks and a separately fixed design; it is not established by these diagnostics.

J SHARED-ORACLE STOP DIAGNOSTIC (POST-HOC)

Estimand and gate. The original frozen comparison requires both empirical VIE and RIG to select a region. This post-hoc diagnostic retains their frozen selections, fresh outcomes, and primary interval while applying one common stopping gate. Define

$$g(s) = \mathbf{1}\left\{ \max_{r \in \mathcal{R}} \widehat{\mathrm{VIE}}_{\mathrm{discovery}}(s, r) > 0 \right\}.$$

For each selector $k \in \{\mathrm{VIE}, \mathrm{RIG}\}$, let $r_k(s)$ be its original frozen region. The diagnostic executes $\mathrm{LOOK}(r_k(s))$ if $g(s) = 1$, and the original NOLOOK continuation otherwise. A zero maximum takes STOP. This is a *shared outcome-supervised oracle gate*, including for the RIG selector; it is not an unsupervised RIG stopping policy. STOP means no further visual acquisition, not truncating textual reasoning. No cost penalty is introduced.

Source audit and coverage. All 120 original questions and 360 states remain. The original selection manifest, state/question mapping, and 1,065 unique VIE/RIG/NOLOOK fresh treatments pass restored-inventory hash checks. These contain 4,260 continuations; 15 states share the same VIE/RIG selection. All required outcomes are available. We verify treatment identity, four rollout indices, deterministic fresh seeds, their separation from discovery seeds, and recorded parsed-answer rewards. The 34 missing-final outcomes remain zero reward, with no parser change. No discovery reward substitutes for a fresh outcome, and no new rollout is generated. The gate has 57 LOOK states from 27 questions and 303 STOP states from 107 questions; questions may occur in both subsets.

Table 23: Shared-oracle STOP policies on the complete original cohort. Accuracy brackets are new descriptive 95% question-cluster intervals; original primary inference is unchanged. Both oracle-STOP rows use VIE-based outcome supervision for stopping.

Policy	LOOK (%)	STOP (%)	Fresh accuracy (%)
Forced-LOOK VIE	100.00	0.00	96.25 [94.24, 97.99]
Forced-LOOK RIG	100.00	0.00	89.31 [85.28, 93.06]
Oracle-STOP + VIE	15.83	84.17	96.32 [94.38, 98.06]
Oracle-STOP + RIG	15.83	84.17	91.94 [88.12, 95.35]
Always NOLOOK	0.00	100.00	87.29 [81.80, 92.29]

Paired contrasts. Table 24 shows the shared-gate gap and its reduction. Adding the gate changes VIE accuracy by only +0.07 pp, with an interval crossing zero, while RIG improves by +2.64 pp. The RIG improvement does not eliminate the difference under the same gate. The original +6.94 pp gap and interval $[4.10, 10.28]$ are exactly reproduced using the original bootstrap; the same point estimate under the new common draws has interval $[4.03, 10.21]$. This Monte Carlo difference does not replace the historical primary interval.

Where the original difference remains. For the fresh state-level forced difference $d(s)$, the gated difference is exactly $g(s)d(s)$, because both policies share the NOLOOK reward whenever $g(s) = 0$.

Table 24: Paired post-hoc STOP contrasts in percentage points, with common question-bootstrap draws. Positive stopping changes mean higher accuracy after adding the oracle gate.

Contrast	Estimate [95% CI]
Forced VIE − RIG (diagnostic draws)	+6.94 [4.03, 10.21]
Shared oracle-STOP VIE − RIG	+4.38 [2.01, 7.29]
Forced gap − shared-gate gap	+2.57 [0.97, 4.58]
VIE: oracle-STOP − forced	+0.07 [-0.42, 0.49]
RIG: oracle-STOP − forced	+2.64 [1.11, 4.58]

Consequently,

$$\overline{d} = \overline{gd} + \overline{(1-g)d} = 4.375 \text{ pp} + 2.5694 \text{ pp}.$$

The two terms are weighted contributions to the original full-cohort difference. Table 25 separates them from the conditional differences within each subset; the gate-on +27.63 pp uses only 57 states, not all 360.

Table 25: Fixed discovery-gate subsets (post-hoc). Brackets are descriptive 95% intervals. Questions overlap across rows. The contribution uses the full-cohort denominator; the conditional gap uses eligible states only.

Subset	States / questions	Conditional gap (pp)	Contribution (pp)
Gate-on	57 / 27	+27.63 [15.43, 40.91]	+4.38 [2.01, 7.29]
Gate-off	303 / 107	+3.05 [1.14, 5.42]	+2.57 [0.97, 4.58]

Inference and verification. The post-hoc configuration was fixed before computing these diagnostic results, not before the original study. Each treatment contributes its mean of four paired fresh rewards. We resample all 120 question IDs 10,000 times using NumPy seed 20260912, preserving each question's checkpoints, policy pairing, and shared outcomes. All full-cohort states are equally weighted. For subset statistics, each draw divides the sum of eligible-state values by its eligible-state count; this retains unequal numbers of eligible checkpoints within questions. No draw has a zero subset denominator. Intervals are percentile 2.5/97.5 bounds without multiplicity correction. Gap reduction is computed within each paired draw, not by subtracting interval endpoints. Independent question-count-weighted reconstruction verifies all 32 metric summaries and the integer success totals.

The common LOOK fraction is 15.83% $[10.28, 21.94]$ and STOP fraction 84.17% $[78.06, 89.72]$. These fractions do not account for the expensive construction of an outcome-supervised gate and imply no accuracy–compute advantage. The empirical nonpositive gate does not identify states with truly nonpositive acquisition value. This analysis conditions on realized discovery labels and selections; it does not integrate repeated discovery, establish causal mediation, or compare independently calibrated stopping policies. It supports a remaining region-ranking difference under one shared oracle gate, alongside a measurable gate-off contribution to the forced-LOOK gap.

K Figure construction and illustrative cases (post-hoc)

K.1 Tie strata and observed alternative-policy coverage

Figure 1 and Table 26 reuse the original discovery labels, frozen selections, and fresh outcomes. Exact-max strata are defined before examining their fresh contrasts. They partition states, not questions. Bootstrap draws resample all 120 question clusters and keep all eligible states within each sampled cluster; the stratum estimate is the resulting state-weighted ratio. In the unique-max stratum, 13 of 10,000 draws contain no eligible state and are omitted, not assigned zero. The pooled interval is the unchanged primary interval. These intervals do not include independent repeated discovery.

The unique-max, partial-tie, and all-tied strata contribute 0.625, 6.111, and 0.208 pp to the full-cohort 6.944 pp gap. Thus the all-tied stratum does not account for most of the gap. Nevertheless,

Table 26: Fresh accuracy by discovery tie stratum. Accuracies are percentages; gaps and 95% question-cluster intervals are percentage points.

Stratum	States	Q	VIE	RIG	Gap [95% CI]
Unique max	8	7	46.88	18.75	+28.12 [-5.00, 58.33]
Partial tie	164	89	95.12	81.71	+13.41 [8.03, 19.04]
All 35 tied	188	91	99.34	98.94	+0.40 [-0.15, 1.17]
All (primary)	360	120	96.25	89.31	+6.94 [4.10, 10.28]

partial ties average 23.57 maximizers, so their gap does not identify a uniquely informative crop. The current deterministic VIE selector chooses `g4_r3_c3` in 296/360 states. Figure 3 makes this geometry concentration and missing alternative outcomes explicit.

Coverage is the mean, over all states, of policy probability mass assigned to regions with observed original fresh outcomes. A complete state has outcomes for every action in the policy's support; the fraction of complete states is different from mean coverage. Random-among-ties is uniform over exact VIE maximizers. Canonical max ID always selects the lexicographically largest ID in the entire grid, without consulting VIE labels. The current VIE policy instead chooses the largest ID among exact maximizers. Area- and position-matched ties are uniform over exact maximizers with, respectively, the same normalized area or center quadrant as the original VIE selection. Scale + position is uniform over all regions matching both the selected source scale and center quadrant; these supports also match its area. These geometry controls are conditional on the original selection, not independent deployable policies. The original fresh bank contains outcomes for several frozen policies but does not cover every alternative support. We therefore report coverage rather than full-cohort alternative accuracy and do not impute missing outcomes. This original-bank audit is separate from the new-bank sampled follow-up in Appendix L. The latter does not complete the original bank's exact-mixture outcomes.

K.2 Fixed illustrative-case selection and outcome audit

The illustration frame is all 60 programmatically generated Geometry/Search questions and their checkpoints; source-document and ChartQA images are excluded. The selection rule was saved before opening the selected images. Case 1 requires different selections and higher fresh VIE success. Case 2 requires a high-RIG selected region (within-state top nine), nonpositive discovery VIE, and nonpositive fresh RIG–NOLOOK difference. These pools are ordered by decreasing VIE–RIG fresh gap, unique maximum first, fewer invalid outcomes, question ID, and checkpoint ID. Case 3 requires RIG to outperform VIE and reverses the gap order. Previously selected questions are excluded from subsequent choices. All eligible lists and the rule are retained. These are outcome-selected extremes, not representative samples or independent tests.

All three selected states are Geometry. Cases 1 and 2 have 32 and 28 tied maxima, respectively, and zero discovery VIE for their VIE choices. Case 1 has a blank VIE crop. Neither supports the claim that VIE discovered new useful visual information. Case 2 is high-RIG by within-state rank despite its small absolute KL; a common absolute high-RIG threshold is not assumed. Case 3 favors RIG on fresh seeds even though the selected VIE discovery value is higher. Four-seed success counts remain noisy. Visual inspection also exposes a dataset-label defect: Cases 2 and 3 show base/height pairs 9/11 and 15/7, whose exact areas are 49.5 and 52.5, while the stored keys are 49 and 52. The generator computes `base * height // 2`; these exact areas are absent from the displayed answer options. We retain the original keys and rewards for traceability and mark this defect in the figures. Therefore these two cases illustrate agreement with flawed stored labels, not mathematical correctness. Their keys are not silently repaired, and they cannot establish that either crop supplied correct mathematical evidence. The subsequent full original-cohort audit identifies 13/30 affected Geometry items (Appendix O).

The displayed originals match the archived source-image SHA256 checksums. Regions use the original normalized bounding boxes, rounding, and source pixels. The original image overview is accompanied by a diagram-detail inset and the two frozen crops; display-only white padding preserves image aspect ratios and does not claim to reproduce the inference canvas. Locally re-encoded padded PNGs do not match the archived observation-byte hashes; exact observation-file

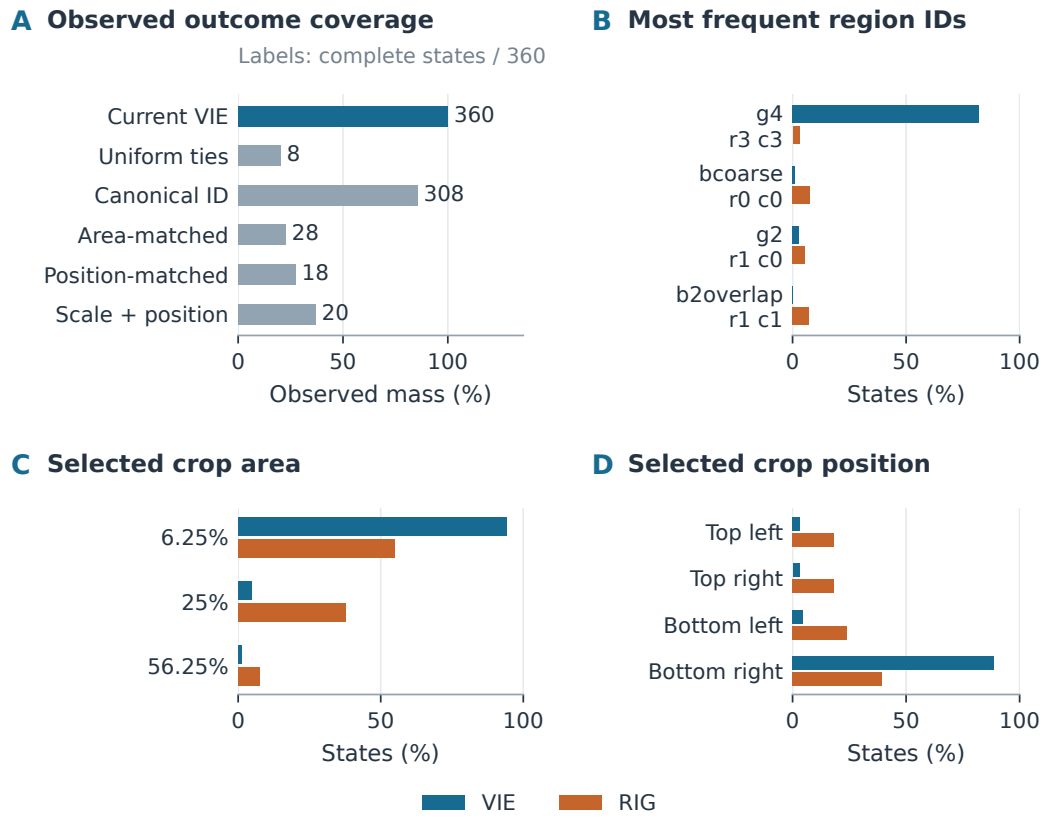

Figure 3: Post-hoc coverage and selection frequency audit. **A:** Bars show observed probability mass; labels count complete states, not successful states. **B:** Four most frequent region IDs by combined VIE/RIG count; the full frequency table is in the numerical artifact. **C–D:** All normalized areas and center quadrants. Different deterministic tie rules and geometry-matched alternatives lack complete fresh coverage; these plots establish neither alternative accuracy nor causal attribution.

identity is therefore not claimed. Original and regenerated hashes, bounding boxes, probability vectors, and all fresh seed IDs are retained in the artifact. No reasoning trace or outcome-based seed filtering is used. The twelve fresh outputs per case are parser-valid and uncapped. Table 27 preserves all rewards in recorded rollout order.

Table 27: Complete fresh reward vectors for the illustrative cases (four recorded seeds per treatment). Missing/invalid outcomes would remain failures; none occurs in these three cases.

Case	VIE	RIG	NOLOOK
1	1 1 1 1	0 0 0 0	1 1 1 1
2	1 1 1 1	0 0 0 0	1 1 1 1
3	0 0 1 0	1 0 1 0	0 0 0 0

K.3 TIE-AWARE PROBE ASSOCIATIONS AND PAIRED FRESH DISTRIBUTIONS

Figure 2 uses all 12,600 candidates. Four LOOK and eight shared NOLOOK rewards put absolute VIE on a 0.125 grid, although within-state differences between regions lie on a 0.25 grid. The count heatmap uses 40 equal-width KL bins and the exact VIE levels, without jitter or an absolute high-RIG boundary. Its extent includes both outer half-bins in full; the paired-gap axis contains every observed question gap. Spearman uses average ranks for ties; Kendall uses tau-b. We average the correlations within the 172 states (89 questions) where both vectors vary. The 188 constant-VIE states

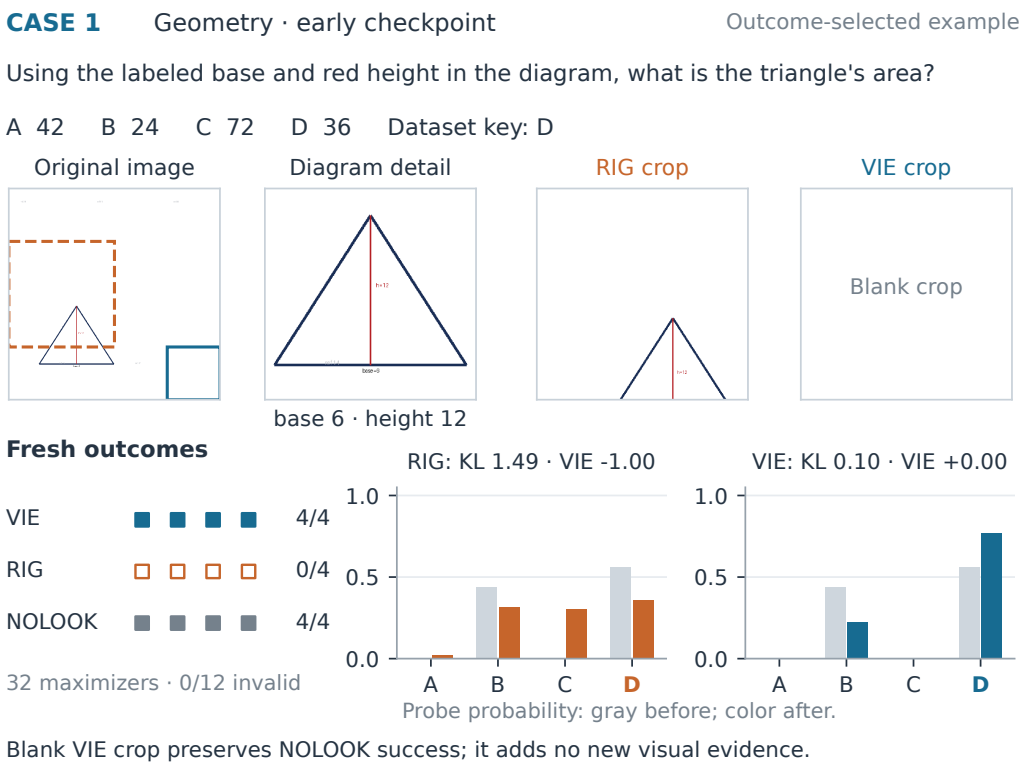

Figure 4: An actual frozen intervention, selected for illustration rather than statistical evidence. Orange dashed and blue solid boxes identify the RIG and VIE selections; the full original, a diagram-detail inset, and the two frozen crops are shown using original source pixels. Filled/empty squares retain each fresh seed as dataset-key match/mismatch, respectively. Bars show forced-choice probabilities before/after the discovery intervention; bold option D is correct. KL and empirical VIE are discovery quantities, while success counts use all four original fresh seeds. No reasoning trace is displayed.

have undefined correlations and are excluded, not set to zero. Mean Spearman is -0.0184 with 95% question-cluster interval $[-0.0920, 0.0558]$; mean tau-b is -0.0138 with interval $[-0.0751, 0.0486]$. Both describe weak average within-state association with substantial uncertainty, rather than proving absence of all relationships.

For these new intervals and the KL bins below, a fixed NumPy bootstrap seed of 20260918 generates 10,000 draws of 120 question clusters. Each estimate is the ratio of sampled sums to eligible-state counts, preserving checkpoint pairing. No draw has a zero denominator. The paired question plot averages all three checkpoints and all four fresh seeds per treatment: 32 questions favor VIE, 82 tie, and six favor RIG. Each plotted line is one question, not an independent checkpoint or continuation.

Figure 7 sorts the 360 RIG-selected states by frozen KL, then checkpoint ID, into four equal rank bins before estimating the paired fresh RIG–NOLOOK contrast. Table 28 gives their ranges and question counts; questions may contribute to several bins. All four intervals cross zero. These exploratory bins do not establish a monotone or causal probe-to-benefit relationship. The KL partition reuses the supervised correct-option/wrong-options decomposition in Appendix G; it requires the answer label and is not causal mediation or a comparison against a newly evaluated probe.

K.4 Contrasts with distinct evidence roles

The original cross-study contrast is retained below to distinguish the offline and learned-policy evidence.

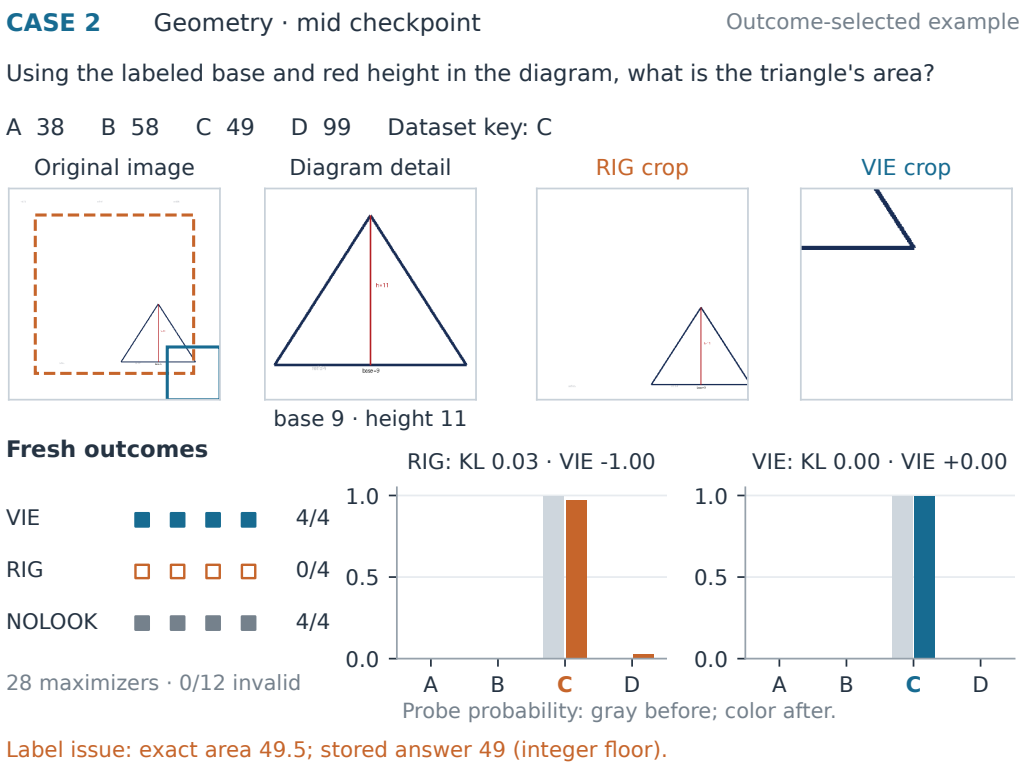

Figure 5: Illustrative Case 2: the RIG-selected region is high by within-state rank and has nonpositive discovery VIE; fresh RIG is 0/4 versus NOLOOK and VIE at 4/4. The absolute KL is small. The stored key is 49 although the exact area is 49.5; counts measure original dataset-key agreement. Formatting and probability/seed separation follow Figure 4; this is an outcome-selected example.

Table 28: Frozen RIG-selected KL rank bins. Every bin has 90 states. Fresh gaps and question-cluster intervals are percentage points.

Bin	KL range (nats)	Questions	RIG − NOLOOK [95% CI]
1	0.00004–0.09588	59	-1.39 [-4.55, 0.99]
2	0.10659–0.57738	63	+2.22 [-2.56, 7.30]
3	0.57741–1.76765	62	+0.28 [-6.25, 8.04]
4	1.81716–12.79067	55	+6.94 [-2.65, 16.89]

L Bounded randomized tie follow-up (post-hoc design)

Design and estimand. After examining the original results, we froze this follow-up before generating any new outcomes. All 120 questions and 360 saved states remain, with unchanged exact empirical maxima. Each state has four independently keyed draws with replacement from each of three supports: all exact VIE maxima, all 35 regions, and all grid regions matching the original VIE selection's source scale and center quadrant. The quadrant boundary is 0.5, with equality assigned right/bottom; matched areas are verified equal. The last control inherits the original selected geometry and is not an independent deployable algorithm. Eight deterministic policies are retained: original VIE/RIG, tie lex-min, three fixed hash tie-breaks, and canonical min/max ID over the entire grid. No rule is selected or added based on the new outcomes.

Every selected action receives four new fresh seeds, shared across actions within a state and disjoint from the original evaluation/discovery namespaces. Duplicate action draws retain their multiplicity in policy scores but share generated outcomes. This gives 4,619 distinct state–action treatments and

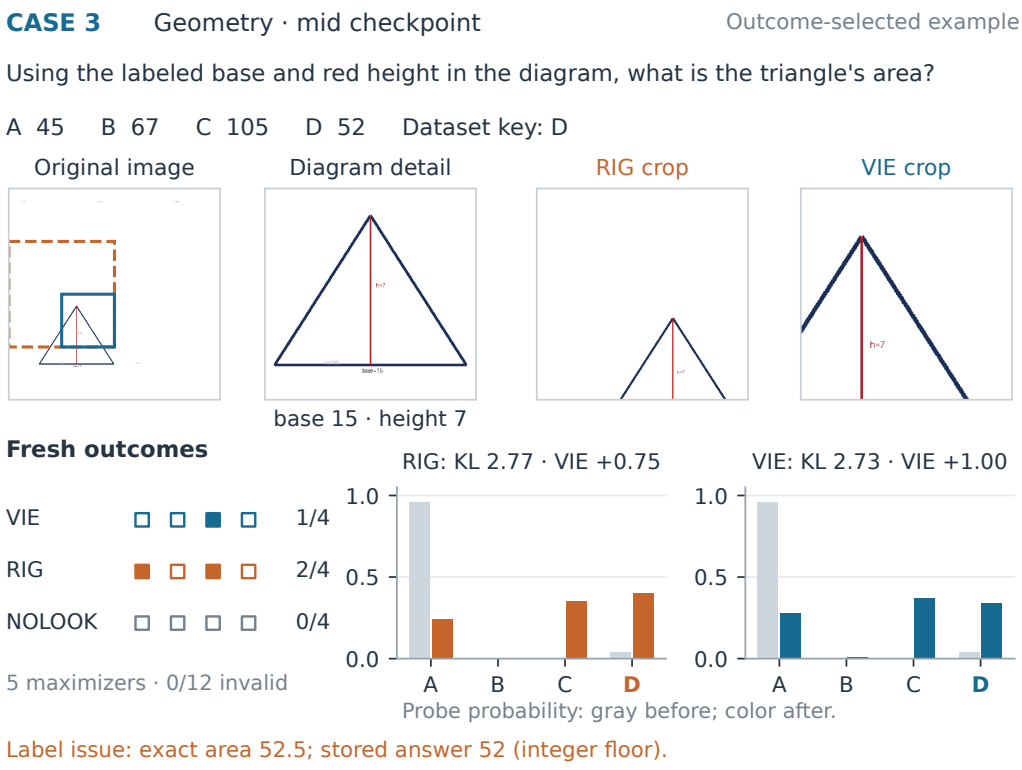

Figure 6: Illustrative counterexample: RIG succeeds on 2/4 fresh seeds, VIE on 1/4, and NOLOOK on 0/4, despite discovery VIE being larger for the VIE choice. The stored key is 52 although the exact area is 52.5; these counts measure original dataset-key agreement. All seeds are retained; this is not a general failure-rate estimate. Formatting follows Figure 4.

18,476 primary continuations. Original VIE and RIG are also reevaluated in this entirely new bank; old outcomes are not pooled with it. The estimand is the average of four frozen sampled policy realizations, conditional on the saved states and original discovery. It is not the exact uniform-mixture accuracy or an independent new-question replication. All arms execute one LOOK; this follow-up has no STOP gate.

Measurement and independent verification. Generation preserves the original prompt, crop, parser, temperature 0.7 and top-p 0.9. Only a truncated 2,048-token response receives one same-seed replay at 4,096 tokens. Nine such replays yield 18,485 total calls, below the frozen 36,952-call ceiling. Seven invalid-final and 93 missing-final responses retain zero reward; no slot is missing. An independent implementation reparses all responses and repair records, verifies 55,446 slot files and reproduces 676 numerical checks, including direct-index question bootstraps, without importing the execution analyzer. The calls generated 1,390,975 tokens with summed recorded latency 55.73 hours; this is neither GPU-hours nor an end-to-end cost estimate. Historical model config/index/processor hashes match; full weight-shard hashes were first recorded for this follow-up, so historical bitwise weight equality cannot be established retrospectively.

Paired comparisons. State scores average the four continuation rewards per action and then the four draw slots. Question-cluster bootstrap resamples all 120 questions 10,000 times with NumPy seed 20260915. The three fixed primary contrasts use Bonferroni-adjusted 98.333% individual intervals (a 95% family); ordinary 95% intervals are also reported. Subgroups use the eligible-state ratio and common draws; 13 unique-max draws have zero denominator and are omitted. All secondary contrasts are descriptive.

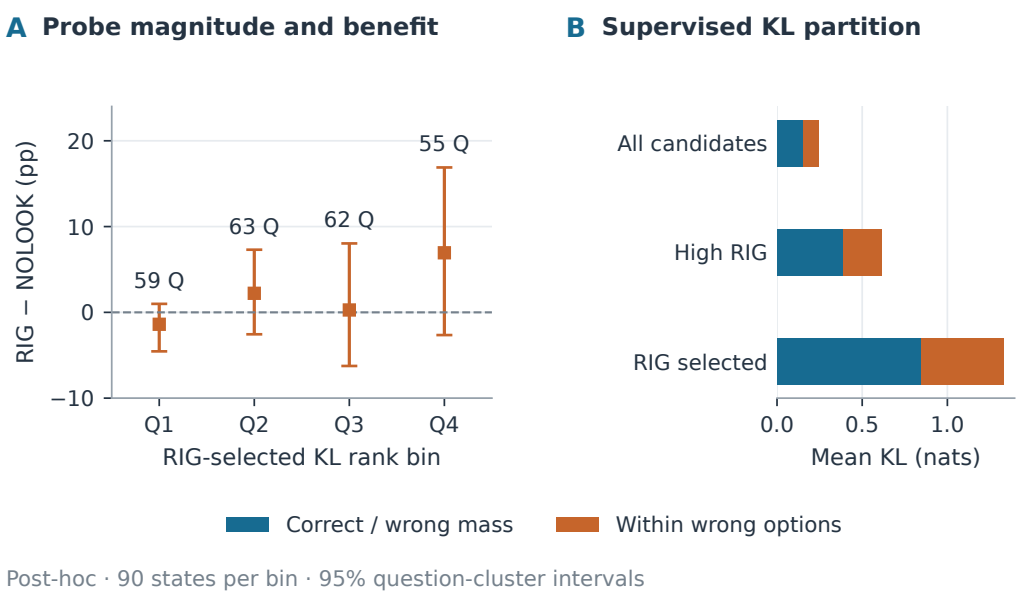

Figure 7: Post-hoc probe diagnostics. **A:** Paired fresh RIG–NOLOOK benefits in four fixed KL rank bins, with 95% question-cluster intervals and contributing question counts. **B:** Mean KL split into correct/wrong mass change and redistribution among wrong options, weighted by after-probe wrong mass. High RIG denotes the top nine candidates per state; the decomposition is label-supervised.

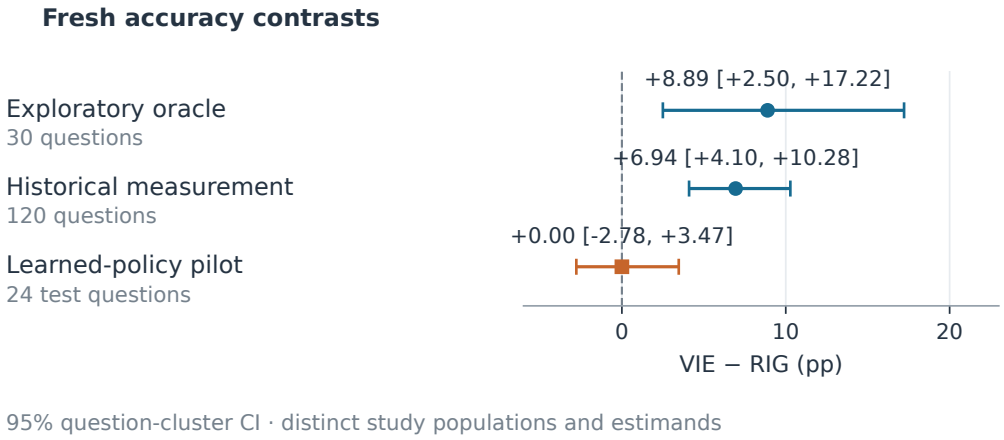

Figure 8: Fresh VIE–RIG accuracy contrasts, with 95% question-cluster bootstrap intervals. The earlier exploratory and historical measurement rows compare empirical region selectors. The pilot compares VIE-balanced with RIG-VIE-calibrated on a subset of the historical 120-question corpus. The rows have different estimands and are not pooled or independent replications of a single learned policy.

All three adjusted intervals are positive. This supports the tested sampled-max procedure over these controls on the stored-label cohort; it does not rule out all spatial confounding or establish reliable high-value set discovery across new discovery banks.

Original lex-max VIE minus sampled uniform-max is -0.42 pp, 95% CI $[-1.32, 0.36]$: no advantage of this particular original tie-break is established, and the interval is not an equivalence test. The four draw-index realizations have sampled-max accuracies 96.53, 96.94, 96.46, and 97.57%; their gaps versus RIG range from +6.39 to +7.50 pp and versus geometry from +5.97 to +6.46 pp. They share a continuation bank and are not four independent experiments. The question bootstrap

Table 29: New-bank sampled uniform-max minus each fixed primary comparator (pp). Intervals condition on discovery and sampled actions.

Comparator	Gap [95% CI]	Gap [family-adjusted CI]
RIG	+6.81 [3.99, 10.00]	+6.81 [3.45, 10.71]
Sampled uniform-all	+10.63 [6.58, 15.24]	+10.63 [5.76, 16.44]
Sampled geometry-matched	+6.22 [3.54, 9.27]	+6.22 [3.04, 10.05]

Table 30: All eleven frozen policies on the new bank. Secondary differences versus RIG retain negative and null results; no best tie-break is selected.

Policy	Accuracy (%)	Difference vs RIG [95% CI]
Original VIE (lex-max)	96.46	+6.39 [3.68, 9.51]
RIG	90.07	Reference
Tie lex-min	97.08	+7.01 [4.03, 10.28]
Tie hash 0	97.08	+7.01 [4.31, 10.00]
Tie hash 1	96.46	+6.39 [3.61, 9.44]
Tie hash 2	96.74	+6.67 [3.75, 9.93]
Canonical max ID	86.39	-3.68 [-7.85, 0.21]
Canonical min ID	87.78	-2.29 [-6.18, 1.60]
Sampled uniform-max	96.88	+6.81 [3.99, 10.00]
Sampled uniform-all	86.25	Not a fixed contrast
Sampled geometry-matched	90.66	Not a fixed contrast

conditions on these draws, not the complete action-randomization mechanism. For a fixed potential-reward bank, action randomization alone has worst-case pooled SD 1.318 pp for one sampled arm and 1.864 pp for two independent sampled arms' difference; these are variance bounds, not empirical standard errors or confidence intervals.

Table 31: Descriptive sampled-max gaps within original tie strata (pp, ordinary 95% intervals). Questions may occur in multiple strata.

Stratum	vs RIG	vs uniform-all	vs geometry
Unique (8 states)	+28.13 [0.00, 56.40]	+39.06 [12.50, 67.50]	+10.16 [0.00, 21.09]
Partial (164)	+13.41 [8.07, 18.98]	+21.46 [13.92, 29.49]	+13.22 [7.78, 19.07]
All tied (188)	+0.13 [-0.15, 0.53]	-0.03 [-0.15, 0.07]	-0.07 [-0.20, 0.06]

The partial-tie gap versus geometry is +13.22 pp; the all-tied gap is -0.07 pp with an interval crossing zero. In all-tied states the two uniform supports coincide, although independent finite action draws need not produce identical scores. An additional descriptive domain check, added at numerical review rather than preregistered, gives sampled-max minus RIG gaps of +8.61/+4.65/+14.72/-0.76 pp in Chart/Doc/Geometry/Search. The Search interval is $[-1.39, -0.25]$ pp (uncorrected), so the pooled result does not imply every-domain improvement. The geometry-matched gaps in those domains are +6.46/+3.96/+13.47/+0.97 pp. Geometry retains the known integer-floor label defect discussed in Appendix K; these outcomes are stored-key agreement, not a repaired mathematical benchmark. Neither this diagnostic nor excluding a domain silently replaces the original primary result.

M PROBE DIAGNOSTICS, CONTINUATION FIDELITY, AND INTERFACE SENSITIVITY

This post-hoc analysis fixes selectors and scoring before their new summaries, retaining all 120 questions, 360 states and 12,600 LOOK candidates. We normalize the saved four-choice vectors, compute JS, TV and $H(p_{\text{before}}) - H(p_{\text{after}})$, and maximize each score with exact equality and fixed lex-max/min plus three hash tie rules. All twenty rules, including KL, are retained in the numerical artifact. Normalization changes KL by at most 1.249×10^{-6} and preserves all 360 original lex-max

RIG selections. Independent JS-entropy and TV-positive-mass identities pass for all candidates; a second direct-index implementation reproduces 686 metric checks and all 7,200 choices. JS divergence and total variation measure differences between the saved probe distributions; neither is an expected-information-gain acquisition rule.

Table 32: Complete fresh support for lex-max selectors, reported separately for the original and new tie banks. Counts are states out of 360.

Measure	Original	New tie	Full-cohort point/CI
KL	360	360	Identified in both
JS	337	334	Unidentified in both
TV	332	329	Unidentified in both
Entropy decrease	360	152	Original bank only

Incomplete selectors have null full-cohort accuracy/gap/CI, rather than being renormalized on available actions. Finite-bank accuracy bounds for JS are [83.75,90.14]%/[83.26,90.49]% in original/new banks, and for TV [82.71,90.49]%/[82.57,91.18]%; these are not confidence intervals. Complete-subset summaries retain same-subset original VIE/RIG comparisons in the artifact and cannot replace the full cohort. Entropy decrease has original-bank accuracy 88.54% and a -0.76 pp difference versus RIG, 95% CI $[-5.00, 3.19]$; its new-bank value is unidentified. Banks are never pooled.

Correctness calibration and continuation fidelity. Four-class Brier against stored answer keys, summed over classes, is 0.3862 [0.2941,0.4853] across all candidates. Probe top-label accuracy is 76.29%, while mean maximum probability is 90.00%. Stored-key calibration inherits the known Geometry label limitation. Continuation fidelity is a separate target: we append probability zero for a fifth invalid/missing-final category and average the five-class Brier over all four continuations per available treatment. Argmax agreement counts invalid/missing final as disagreements. Treatment-weighted ratios use question-cluster bootstrap with 10,000 shared draws and seed 20260920; availability is determined by each bank's selected-action union, not random missingness.

Table 33: Probe versus continuation outcomes; 95% question-cluster intervals. Every continuation, including invalid/missing final, contributes to the denominator.

Bank	Treatments	Agreement (%)	Five-class Brier
Original LOOK	1936	80.59 [76.29, 84.56]	0.3094 [0.2469, 0.3775]
New tie LOOK	4619	80.50 [76.25, 84.50]	0.3054 [0.2446, 0.3700]
Original NOLOOK	360	84.38 [79.86, 88.75]	0.2520 [0.1843, 0.3237]

Original LOOK has 59 missing-final responses, new LOOK has 93 missing-final and seven invalid-final responses, and original NOLOOK has 12 missing-final responses. Ten fixed equal-width confidence bins include empty-bin nulls. In the new LOOK [0.9,1.0] bin (3,204 treatments), mean confidence is 98.56%, stored-label top accuracy 89.20%, and continuation agreement 92.26%. These three targets must not be conflated. Agreement is not equivalence, and four continuations per treatment leave a sparse empirical distribution.

Completed bounded prompt/readout sensitivity. The saved bank above contains one prompt/readout protocol; JS/TV recomputation cannot test another interface. We therefore executed the separately pre-specified Qwen development study on 12 existing questions (three per domain), 36 states and all 35 crops plus NOLOOK. Questions were selected by a fixed ID hash within domain, without filtering on outcomes. We reran the original probe in the same runtime, changed only its prompt wording in one condition, and removed the leading space from answer-label candidates in another. Each condition retains four forced-choice candidates. Full-string conditional log probabilities use summed full-vocabulary log-softmax, without length normalization; separate suffix encoding preserves the frozen prompt boundary. The readout change also changes candidate token identity.

All 3,888 probability records were obtained with 15,552 candidate forwards, zero generation and no retries. A prompt paraphrase preserves the selected region in 10/36 states (27.8%; question-bootstrap 95% CI [13.9,44.4]), with mean region-score Spearman correlation 0.426. Removing the

leading space preserves 26/36 selections (72.2%; [58.3,86.1]), with mean correlation 0.846. These are descriptive comparisons on existing development data, using one fixed variant of each kind; they do not establish a general difference between classes of perturbation.

No fresh continuation outcomes were collected for variant-selected actions in this 12-question development study. The separate 48-question exploratory follow-up supplies such outcomes on its own cohort (Section 5.4); the studies are not pooled. Selection changes therefore do not establish changes in accuracy or the VIE–RIG gap, and cannot be used as evidence that VIE is generally superior. Open-ended readouts, option-order changes and broader template/model robustness were not tested. The stored-key validity concerns apply independently (Appendix O).

N SAME-TASK MISTRAL SECOND-BACKBONE REPLICATION

Design and scope. We fixed Mistral-Small-3.2-24B-Instruct-2506 (revision `95a6d26c4bfb`; full hash in the artifact) before its measurements. An engineering development cohort contains ten previously admitted pilot questions, disjoint from the 120 replication questions by question ID, image bytes and non-null source-record hashes. The two natural-image sources also have distinct source-image IDs. Synthetic source-record hashes are absent; image-byte isolation does not establish distinct template families or semantic independence. The replication reuses the original tasks and stored answer keys, with 30 questions per domain. Mistral independently generates its three checkpoints, discovery outcomes, and fresh continuations. This tests the procedural contrast on a second backbone using the same tasks, rather than providing new-data validation or standard benchmark accuracy.

Compatibility and execution. The low-resolution canvas remains 224×224, with B35 regions and 512×512 crop canvases. The native Mistral tokenizer maps the latter to 532×532 image tensors. Thus, equal crop canvases do not mean equal visual tokens or FLOPs across backbones. Both alternatives retain the original generated prefix before the intervention turn. Development exposed a token-boundary issue: Mistral can merge the closing angle bracket of a step with its following newline. The correction retains the first complete original-token boundary that includes the step ending. It permits only trailing whitespace and rejects any subsequent reasoning step or final-answer content. The first saved generation is reused, not resampled; the earlier code and failure are archived before refreezing. All ten development questions then complete: 70 generations, no repairs and 60 four-choice probes (240 forward calls), including one invalid final. Correctness is not a development selection gate.

The continuation and probe prompts match the primary study's text. Four space-prefixed labels have native token IDs 1349, 1398, 1359 and 1407. We compute full-string teacher-forced log likelihood, sum over tokens without length normalization, and normalize over four labels. The implementation supports multiple tokens, although these four labels are single tokens. Generation fixes temperature 0.7, top-p 0.9 and 2048 new tokens. Only a parser-classified truncation permits one same-seed replay at 4096; invalid/missing finals are never resampled. Remaining nonvalid answers receive zero stored-key reward.

Frozen estimand and completeness. Each of 360 states has eight shared NOLOOK discovery continuations and four for each of 35 LOOK actions. VIE is mean LOOK reward minus the shared NOLOOK mean; RIG is posterior-to-prior KL on four choices. Exact score ties use lexicographic maximum region ID. After all discovery, selections are frozen before opening four new fresh seeds per distinct VIE/RIG/NOLOOK/gray-placebo treatment. These four core policies exclude the original auxiliary selectors. The paired primary contrast is fresh VIE minus RIG accuracy, averaging four seeds per state and three states per question. The 10,000 question-cluster draws use seed 20260914; the percentile interval conditions on realized discovery selections.

There are 120 preparation generations, 53,280 discovery continuations and 5,668 distinct fresh continuations. Nine discovery truncations each receive one repair and remain truncated; total replication generation calls are 59,077. The 12,960 probe distributions require 51,840 forwards. Discovery retains 596 invalid finals, three missing finals and nine unresolved truncations; fresh retains 56 invalid finals. All expected records are present, with no orphan claims or additional generation slots. Invalid rewards remain in every applicable policy denominator. The two-GPU scheduling change reuses all

completed serial slots, partitions the original question order into two fixed 60-question groups, and preserves a global discovery-to-fresh barrier. The parallel phase lasts 19.65 hours; summed recorded replication generation/probe latency is 34.81 hours, neither a complete cost accounting nor a GPU-hour measurement.

Table 34: Mistral fresh evaluation on the same 120 tasks. Accuracies are stored-key agreement; paired differences and 95% question-cluster intervals are in percentage points.

Policy	Accuracy (%)	VIE minus policy [95% CI]
Empirical VIE	83.68	Reference
RIG-KL	73.13	+10.56 [6.39, 14.86]
NOLOOK	67.71	+15.97 [10.97, 21.25]
Gray placebo	67.64	+16.04 [10.83, 21.46]

The primary result supports the tested empirical selection procedure over this forced-choice KL probe on a second system. It does not isolate the contribution of the VIE signal from the tie-break or establish superiority of a deployable learned selector. Absolute accuracies and gap differences across models are not a controlled model ranking: reasoning states, native visual processing and model internals differ.

Table 35: Post-hoc Mistral domain summaries (30 questions each). These unadjusted descriptive intervals were added during audit, not selected as confirmatory subgroup hypotheses.

Domain	VIE (%)	RIG (%)	Gap [95% CI] pp
Chart	76.67	64.72	+11.94 [1.94, 22.22]
Doc	83.06	71.67	+11.39 [4.72, 18.89]
Geometry	78.89	61.11	+17.78 [8.06, 27.22]
Search	96.11	95.00	+1.11 [-2.78, 5.28]

Search has a small positive point estimate with an interval crossing zero, so the pooled result does not imply every-domain benefit. Excluding Geometry leaves 90 questions and a +8.15 pp gap [3.80,12.78]; this is an additional descriptive sensitivity analysis, not a replacement primary. Geometry still inherits the integer-floor stored-key defect. Neither the exclusion nor the second backbone repairs distractor construction or mathematical label validity. Appendix O reports the completed label audit; the Qwen probe study does not establish downstream robustness on either backbone.

Tie structure and attribution. There are 29 unique-max, 268 partial-tie and 63 all-35-tied states. VIE chooses the canonical bottom-right quarter `g4_r3_c3` in 187/360 states. These counts describe the frozen procedure; no randomized-max or geometry-matched fresh policy was run for Mistral. The Qwen tie follow-up therefore cannot be transferred as proof of Mistral signal-specific attribution or repeated-discovery reliability.

Independent audit and reproducibility. An independently written parser/seed/selector/statistics implementation checks the complete 203,676-file private snapshot, verifies all 12,600 candidate VIE/KL values and all 360 choices, and reproduces the primary through multinomial-weight question bootstrap rather than the execution analyzer's direct indexing. A separate native-tokenizer pass reconstructs input contexts and checks original token prefixes, decoded outputs and truncation flags without importing the execution adapter. Model, source-image, code and design hashes are checked, and filesystem ordering agrees with discovery-before-freeze-before-fresh. These checks validate recorded evidence and implementation consistency, not an independent stochastic model rerun or proof of historical immutability. The numerical supplement contains discovery and fresh reward banks with a separate recomputation entry point, but excludes model weights, source images, private paths, and raw responses. The primary Qwen result, STOP diagnostic, and learned-policy null results remain unchanged.

O TASK-CONSTRUCTION AUDIT AND FIXED-INPUT LABEL SENSITIVITY

Audit scope and construction shortcuts. The subsequent audit covers the original 120 questions and ten Mistral engineering-development questions. All 130 image byte hashes were checked; 66 synthetic images were regenerated byte-for-byte, and source-record hashes and question/answer consistency were checked for 64 real-source records. These are provenance and construction checks, not independent semantic adjudication. Source-conditioned deterministic transformations of numeric/text options and synthetic option patterns recover every stored key in the original 120 questions without images. This establishes an available construction shortcut, not a neural text-only baseline or evidence that either backbone exploited it. New synthetic indices reuse template families; byte-distinct images do not establish template independence.

Geometry defect. Thirteen of the original 30 Geometry questions use integer division for triangle area and have no exact correct option. Their synthetic indices are 0005, 0006, 0009, 0010, 0012, 0014, 0015, 0016, 0017, 0018, 0020, 0024 and 0026. The original inputs, keys, discovery measurements and primary scores remain unchanged for traceability. This is a systematic task-construction defect, not merely uncertainty about an otherwise valid answer label.

Fixed-input, fixed-selection diagnostic. A separately frozen diagnostic retains all 120 questions and 360 states per backbone. It scores options by exact numeric value on the 13 affected questions: all available answers are wrong, so both selectors receive zero there; other stored-key rewards remain unchanged. A shared question bootstrap uses 10,000 draws with seed 20260915. The resulting VIE–RIG differences are +4.65 pp [2.43,7.22] for Qwen and +8.61 pp [4.93,12.43] for Mistral. Paired changes from original scoring are -2.29 pp $[-4.72, -0.49]$ and -1.94 pp $[-4.17, 0.07]$, respectively. These diagnostic intervals do not replace historical primary intervals.

These scores are not accuracies on repaired questions. Replacing an option would change the model input and could alter checkpoints, discovery rankings, and continuation outcomes. The existing outcome banks therefore identify neither fixed-action nor newly selected results on repaired inputs. Under the original inputs, exact-value discovery rewards on the affected states are zero for every action, producing all-action ties. This diagnostic neither validates their original rankings nor removes construction shortcuts in the remaining questions. No question was removed or selected for its effect on the gap.

Semantic review and pilot extension. A single-agent visual review of all 64 real-source records supported displayed values or semantics for 51, retained unit/scale qualifications for eight, and left five reference, precision, index-interpretation or handwriting issues unresolved. This was not independent human double-blind adjudication, and source-record consistency is not proof of correctness. A separate pilot synthetic extension regenerated 60 images and checked 180 checkpoint targets without opening the sealed test-containing source manifest. Seven of its 30 Geometry constructions have the same truncation defect (five training and two validation). Reconstructed options are distinguished from directly inspected stored rows; these findings are not pooled with the original 120-question analysis. They further limit claims about the learned-policy study.

Interpretation. The historical results remain reproducible contrasts between frozen procedures under stored-key rewards. Neither the second backbone nor fixed-input rescoring repairs invalid options or establishes corrected mathematical reasoning, robust distractor handling, visual necessity, or standard-benchmark generalization. Those claims require a separately audited cohort and fresh evaluation.

P COMPLETED EXPLORATORY MULTI-INTERFACE PROTOCOL AND AUDIT

This appendix documents the 48-question exploratory follow-up in Section 5.4. It is separate from the 12-question development study in Appendix M, which collected no fresh outcomes for variant-selected actions.

Data and admission provenance. The parent preparation froze 192 questions, 48 per domain, with revised real-source options and newly generated synthetic items. Recorded agent review is not

independent human adjudication. The no-image diagnostic completed 1,536 candidate forwards and retained the aggregate gate failure: document options-only 20/48, document question-plus-options 19/48, chart options-only 13/48, and chart question-plus-options 16/48. The two synthetic domains each scored 12/48 in both modes. The rule required each real-domain/mode proportion to be at most 0.40; the document options-only result failed it. The diagnostic was not rerun on the reduced cohort, and no passing subset gate is claimed.

After the parent cohort failed the admission rule, a result-independent hash rule reduced the exploratory follow-up to 12 questions per domain. Complete four-item groups were retained in each synthetic domain, yielding three resampling groups per domain. This post-launch amendment changed the originally specified sample size. The parent-cohort gate failure remains in force; the reduced cohort was not re-admitted for clean-task confirmation.

Interfaces, selection and evaluation. The baseline P_0 prompt is: "Based on the evidence available so far, if you had to answer now, choose one option: A / B / C / D Answer:" The P_1 prompt is: "Using only the evidence currently available, which answer option is best supported? Select A, B, C, or D. Answer:" The archived design preserves the exact line breaks. R_0 uses candidates ␣A, ␣B, ␣C, ␣D (one leading space); R_1 removes that space. KL uses natural logarithms and a probability floor of 10^{-12}. Each of the three interfaces scores 35 crops against its own NOLOOK probe. VIE uses four LOOK rewards minus the shared mean of eight NOLOOK rewards. Exact ties select the lexicographically largest region ID for every selector. No interface is selected retrospectively by its fresh performance.

The 144 saved states are frozen before discovery. Prefix recovery amendments preserve actual generated token prefixes and are archived separately from the original boundary rule; they do not synthesize reasoning. Discovery requires $144(35 \times 4 + 8) = 21{,}312$ primary continuations and $144 \times 36 \times 3 \times 4 = 62{,}208$ probe candidate forwards. The frozen action union contains 415 distinct state–region pairs, each evaluated on four fresh seeds. Coincident selectors share the same treatment outcomes. Temperature 0.7, top-p 0.9, an initial 2,048-token cap and one same-seed 4,096-token truncation repair are fixed; terminal failures score zero and are not deleted. Integrity failures block analysis rather than induce complete-case evaluation.

Secondary interface and cross-cohort interpretation. The paraphrased-prompt and altered-readout RIG selectors differ from baseline RIG by +2.78 pp $[-1.56, 7.29]$ and +0.69 pp $[-1.91, 3.47]$, respectively. These paired 95% intervals establish neither improvement nor equivalence. The historical Geometry gap is +15.00 pp, whereas the follow-up Geometry VIE–baseline point gap is zero; the aggregate separation is therefore not consistently driven by Geometry across cohorts. Because the cohorts differ in questions, saved states, and sampling, the larger follow-up aggregate gap cannot be interpreted as an improvement over the historical study.

Clustered inference. Let y_{qs} be the mean reward for selector s over the three checkpoints and four fresh continuations of question q. For interface c, the reported contrast is

$$\widehat{\Delta}_c = \frac{1}{4}\sum_{d=1}^{4}\frac{1}{12}\sum_{q\in d}(y_{q,\mathrm{VIE}} - y_{qc}).$$

For each of 10,000 bootstrap draws, clusters are sampled with replacement within domain (12, 12, 3 and 3 clusters); all questions of a sampled synthetic group remain together. Equal cluster sizes within each domain make this equivalent to averaging group means before equal-domain aggregation. The NumPy default random generator uses seed 20260916, in archived domain order. The same draws generate all three paired contrasts and the secondary alternate-minus-baseline differences. Marginal bounds are the 2.5/97.5 percentiles; the three-member Bonferroni bounds are the $0.05/6$ and $1 - 0.05/6$ quantiles. These conditional bootstrap intervals neither integrate repeated discovery nor establish clean-confirmatory coverage after a failed data gate.

Integrity and cost accounting. Archived integrity checks confirm that discovery and fresh records passed validation and that actions were frozen before fresh evaluation. Selected-action and score-file hashes match the files used for the reported analysis. The fresh bank contains 1,660 final records and 1,688 generation attempts, including repairs. The local retained ledger contains 23,417 generation

Table 36: Descriptive follow-up domain results. Each domain has 12 questions and 36 states; G is its number of resampling clusters. Values are fresh stored-key agreement (%), not standard source-benchmark accuracy. No domain-wise superiority test is claimed.

Domain	G	VIE	P_0R_0	P_1R_0	P_0R_1
Document	12	90.97	61.81	63.89	61.81
Chart	12	79.86	35.42	47.22	35.42
Geometry	3	95.14	95.14	95.14	94.44
Search	3	72.22	40.97	38.19	44.44
Equal-domain mean	30	84.55	58.33	61.11	59.03

attempts, or 23,418 charged calls including one historical abandoned attempt; it records 62,208 probe forwards and 4,275,454 generated tokens. The wider campaign ledger retains superseded work: 24,858 charged generation calls and 65,972 probe forwards. The original 1,536 no-image forwards are separately recorded as reused diagnostic cost. These scopes must not be added as disjoint runs. Attempt-level parser counts (22,090 valid, 810 missing-final, 510 truncated, seven invalid-final) include superseded attempts and therefore are not final-outcome failure rates.

Reconstruction boundary. From the 48 archived question summaries, the writing audit independently recomputed the three point contrasts, their marginal and Bonferroni intervals, both alternate-minus-baseline comparisons, and every domain accuracy. It also checked selected-action, score-file, and discovery-audit hashes against the frozen action record. This verifies numerical consistency of the reported summaries, not full raw-output reconstruction, a new GPU run, or independent semantic adjudication. The follow-up provides fresh evidence for the specified interfaces but remains exploratory after the failed admission gate. Its small synthetic-cluster count, outcome-supervised VIE selection, and absence of a fresh NOLOOK control further bound interpretation.