COMMITMENT IS NOT GROUNDING: VISUAL-SEMANTIC REREADING FOR DIFFUSION MLLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion multimodal large language models iteratively resolve answer tokens, yet committing to an answer does not establish visual support for the queried target. We introduce Internal Visual-Semantic Rereading (IVSR), an inference-time controller that audits the native answer at its first valid transfer. A maturity gate selects low-margin commitments for a short visual reread. The reread answer state retrieves an internal visual region, while a contextual question-target representation evaluates its support. A calibrated policy combines this alignment with reread answer margins to keep, reject, or recover the native answer. Backbone weights and prompts remain unchanged; thresholds use labeled held-out calibration. Across DiffusionVL, MMaDA, and LaViDa, IVSR improves all 18 POPE settings by 2.34 F1 points on average and reduces AMBER-D existence hallucination rates by 16.1–47.7%. Separate inference profiling measures 1.4–9.8% gated latency overhead. Mechanism diagnostics and ablations support separating inspection priority, answer preference, and target-specific visual evidence in finite-answer verification.

1 INTRODUCTION

Resolving an answer token is a decoding event, not in itself a verification that the image supports the answer. Diffusion multimodal large language models (MLLMs) generate responses by iteratively parsing masked tokens (Sahoo et al., 2024; Nie et al., 2025; Yang et al., 2025; Li et al., 2025; Zeng et al., 2025). While this iterative process smoothly forms a decoding decision, high internal confidence in a resolved token does not inherently guarantee rigorous grounding in the visual input. Therefore, tackling visual hallucinations requires looking beyond when an answer is formed, to evaluating whether that specific answer possesses sufficient, target-relevant visual evidence.

Existing inference-time interventions, such as contrastive decoding, attention adjustment, and intermediate-layer correction, typically modify predictions by enhancing general visual signals or altering the global decoding trajectory (Leng et al., 2024; Favero et al., 2024; Liu et al., 2024). However, enhancing overall visual influence is distinct from deciding whether to revise an already selected answer. Inspecting every generated token incurs unnecessary computational costs, and blindly penalizing uncertain answers risks introducing new errors. Our focus is the decision rule that connects internal evidence to the retention or revision of a specific native answer.

The masked diffusion decoding interface provides a precise boundary for this decision. Specifically, the first valid answer transfer at the designated answer position exposes the concrete candidate the native decoder is about to submit. By intercepting the generation at this exact event, the candidate and its associated internal states can be captured. Consequently, the native answer remains the reference for every subsequent inspection and revision. This structural advantage leads to our core question: *How can a diffusion MLLM selectively re-examine the answer it is about to submit, and use target-specific visual support to decide whether to retain or revise it?*

Our internal diagnostics address this question by revealing why different signals must be disentangled (Section 3). First, native errors heavily concentrate among low-margin commitments, suggesting that initial answer separation is an effective proxy for allocating inspection priority. Second, a paired visual-access probe illustrates that a weak, incorrect preference can persist through equal-step repetitions, yet reverse when visual access is actively modified. Finally, correct and incorrect object queries often attend to similar overlapping spatial regions, demonstrating that visual focus alone

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.

cannot confirm target identity. These observations motivate separating inspection priority, reread answer preference, and target-specific visual support, rather than treating any one of them as a sufficient basis for revision.

We introduce Internal Visual-Semantic Rereading (IVSR), a selective inference-time controller that re-examines native answer commitments using reread preferences and target-specific visual support. A maturity gate first isolates low-margin commitments. For these selected cases, a short visual reread provides updated answer logits and internal visual states. Crucially, the reread answer state determines *which* visual content to retrieve, while the contextual question target determines *what* that content should support. A calibrated policy then combines the reread answer margin and the region-target alignment to decisively keep, reject, or recover the native answer. Model weights and prompts remain entirely unchanged, with decision thresholds calibrated on labeled held-out data.

Our contributions are threefold:

1. **Commitment-level visual verification.** We formulate visual verification in diffusion MLLMs around the native answer selected at the first valid transfer. Internal diagnostics motivate a distinction between prioritizing inspection, re-evaluating answer preference, and assessing target-specific visual support, providing the basis for selective answer revision.
2. **Selective visual-semantic rereading.** We introduce IVSR, which preserves the native answer as its reference and separates answer-conditioned visual retrieval from target-conditioned evaluation. A calibrated policy combines regional alignment with reread answer margins to keep, reject, or recover the answer, without changing model weights or prompts.
3. **Cross-backbone effectiveness and efficiency.** Across three diffusion MLLMs, IVSR improves all 18 evaluated POPE settings over native decoding by an average of 2.34 F1 points and reduces AMBER-D existence hallucination rates by 16.1–47.7% relative. Separate inference profiling measures 1.4–9.8% additional latency, while ablations and matched-calibration comparisons further evaluate the controller.

2 Related Work

Hallucination and visual grounding. POPE and AMBER measure visual hallucination, while MME evaluates broader perception and cognition (Li et al., 2023; Wang et al., 2023; Fu et al., 2025). LURE and Woodpecker revise generated content (Zhou et al., 2024; Yin et al., 2024); VCD and M3ID contrast visual evidence with degraded inputs or language priors (Leng et al., 2024; Favero et al., 2024); PAI adjusts visual attention (Liu et al., 2024); and DeCo incorporates intermediate-layer predictions (Wang et al., 2025a). Internal-grounding studies examine semantic interventions, absent concepts, and visual counterfacts (Wang et al., 2025b; Kim et al., 2025; Golovanevsky et al., 2025). Closely related work decodes attended visual features with Logit Lens to detect semantic mismatch (Wang et al., 2026). IVSR instead evaluates pooled visual states against the contextual question target at a diffusion answer's first transfer, retaining the native candidate as the reference for revision.

Diffusion decoding. Masked diffusion language models iteratively resolve tokens (Sahoo et al., 2024; Nie et al., 2025); MMaDA, LaViDa, and DiffusionVL extend diffusion generation to multimodal tasks (Yang et al., 2025; Li et al., 2025; Zeng et al., 2025). Attend-and-Excite guides attention during text-to-image synthesis to improve prompt-subject coverage (Chefer et al., 2023). Our setting concerns verifying an answer about an input image: the first valid transfer supplies a candidate and checkpoint for selective visual reinspection.

Selective decisions and calibration. Selective classification trades coverage for prediction risk (Geifman & El-Yaniv, 2017), and post-hoc confidence estimation ranks predictions for selection (Cattelan & Silva, 2024). Probability calibration separately aligns confidence with correctness likelihood (Guo et al., 2017). IVSR uses margins to allocate inspection and held-out labels to fit revision thresholds. It returns an answer for every query, combining reread preference with target-specific visual support to decide whether the native answer should change.

3 DIAGNOSING NATIVE ANSWER COMMITMENTS

We examine three questions that precede audit design: which commitments merit inspection, whether native iteration revises an emerging error, and whether visual focus identifies the queried target. The diagnostics use DiffusionVL on POPE–MSCOCO adversarial (Appendix D.4). For image I and question q, let $\mathcal{A} = \{\text{Yes}, \text{No}\}$, a be the answer position, and c_0 the candidate selected at its first valid native transfer. A *preference* compares logits; a *commitment* transfers the token.

3.1 LOW-MARGIN COMMITMENTS CONCENTRATE ERRORS

An answer may transfer before its alternatives are well separated. We measure *commitment maturity* by the margin aligned with the actual candidate:

$$m_0 = z^0_{c_0}(a) - \max_{y \in \mathcal{A} \setminus \{c_0\}} z^0_y(a). \tag{1}$$

Here $z^0(a)$ is read immediately before transfer. Alignment with c_0 retains negative margins if a sampler selects a non-argmax candidate. In the measured ordering of 3,000 answers (Figure 1, right), most errors lie toward the low-margin end. The overlap with correct answers makes margin a selection signal: it directs inspection toward less mature commitments without itself deciding which answers to revise.

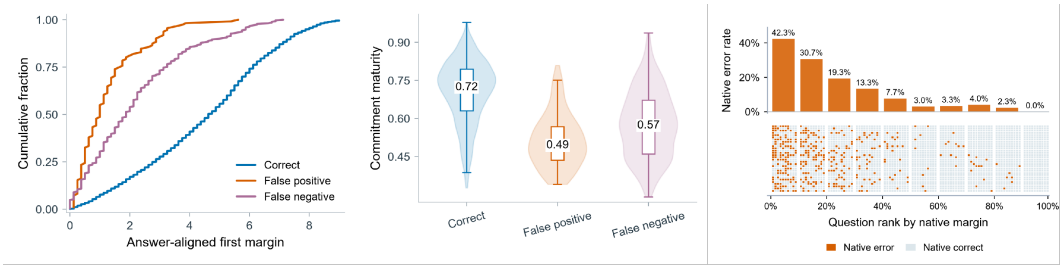


Figure 1: **Low-margin commitments concentrate errors.** Left and center: archived margin and maturity summaries. Right: independently measured native outcomes on 3,000 questions, ordered by candidate-aligned margin. The error concentration motivates selective inspection (sources and protocol in Appendix D.4).

3.2 NATIVE ITERATION CAN PRESERVE INCORRECT PREFERENCES

Continued decoding need not reconsider an emerging error. In Figure 2, left, other response positions resolve while the answer remains masked, yet successive readouts favor the same incorrect Yes candidate until transfer. Such trajectories recur in the fixed diagnostic cohort. Moreover, under the evaluated transfer rule, a written answer leaves the set of masked positions eligible for later updates.

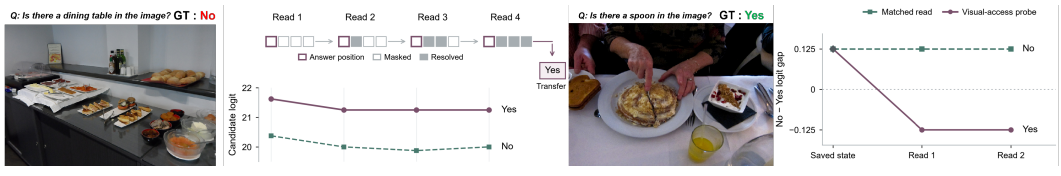


Figure 2: **Persistent preference and responsiveness to visual access.** Left: an incorrect Yes preference survives native updates until transfer. Right: two matched reads preserve No, while a $+2$ answer-to-visual-key mask bias favors Yes from the same saved state. The gap stays aligned with native No. Protocols and selection appear in Appendix D.4.

The separate spoon example probes whether a saved decision remains responsive to visual access. Both branches start from the same pre-transfer state with the answer masked; images, prompts, and cached states remain fixed. Matched reads preserve the incorrect No preference, whereas adding a

bias from the answer query to visual keys favors Yes. Native persistence and this paired response motivate explicit visual reinspection while the commitment can still be intercepted. Appendix D.4 gives the complete trajectories, matched conditions, and example-selection rules.

3.3 Shared Visual Focus Does Not Identify the Target

Reinspection also needs a semantic criterion. DiffusionVL answers Yes to both fork/knife and suitcase/backpack in Figure 3. The first query in each pair is a true positive and the second a false positive, yet their target-to-visual attention shares a dominant utensil or luggage region. Spatial overlap thus accompanies different semantic correctness.



Figure 3: **Shared focus leaves target identity ambiguous.** Fork/knife and suitcase/backpack receive Yes, although only the first query in each pair is correct. Native target-to-visual attention averages all heads and 28 layers, with a shared scale per pair. Red boxes guide presentation. This attention readout differs from the region–target score e.

These observations lead to distinct design requirements: use margin to allocate inspection, re-examine persistent answer preferences, and evaluate retrieved visual content against the named target. The resulting controller separates selecting an audit from authorizing a revision.

4 IVSR: Selective Commitment Auditing

IVSR intercepts a native commitment, collects additional evidence only when selected by a gate, and combines two readouts before submission (Figure 4). The backbone, prompt, and native transfer rule are unchanged. Appendix A provides the operational definitions and Algorithm 1.

4.1 Interception and Selection

Let x_t be the masked response canvas, $\tilde{x}_t$ the native token proposals, and $\mathcal{T}_t$ the positions selected for transfer. With lexical map λ from task candidates to accepted single-token encodings, the audit reference is

$$t^\star = \min\{t : a \in \mathcal{T}_t,\ \tilde{x}_{t,a} \in \lambda(\mathcal{A})\}, \qquad c_0 = \lambda^{-1}(\tilde{x}_{t^\star,a}). \tag{2}$$

We save $x_{t^\star}$ and the logits $z^0(a)$ before transfer. The maturity gate uses Eq. (1):

$$G = \mathbf{1}[m_0 < \tau_0(c_0)]. \tag{3}$$

For $G = 0$, the controller returns c_0 immediately. For $G = 1$, it initializes the reread from the saved state. Candidate-specific cutoffs accommodate the affirmative and negative margin distributions; selection alone never changes an answer.

4.2 Complementary Evidence Readouts

Reread answer evidence. A backbone-specific operator performs a fixed K-step visual reread:

$$(z^r, h_a^r, H_{\mathcal{V}}^r) = \mathcal{R}_\theta(I, q, x_{t^\star}, a; K, \ell). \tag{4}$$

The final readout supplies answer logits $z^r(a)$, answer state $h_a^r \in \mathbb{R}^d$, and visual states $H_{\mathcal{V}}^r \in \mathbb{R}^{|\mathcal{V}|\times d}$ at layer ℓ. The image, question, and parameters stay fixed while auxiliary states may evolve. We

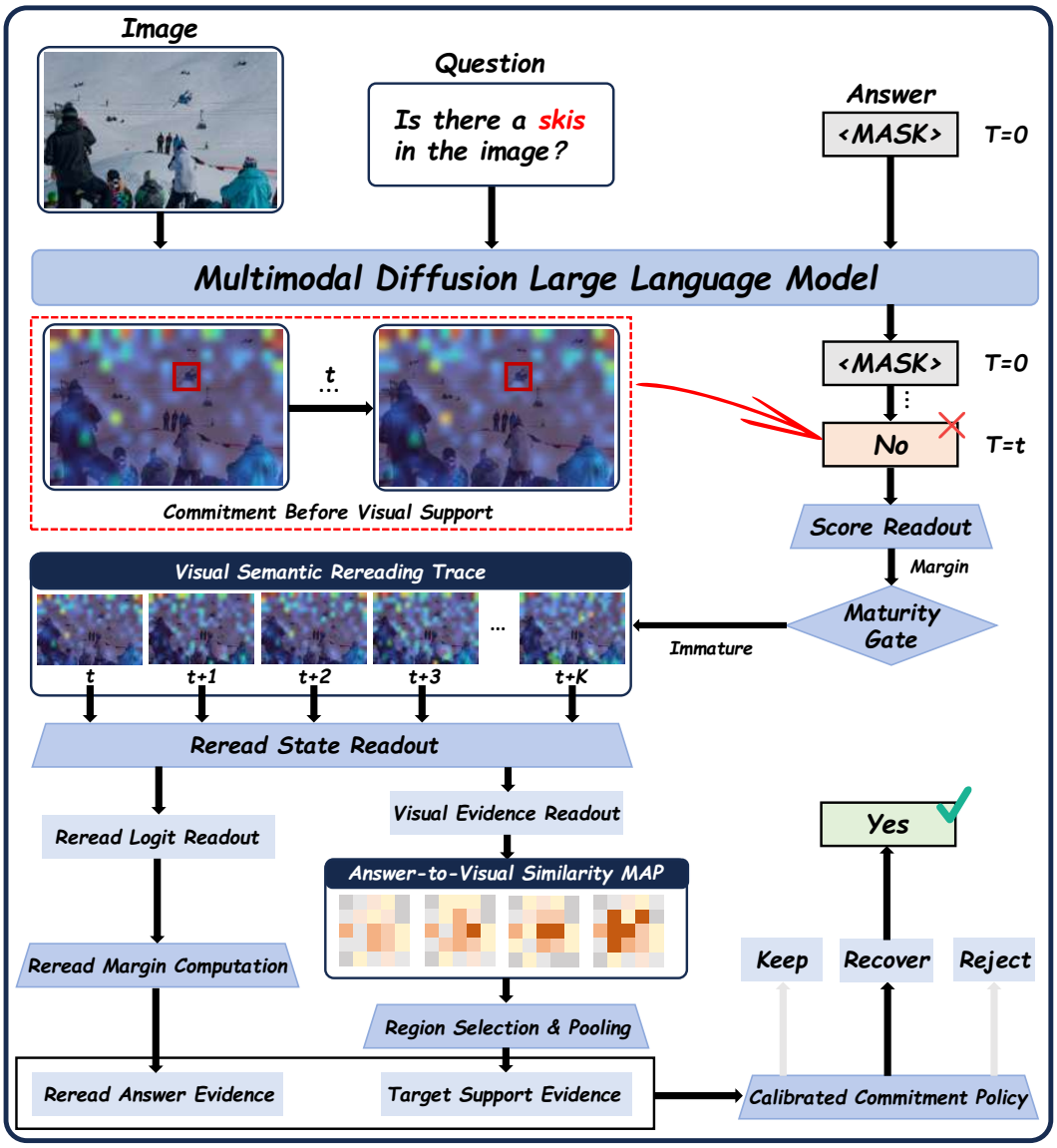


Figure 4: **IVSR framework.** The maturity gate selects native commitments for visual rereading. Final answer logits supply candidate margins; an answer-conditioned visual region supplies target alignment. Their calibrated combination determines retention, rejection, or recovery. The illustrated recovery path depicts schematic control and data flow.

retain the favored candidate and candidate-aligned margins:

$$c_r = \arg\max_{y \in \mathcal{A}} z_y^r(a), \tag{5}$$

$$m_r(c) = z_c^r(a) - \max_{y \in \mathcal{A} \setminus \{c\}} z_y^r(a). \tag{6}$$

The policy uses both $m_r(c_0)$ and $m_r(c_r)$: a negative $m_r(c_0)$ records that the alternative has overtaken the native commitment, information that a winner-only margin would discard.

Target-grounded regional evidence. A deterministic selector $P_q = \Omega(q)$ extracts the queried object, or the entity together with its queried attribute or relation. All tokenizer pieces contribute to

the contextual target representation:

$$u_q = \mathrm{norm}\left(\frac{1}{|P_q|}\sum_{j\in P_q} h_j^{\ell}\right), \qquad \mathrm{norm}(v) = v/\|v\|_2. \tag{7}$$

The reread answer state retrieves visual content by cosine relevance. We select the K_r highest-scoring visual tokens and pool their states:

$$\rho_i = \cos(h_a^r, h_i^r), \qquad V_r = \mathrm{TopK}_{K_r}(\mathcal{V}; \rho), \tag{8}$$

$$\alpha_i = \frac{\exp(\rho_i/\eta)}{\sum_{j\in V_r}\exp(\rho_j/\eta)}, \qquad r = \mathrm{norm}\left(\sum_{i\in V_r}\alpha_i h_i^r\right). \tag{9}$$

This internal region requires no annotated box. Its support for the named target is

$$e = \cos(r, u_q). \tag{10}$$

Retrieval is conditioned on the answer; evaluation is conditioned on the target. Thus, e complements logit separation with regional semantic compatibility, rather than reusing attention concentration as evidence of target identity.

4.3 Calibrated Retention and Revision

The binary policy requires agreement between reread answer evidence and target support:

$$\Gamma_{\mathrm{rej}} = G \wedge (c_0 = \mathrm{Yes}) \wedge [m_r(c_0) < \tau_r^-] \wedge [e < \tau_q^-], \tag{11}$$

$$\Gamma_{\mathrm{rec}} = G \wedge (c_0 = \mathrm{No}) \wedge (c_r = \mathrm{Yes}) \wedge [m_r(c_r) > \tau_r^+] \wedge [e > \tau_q^+]. \tag{12}$$

Rejection revises an affirmative commitment with weak support on both axes. Recovery requires a supported affirmative alternative to a native No. The submission is

$$\hat{y} = \begin{cases} \mathrm{Yes}, & \Gamma_{\mathrm{rec}} = 1, \\ \mathrm{No}, & \Gamma_{\mathrm{rej}} = 1, \\ c_0, & \text{otherwise.} \end{cases} \tag{13}$$

The edit predicates are mutually exclusive, and a changed reread preference alone cannot authorize revision. Fit-600 fits candidate cutoffs; disjoint Select-600 selects thresholds and K, K_r, ℓ, η before evaluation. Margin thresholds have logit units and target thresholds have cosine-similarity units. Appendix B reports the search and calibration protocol.

5 Experiments

We evaluate DiffusionVL, MMaDA, and LaViDa on POPE–MSCOCO, POPE–AOKVQA, MME, and AMBER-D existence (Zeng et al., 2025; Yang et al., 2025; Li et al., 2025; 2023; Fu et al., 2025; Wang et al., 2023), reporting POPE F1, MME scores, and hallucination rate (HR). Both POPE benchmarks include popular, random, and adversarial splits. All tasks use concise binary answers, including MME count, attribute, and reasoning questions.

We compare native decoding, VCD, PAI, DeCo, and IVSR. Baselines use their reported defaults; IVSR uses labeled Fit-600 and disjoint Select-600 calibration. Appendix B compares shared calibration access. Main evaluation, profiling, ablations, and diagnostics use their respective configurations and native references; comparisons are made within each series.

5.1 Main Results

IVSR improves all 18 POPE backbone–split comparisons over native decoding by 0.20–7.79 F1 points (mean: 2.34; Table 1). It leads 23 of 24 backbone–metric comparisons; PAI leads on LaViDa POPE–AOKVQA popular (86.48 versus 86.02).

MME increases by 46.6, 182.1, and 81.9 points for DiffusionVL, MMaDA, and LaViDa, respectively, while AMBER-D existence HR falls by 47.7%, 20.0%, and 16.1% relative to native decoding.

Table 1: Main results. POPE F1 and AMBER-D existence HR are percentages; bold marks each backbone's best result. DiffusionVL MME totals sum Table 8's category scores. Baselines use reported defaults; Table 5 compares shared calibration access.

	POPE–MSCOCO ↑			POPE–AOKVQA ↑			MME	AMBER-D
Inference	Pop.	Rand.	Adv.	Pop.	Rand.	Adv.	↑	HR ↓
DiffusionVL								
Native	88.07	89.51	86.63	90.90	91.70	84.20	2249.6	8.10
+VCD	88.49	90.79	87.04	90.97	92.03	84.31	2264.9	7.51
+PAI	88.78	90.53	87.47	91.07	92.55	84.42	2256.3	6.50
+DeCo	88.82	91.12	87.56	91.03	92.29	84.37	2272.2	7.11
+IVSR	**90.71**	**92.90**	**88.14**	**91.10**	**92.80**	**84.50**	**2296.2**	**4.24**
MMaDA								
Native	79.93	79.31	74.32	84.35	83.33	73.94	1213.4	20.76
+VCD	80.15	80.38	74.61	84.65	83.86	74.72	1250.6	19.50
+PAI	80.77	80.70	75.20	85.26	84.20	75.72	1345.0	17.67
+DeCo	81.01	81.53	74.91	84.96	84.63	75.14	1297.8	18.68
+IVSR	**81.86**	**82.51**	**75.48**	**85.45**	**85.05**	**76.42**	**1395.5**	**16.61**
LaViDa								
Native	81.64	83.24	79.93	84.73	87.87	80.35	1744.6	12.62
+VCD	84.23	86.59	81.57	84.50	89.64	79.31	1725.3	12.08
+PAI	82.87	86.51	81.04	**86.48**	88.67	80.75	1769.3	12.19
+DeCo	83.82	85.23	80.88	84.77	87.82	80.54	1795.6	11.78
+IVSR	**86.88**	**91.03**	**83.22**	86.02	**91.20**	**80.84**	**1826.5**	**10.59**

5.2 MME Category Analysis

Figure 5 compares all five methods across 14 MME categories on DiffusionVL.

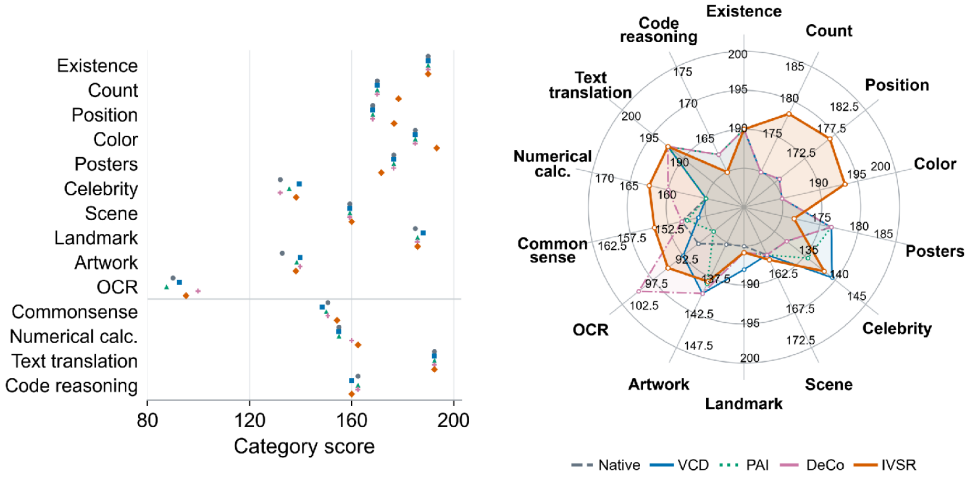


Figure 5: **MME categories on DiffusionVL.** Left: absolute scores; vertical offsets separate methods. Right: the same scores on category-specific radar axes, each spanning 20 points with 5-point rings. The common span preserves score differences; axis origins vary by category. Full scores appear in Table 8.

IVSR improves ten categories and ties two: count, position, and color each gain 8.3 points; numerical calculation gains 7.5; existence is unchanged. Posters and code reasoning decline by 4.9 and 2.5

points, and DeCo leads on OCR. Appendix D.1 gives the category scores whose sums supply the DiffusionVL MME totals.

5.3 Inference Efficiency

Table 2 reports runtime, resources, audit exposure, and profiling-series F1. Per-question latency excludes calibration.

Table 2: Inference profiling with hardware, batch size, generation length, and diffusion steps fixed within each backbone. Relative latency and extra FLOPs use native decoding as reference. Read/extra is the reread or extra-forward rate; F1 is POPE–MSCOCO adversarial.

	Runtime			Resources		Readout	
Method	Latency (ms)	Relative (×)	Throughput (samples/s)	Extra FLOPs (%)	Peak memory (GiB)	Read/extra (%)	F1 (%)
DiffusionVL							
Native	464.33	1.000	2.154	0.00	15.917	0.00	86.63
VCD	903.50	1.946	1.107	100.00	16.053	100.00	87.04
PAI	980.34	2.111	1.020	100.07	16.052	100.00	87.47
DeCo	644.76	1.389	1.551	4.92	15.808	0.00	87.55
IVSR (gated)	481.04	1.036	2.079	0.95	15.808	17.33	**88.14**
IVSR (all)	567.29	1.222	1.763	5.49	15.808	100.00	87.76
MMaDA							
Native	1889.59	1.000	0.529	0.00	18.893	0.00	74.33
IVSR (gated)	2073.86	1.098	0.482	3.13	17.827	18.87	**75.49**
LaViDa							
Native	1059.67	1.000	0.944	0.00	16.976	0.00	79.98
IVSR (gated)	1074.37	1.014	0.931	1.12	18.663	8.93	**83.16**

Gated IVSR adds 1.4–9.8% latency. On DiffusionVL, it takes 481.04 ms, below VCD (903.50), PAI (980.34), and DeCo (644.76). Rereading every example takes 567.29 ms and reduces F1 from 88.14 to 87.76, favoring selective inspection at this operating point.

Costs differ by resource: DiffusionVL adds 0.95% FLOPs and 3.6% latency; LaViDa memory rises from 16.976 to 18.663 GiB. Read/extra counts procedure-specific events, so DeCo's zero does not imply zero overhead. Appendix C details the accounting and protocol.

5.4 Ablation

Table 3: Ablations on POPE–MSCOCO adversarial F1 and AMBER-D existence HR (%). "No reread check" removes the margin predicate while retaining reread computation.

Backbone	Metric	Native	No gate	No reread check	No target evidence	Global	Full
DiffusionVL	F1 ↑	86.63	87.14	88.10	87.61	87.85	**88.14**
MMaDA	F1 ↑	74.32	75.34	75.21	74.87	75.04	**75.48**
LaViDa	F1 ↑	79.93	80.75	82.72	81.41	82.06	**83.22**
DiffusionVL	HR ↓	8.10	6.82	4.35	5.63	5.02	**4.24**
MMaDA	HR ↓	20.76	17.22	17.85	19.52	18.68	**16.61**
LaViDa	HR ↓	12.62	12.10	10.89	11.70	11.29	**10.59**

Table 3 evaluates POPE–MSCOCO adversarial F1 and AMBER-D existence HR. *No gate* inspects every commitment. *No reread check* removes only the answer-margin predicate, retaining the reread computation and visual-state readout. *No target evidence* drops the region–target constraint; *global* replaces the answer-conditioned region with a non-regional readout.

Full IVSR leads all six settings. Removing the gate costs 1.00, 0.14, and 2.47 F1 points on DiffusionVL, MMaDA, and LaViDa; global support costs 0.29, 0.44, and 1.16. Individual checks contribute unevenly: removing the reread-margin predicate changes DiffusionVL F1 by only 0.04 points.

Appendix B examines calibration and budgets; Appendix D.3 tests target wording. A fixed edit-level audit records 107 beneficial and 90 harmful changes (Appendix D.2), showing that aggregate gains coexist with intervention errors.

5.5 Reread-Length Sensitivity

Figure 6 shows the five measured reread lengths at a fixed gate. F1 peaks at $K = 4$, whereas longer rereads continue to increase latency. This one-factor sweep is distinct from joint calibration, which selects $K = 2$; Appendix B.3 provides the remaining sensitivity settings.

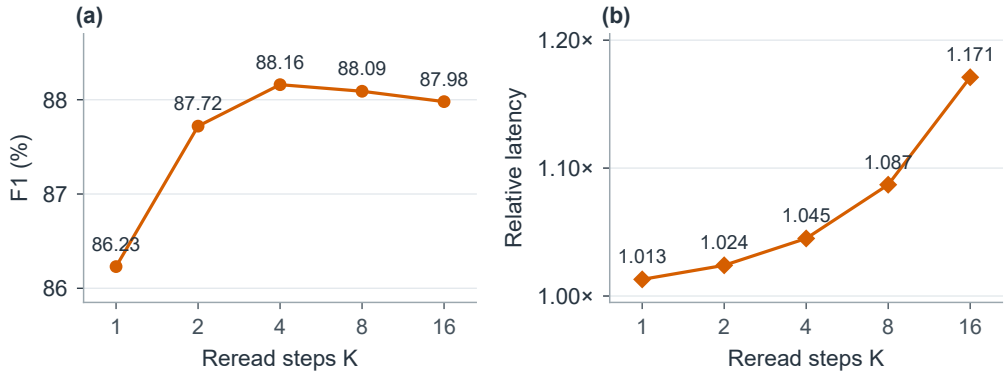


Figure 6: **Fixed-gate reread-length sensitivity on DiffusionVL.** The five measured values of K share a 19.57% reread rate and native budget $T = 8$. (a) POPE–MSCOCO adversarial F1. (b) Latency relative to native decoding. Point labels retain the exact reported values; connecting segments are visual guides. Longer rereads beyond $K = 4$ add latency without improving F1 in this sweep. This one-factor comparison is separate from the jointly selected $K = 2$ configuration in Table 5.

6 Conclusion

IVSR makes verification an explicit decision at a diffusion MLLM's first answer transfer. Selective rereading gathers complementary evidence: the answer margin measures candidate preference, while regional alignment tests compatibility with the question target. Their calibrated combination improves all 18 reported POPE settings and reduces existence hallucinations across three backbones. Separate inference profiling measures 1.4–9.8% latency overhead.

The resulting interface allocates verification computation before submission and ties each revision to the native decision it changes. It separates forming an answer from evaluating the visual evidence needed to retain it.

REFERENCES

Luís Felipe Cattelan and Danilo Silva. How to fix a broken confidence estimator: Evaluating post-hoc methods for selective classification with deep neural networks. In *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, volume 244 of *Proceedings of Machine Learning Research*, pp. 547–584, 2024. URL https://proceedings.mlr.press/v244/cattelan24a.html.

Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics*, 42(4):1–10, 2023. doi: 10.1145/3592116. URL http://dx.doi.org/10.1145/3592116.

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. Multi-modal hallucination control by visual information grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14303–14312, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Favero_Multi-Modal_Hallucination_Control_by_Visual_Information_Grounding_CVPR_2024_paper.html.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. MME: A comprehensive evaluation benchmark for multimodal large language models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-4899. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/d79a27cf2772fe00be7f341efc0eb517-Paper-Datasets_and_Benchmarks_Track.pdf.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/4a8423d5e91fda00bb7e46540e2b0cf1-Paper.pdf.

Michal Golovanevsky, William Rudman, Michael A. Lepori, Amir Bar, Ritambhara Singh, and Carsten Eickhoff. Pixels versus priors: Controlling knowledge priors in vision-language models through visual counterfacts. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 24837–24852, 2025. doi: 10.18653/v1/2025.emnlp-main.1262. URL https://aclanthology.org/2025.emnlp-main.1262/.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Sohee Kim, Soohyun Ryu, Joonhyung Park, and Eunho Yang. Unveiling the response of large vision-language models to visually absent tokens. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 21535–21557, 2025. doi: 10.18653/v1/2025.emnlp-main.1092. URL https://aclanthology.org/2025.emnlp-main.1092/.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13872–13882, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Leng_Mitigating_Object_Hallucinations_in_Large_Vision-Language_Models_through_Visual_Contrastive_CVPR_2024_paper.html.

Shufan Li, Konstantinos Kallidromitis, Hritik Bansal, Akash Gokul, Yusuke Kato, Kazuki Kozuka, Jason Kuen, Zhe Lin, Kai-Wei Chang, and Aditya Grover. LaViDa: A large diffusion language model for multimodal understanding. In *Advances in Neural Information Processing Systems*, volume 38, pp. 105101–105134, 2025. doi: 10.52202/

085713-3511. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/975affbe7b5f3b55fba62247b6877b1c-Paper-Conference.pdf.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305, 2023. doi: 10.18653/v1/2023.emnlp-main.20. URL https://aclanthology.org/2023.emnlp-main.20/.

Shi Liu, Kecheng Zheng, and Wei Chen. Paying more attention to image: A training-free method for alleviating hallucination in LVLMs. In *European Conference on Computer Vision*, pp. 125–140. Springer, 2024. doi: 10.1007/978-3-031-73010-8_8. URL https://link.springer.com/chapter/10.1007/978-3-031-73010-8_8.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 50608–50646, 2025. doi: 10.52202/085713-1689. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/48b383b24230e0e6e649d9c98dae4d8c-Paper-Conference.pdf.

Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 130136–130184, 2024. doi: 10.52202/079017-4135. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/eb0b13cc515724ab8015bc978fdde0ad-Paper-Conference.pdf.

Chenxi Wang, Xiang Chen, Ningyu Zhang, Bozhong Tian, Haoming Xu, Shumin Deng, and Huajun Chen. MLLM can see? dynamic correction decoding for hallucination mitigation. In *International Conference on Learning Representations*, 2025a. URL https://arxiv.org/abs/2410.11779.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and Jitao Sang. AMBER: An LLM-free multi-dimensional benchmark for MLLMs hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023. doi: 10.48550/arXiv.2311.07397. URL https://arxiv.org/abs/2311.07397.

Qidong Wang, Junjie Hu, and Ming Jiang. V-SEAM: Visual semantic editing and attention modulating for causal interpretability of vision-language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 17396–17420, 2025b. doi: 10.18653/v1/2025.emnlp-main.880. URL https://aclanthology.org/2025.emnlp-main.880/.

Zichuan Wang, Songlin Yang, Bo Peng, Zhenchen Tang, Yang Li, Beibei Dong, and Jing Dong. Same attention, different truths: Put logit-lens over visual attention to detect and mitigate LVLM object hallucination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25315–25325, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Wang_Same_Attention_Different_Truths_Put_Logit-Lens_over_Visual_Attention_to_CVPR_2026_paper.html.

Ling Yang, Ye Tian, Bowen Li, Xinchen Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. MMaDA: Multimodal large diffusion language models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 138867–138907, 2025. doi: 10.52202/085713-4636. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/caa934a507a952698d54efb24845fc4b-Paper-Conference.pdf.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. Woodpecker: Hallucination correction for multimodal large language models. *Science China Information Sciences*, 67(12):220105, 2024. doi: 10.1007/s11432-024-4251-x. URL https://link.springer.com/article/10.1007/s11432-024-4251-x.

Lunbin Zeng, Jingfeng Yao, Bencheng Liao, Hongyuan Tao, Wenyu Liu, and Xinggang Wang. DiffusionVL: Translating any autoregressive models into diffusion vision language models. *arXiv preprint arXiv:2512.15713*, 2025. doi: 10.48550/arXiv.2512.15713. URL https://arxiv.org/abs/2512.15713.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.00754.

AI USE STATEMENT

AI tools provided editorial and technical assistance; the authors are responsible for the final manuscript.

ETHICS STATEMENT

This work uses existing visual-language benchmarks, with no new human-subject study or annotation collection. Model and image use remains subject to the corresponding licenses. The controller is evaluated as a research method for visual verification; high-stakes deployment requires independent validation and oversight.

REPRODUCIBILITY STATEMENT

Section 4 and Algorithm 1 define the controller. Appendices A–C specify the recorded readouts, calibration, profiling, and implementation scope. Appendix D gives exact category scores, edit counts, target rewrites, and mechanism protocols; Appendix E derives the accounting identities. The source package includes all tables and figures. Diagnostic records and plotting code are retained separately in the research artifacts.

A Operational Details

Table 4 summarizes the controller's inputs and readouts. Algorithm 1 specifies their use at the native answer transfer.

Table 4: Notation and tensor sources. The tokenizer vocabulary $\mathcal{W}$, visual-token positions $\mathcal{V}$, lexical map λ, and readout layer ℓ are distinct objects.

Symbol	Meaning and source
I, q	Input image and question; unchanged by the controller.
$x_t, \mathcal{T}_t$	Partially masked response canvas and the native decoder's transfer set at execution step t.
$a, \mathcal{A}, \lambda$	Answer position, binary task candidates $\{\mathrm{Yes}, \mathrm{No}\}$, and their accepted lexical encodings.
$t^\star, c_0$	First valid candidate transfer and the native candidate selected at that event.
$z^0(a), z^r(a)$	Vocabulary-logit vectors at the native transfer and the final reread readout, respectively.
$m_0, m_r(c)$	Native and reread margins aligned with a task candidate; both are in logit units.
K, ℓ, L	Number of reread steps, selected readout layer, and backbone depth.
$h_a^r, H_{\mathcal{V}}^r$	Final reread answer state and visual-token states at the chosen layer.
$P_q = \Omega(q), u_q$	Deterministically selected target span and its normalized contextual mean.
K_r, η	Number of retrieved visual tokens and softmax pooling temperature.
V_r, ρ_i, α_i, r	Retrieved indices, answer–visual relevance, normalized pooling weights, and the resulting region vector.
e	Cosine similarity between region vector r and target vector u_q.
$\tau_0, \tau_r^\pm, \tau_q^\pm$	Calibrated decision thresholds; quantile search coordinates are not interchangeable with fitted raw thresholds.

A.1 Native Commitment and Answer Encoding

The native trajectory determines $t^\star$, $x_{t^\star}$, and c_0 before intervention. Every retention, rejection, or recovery decision uses the saved c_0 as its reference. Auxiliary hidden states may change during rereading, while the image, question, and model parameters remain fixed. The controller compares the binary candidates without sampling an independent full answer.

The first-transfer definition assumes the evaluated concise-answer interface. For decoders that later overwrite the answer, this event may be provisional; their transfer and submission events require an explicit correspondence. The single-position equations also require a separate extension for multi-token answers.

A.2 Target Extraction

For "Is there a traffic light in the image?", $\Omega(q)$ selects every tokenizer piece of "traffic light". Attribute and relation queries select the target entity together with the queried attribute or relation phrase. The selected contextual states are mean-pooled and normalized, using one target span per question.

This target representation supports multi-word phrases with single-token Yes/No answers. The rewrite audit evaluates target wording; unrestricted output synonyms and multi-token answer scoring remain unevaluated. Equations (7)–(10) require nonempty spans and nonzero pooled vectors. Exact tokenizer matching and exceptional-case behavior are unspecified in the available implementation record.

A.3 Backbone Configuration

The final reread answer state queries visual tokens by cosine similarity; top-K_r selection and softmax pooling with temperature η produce the region vector. Region and target vectors must occupy compatible hidden spaces. The selected reread length K is fixed within each experiment; the gate determines which examples are inspected.

The benchmark record omits several details required to reproduce the backbone adapters: the tensor-level visual-access modification, mask/cache updates, visual-token indexing, answer aliases, prompt/visual readout synchronization, and some fitted scalar thresholds. Its joint configuration

reports a visual-bias setting of 2.0 without specifying the corresponding tensor operation. Appendix D.4.2 fully specifies the independently implemented cached visual-bias probe; the benchmark controller and this diagnostic use distinct implementations.

Algorithm 1 IVSR at the first valid binary answer transfer

Require: Frozen model f_θ, native decoder, reread adapter $\mathcal{R}_\theta$, image I, question q, answer position a, lexical map λ, target selector Ω
Require: Fixed $K, K_r, \ell, \eta, \tau_0(\cdot), \tau_r^-, \tau_r^+, \tau_q^-, \tau_q^+$
Run native decoding to $t^\star$; save $x_{t^\star}$, c_0, and $z^0(a)$ using Eq. (2)
Compute m_0 and G using Eqs. (1) and (3)
if $G = 0$ **then**
    return c_0
end if
Run the fixed K-step reread from the saved checkpoint; retain the final readout
Compute c_r and m_r using Eqs. (5) and (6)
Extract $P_q = \Omega(q)$ and pool contextual states into u_q using Eq. (7)
Retrieve V_r, pool r, and compute e using Eqs. (8)–(10)
if Γ_{rec} **then**
    return Yes
else if Γ_{rej} **then**
    return No
else
    return c_0
end if

B Calibration and Operating-Point Selection

B.1 Fit, Selection, and Evaluation

Fit-600 supplies candidate thresholds, Select-600 selects the rule and readout configuration, and 3,000 held-out examples evaluate the selected configuration. The partitions are reported as mutually disjoint; sample identifiers are unavailable for an independent overlap check. Model weights remain frozen, and all selected settings, including K, are fixed before evaluation.

For calibration scores $\{s_i\}$ and quantile coordinate p, the fitted threshold is $\hat{Q}_{\mathrm{fit}}(p) = \inf\{v : \hat{F}_{\mathrm{fit}}(v) \geq p\}$. This adapts thresholds to the backbone's score scale while retaining the units of s_i. The coordinate $p \in [0, 1]$ specifies a quantile; the deployed policy uses its corresponding raw threshold and branch assignment in Section 4.3.

B.2 Matched-Calibration Comparison and Search

With the same calibration partitions, IVSR attains F1 88.19 versus 87.79 for PAI, the strongest tuned comparator, at 482.52 versus 999.85 ms per example (Table 5). This evaluation has its own native reference and configuration, separate from Table 1.

Table 6 lists the grids and selected settings. Each baseline grid contains 16 combinations; the IVSR grid contains 256 if all listed coordinates vary independently. The comparison controls calibration-data access; realized trial counts and subsampling were not recorded. Native decoding requires no search.

Table 5: Matched-calibration comparison on DiffusionVL with 3,000 POPE adversarial examples. Tuned methods share calibration partitions. This series has a separate configuration and native reference from Table 1.

Method	F1 (%) ↑	Latency (ms) ↓	Peak memory (GiB)
Native	86.69	470.27	15.918
VCD	86.14	917.76	16.054
PAI	87.79	999.85	16.053
DeCo	86.73	652.22	15.808
IVSR	**88.19**	482.52	15.810

Table 6: Search configurations for the matched-calibration experiment. Baseline grids and the IVSR joint grid each use the same calibration partitions. The threshold entries have an unresolved mapping to fitted raw-score cutoffs.

(a) Baseline searches

Method	Values tested	Selected configuration
VCD	$\alpha \in \{0.5, 1.0, 1.5, 2.0\}$; $\beta \in \{0.01, 0.05, 0.1, 0.2\}$; noise step 500.	$\alpha = 1.0$, $\beta = 0.2$; noise step 500.
PAI	Attention $\alpha \in \{0.1, 0.2\}$; $\gamma \in \{1.0, 1.1, 1.3, 1.5\}$; $\beta \in \{0.05, 0.1\}$; layers $[2, 28]$; blank-image negative branch.	$\alpha = 0.2$, $\gamma = 1.3$, $\beta = 0.05$; layers $[2, 28]$; blank-image negative branch.
DeCo	$\alpha \in \{0.2, 0.4, 0.6, 0.8\}$; top-$p \in \{0.8, 0.9\}$; top-$k \in \{10, 20\}$; early-exit layers 20–28.	$\alpha = 0.2$, top-$p = 0.8$, top-$k = 10$; early-exit layers 20–28.

(b) IVSR joint search

Setting	Values tested	Selected configuration
Reread length K	$2, 4$	2
Region fraction $K_r/\lvert\mathcal{V}\rvert$	$0.05, 0.1$	0.1
Readout layer ℓ	$21, 28$	28
Pooling temperature η	$0.1, 0.3$	0.3
Gate setting	0.2	0.2
Negative reread setting	$0.1, 0.2$	0.2
Negative target setting	$0.1, 0.2$	0.1
Positive reread setting	$0.8, 0.9$	0.9
Positive target setting	$0.8, 0.9$	0.8
Visual-bias setting	2.0	2.0
Reread-budget setting	0.2	0.2

The calibration record describes quantile fitting but leaves the mapping from panel (b)'s threshold settings to final raw-score cutoffs ambiguous. "Visual bias" controls the reread adapter; the decision signals remain those in Eqs. (11)–(12).

B.3 Sensitivity: Score Settings and Reread Length

The one-factor analyses hold all other settings fixed. Table 7 reports the available extrema, and Figure 6 shows all five reread-length measurements. These sweeps use different configurations from the joint search in Table 6. Relative depth and region fraction accommodate different backbone depths and visual-token counts; thresholds remain backbone-specific.

Table 7: Reported one-factor sensitivity ranges. F1 is in percent. Gate candidates are fitted from Fit-600; the joint rule and readout settings are selected on Select-600. The p values below are quantile/decile coordinates rather than raw logit margins.

Setting	Values tested	Selected	Worst F1	Best F1
Region fraction	$0.025, 0.05, 0.1, 0.2, 0.4$	0.1	87.47	87.91
Relative depth ℓ/L	$0.25, 0.5, 0.75, 1$	1	86.52	87.79
Temperature η	$0.03, 0.1, 0.3, 1$	0.3	86.52	87.95
Gate quantile	$0.10, 0.15, 0.20, 0.25, 0.30$	$p = 0.20$	86.68	87.82
Reread thresholds	Fit deciles 0.1–0.9	$p = 0.10$	86.57	87.94
Target thresholds	Fit deciles 0.1–0.9	$p = 0.20$	87.26	88.06

At the fixed 19.57% reread rate, F1 peaks at $K = 4$. Joint selection instead chooses $K = 2$ with the other settings in Table 6, yielding F1 88.19. Here $T = 8$ is the native step budget and K counts auxiliary reread steps; the five K/T values are $0.125, 0.25, 0.5, 1, 2$.

B.4 Budget-Constrained Operating Points

Figure 7 varies the maximum reread rate during rule selection and evaluates each selected rule with fixed settings. The highlighted random and adversarial points reread 9.7% and 18.9% of examples. This sweep, the profiling comparison (Table 2), and the intervention audit (Table 9) use distinct configurations and operating points.

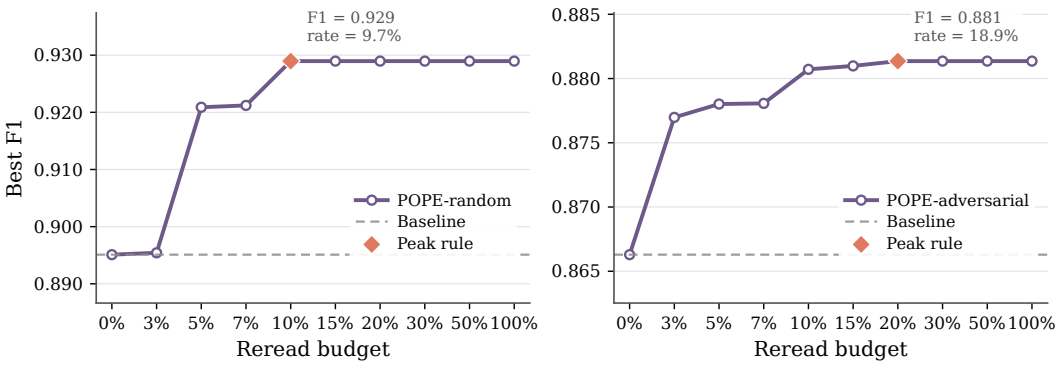


Figure 7: Reread-budget curves on POPE random and adversarial. The dashed line denotes native decoding. Highlighted points attain F1 0.929 at a 9.7% reread rate and F1 0.881 at 18.9%, respectively. F1 is shown as a fraction. "Best F1" denotes a selected operating point, not an uncertainty bound.

C Profiling and Compute Cost

Table 2 holds hardware, batch size, generation length, and diffusion steps fixed within each backbone comparison. The accelerator model, software versions, numerical precision, timing, and warm-up procedure are unavailable, limiting reproduction of absolute latency.

Profiling, main evaluation, and calibration use separate configurations and native references. LaViDa's native/IVSR F1 is 79.93/83.22 in the main evaluation and 79.98/83.16 in profiling. DiffusionVL's gate-removal ablation yields 87.14, while the profiled all-reread configuration yields 87.76. Relative latency uses the native run from the same series.

Cost accounting. Let $r_{\text{read}} = \Pr(G = 1)$, C^r_{step} be the cost of one reread step, and C_{target} the region/target readout cost on an inspected example. The additional arithmetic cost is approximately

$$C_{\text{extra}} = C_{\text{gate}} + r_{\text{read}}\big(KC^r_{\text{step}} + C_{\text{target}}\big). \tag{14}$$

The approximation $r_{\text{read}}K/T$ assumes comparable native and auxiliary step costs; wall-clock latency also depends on caching, memory traffic, synchronization, and readout overhead. DeCo's

intermediate-layer readouts incur FLOPs despite a zero extra-forward rate. The reported FLOPs are percentages above native decoding; absolute operation counts are unavailable.

Units and derived values. Latency is in milliseconds per example, throughput in examples per second, and peak memory in GiB. Relative latency is the displayed latency divided by its native reference; for example, VCD gives $903.50/464.33 = 1.946\times$. Throughput equals $1000/\text{latency (ms)}$ at reported precision. Baseline and all-reread measurements are available only for DiffusionVL.

D Task Coverage and Intervention Diagnostics

D.1 MME Category Coverage

Table 8 gives the 14 category scores for the five DiffusionVL methods in Figure 5. These tasks evaluate binary verification.

Table 8: DiffusionVL MME category scores for Figure 5. Totals sum the 14 displayed scores, matching Table 1.

Category	Native	VCD	PAI	DeCo	IVSR
Existence	190.0	190.0	190.0	190.0	190.0
Count	170.0	170.0	170.0	170.0	178.3
Position	168.3	168.3	168.3	168.3	176.6
Color	185.0	185.0	185.0	185.0	193.3
Posters	176.5	176.5	176.5	176.5	171.6
Celebrity	132.0	139.5	135.5	132.0	138.2
Scene	159.3	159.3	159.3	159.3	160.0
Landmark	185.0	188.0	185.8	185.8	185.8
Artwork	132.8	139.8	138.4	139.8	138.1
OCR	90.0	92.5	87.5	99.8	95.0
Commonsense reasoning	150.7	148.5	150.0	150.7	154.3
Numerical calculation	155.0	155.0	155.0	160.0	162.5
Text translation	192.5	192.5	192.5	192.5	192.5
Code reasoning	162.5	160.0	162.5	162.5	160.0
Total	2249.6	2264.9	2256.3	2272.2	2296.2

The radar in Figure 5 uses a common 20-point span with category-specific origins. For scores s_{mj} of method m on category j, the origin is $b_j = \min\{180,\ 2.5\lfloor \min_m s_{mj}/2.5 \rfloor - 5\}$ and the plotted radius is $(s_{mj} - b_j)/20$. All methods share the same axis for each category, and adjacent labeled rings differ by five points. Comparisons are made within spokes; the tabulated sum defines the MME total. Summing all 14 categories gives 2249.6, 2264.9, 2256.3, 2272.2, and 2296.2 for native decoding, VCD, PAI, DeCo, and IVSR, respectively. Table 1 uses these same category-derived totals. The table and figure use one reported five-method series throughout.

D.2 Intervention Outcomes

On the fixed diagnostic split in Table 9, IVSR increases affirmative predictions and recall while lowering precision. F1 rises by 1.33 points and accuracy by 0.57 points. This configuration is separate from the main evaluation, profiling, and matched-calibration series.

Table 9: Intervention outcomes on a fixed diagnostic split. (a) Aggregate percentages and percentage-point changes. (b) Edit counts and resulting F1; net edits are beneficial minus harmful changes. Native F1 is 86.71.

(a) Aggregate behavior

Metric	Native	IVSR	Change (pp)
Accuracy	87.40	87.97	+0.57
Precision	91.74	87.49	-4.25
Recall	82.20	88.60	+6.40
F1	86.71	88.04	+1.33
Yes ratio	44.80	50.63	+5.83
Reread rate	0.00	9.60	+9.60

(b) Beneficial and harmful edits

Policy	F1 (%)	Changed	Wrong→ correct	Correct→ wrong	Net
Reject only	86.83	11	8	3	+5
Recover only	87.92	186	99	87	+12
Full IVSR	88.04	197	107	90	+17

Rejection changes 8 false positives to true negatives and 3 true positives to false negatives. Recovery changes 99 false negatives to true positives and 87 true negatives to false positives. The disjoint branches therefore produce 197 edits: 107 beneficial and 90 harmful.

These edits yield 17 net corrected decisions and edit precision $107/197 \approx 54.3\%$. Equation (16) relates this net count to accuracy; F1 also depends on precision and recall. The higher Yes ratio shows that the gain accompanies increased affirmative coverage.

D.3 Controlled Target Rewrites

The rewrite diagnostic contains four groups of 12 examples (Table 10). Overall accuracy remains 83.3%, while consistency falls by 4.2 points. Synonym and relation accuracy each increase by one example; multi-word/template and attribute accuracy each decrease by one.

Table 10: Target-rewrite diagnostic. Scores are percentages; consistency measures agreement under rewrites. Each group has 12 examples, so one correct answer corresponds to approximately 8.3 accuracy points.

Target rewrite	N	Native consistency	IVSR consistency	Native accuracy	IVSR accuracy
Synonym	12	83.3	83.3	66.7	75.0
Multi-word / template	12	100.0	91.7	83.3	75.0
Attribute	12	100.0	91.7	100.0	91.7
Relation	12	83.3	83.3	83.3	91.7
Overall	48	91.7	87.5	83.3	83.3

Pooling prompt states represents multi-token targets, but span extraction and contextual representations remain sensitive to wording. The individual rewritten pairs and full tokenizer-level parsing specification are unavailable, preventing exact reconstruction of this evaluation.

D.4 Mechanism Diagnostics

Native commitment protocol. The sample-level diagnostics use DiffusionVL-Qwen2.5VL-7B in BF16 evaluation mode, generation length 8, eight denoising steps per block, temperature 0, and the official low-confidence-static transfer rule. The prompt requests a Yes/No answer to the unmodified POPE question. An observer records logits and the selected token immediately before answer transfer. The lexical candidates are Yes (9454) and No (2753); the native full-vocabulary sampler is unchanged.

Margins align with the transferred c_0. Pilot checks verified identical outputs to the uninstrumented generator and agreement between c_0 and the completed answer.

Measured margin ranking. The right panel of Figure 1 uses all 3,000 POPE–MSCOCO adversarial questions on 500 images, with no missing records. Native counts are 1,233 true positives, 1,389 true negatives, 111 false positives, and 267 false negatives. Questions are ordered by increasing m_0, breaking exact ties with a label-independent permutation (seed 3407). Ten consecutive blocks of 300 questions have error counts $127, 92, 58, 40, 23, 9, 10, 12, 7, 0$. Each mosaic cell represents one question; rank increases down columns, then left to right within each block. Labels determine color only. The first two blocks contain $219/378 = 57.9\%$ of errors, leaving 159 elsewhere. Deployment-gate calibration follows Appendix B.

Figure 8 shows the complete raw-margin empirical distributions and error-capture curve. Inspecting 10% and 20% of questions captures 127 and 219 of the 378 errors. The diagonal gives expected capture under random selection.

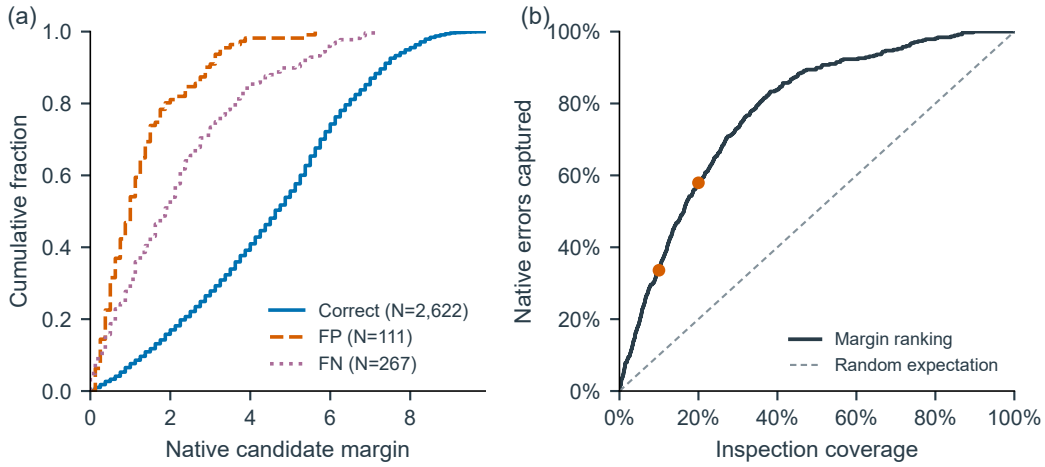


Figure 8: **Distribution and inspection value of the native margin.** (a) Empirical cumulative distributions for native-correct answers ($N = 2{,}622$), false positives ($N = 111$), and false negatives ($N = 267$), including their full observed ranges. (b) Fraction of all 378 native errors captured by inspecting questions in increasing margin order. Markers indicate 10% and 20% inspection coverage; the dashed diagonal is the expectation under random selection. Both panels use the same 3,000-question diagnostic as the right panel of Figure 1.

Archived distribution graphics. The left and center panels of Figure 1 are archived summaries recreated from vector paths. Their sample records and normalized-maturity transformation are unavailable. Quantitative margin analysis uses the independently measured raw margins in the right panel and Figure 8.

D.4.1 Native Preference Persistence

The left example in Figure 2 comes from 300 questions on 50 images sampled uniformly without replacement from the same 500-image pool (NumPy seed 20270925). All six questions per image were retained, giving 150 Yes and 150 No labels. This exploratory cohort was reused from earlier diagnostics in full.

The observer records every answer-block forward while the answer remains masked, through its native transfer. Pre-transfer preferences are provisional. For analysis only, gaps align with the official label: $g_t = z_y^t(a) - z_{\bar{y}}^t(a)$, where $\bar{y}$ is the other candidate. Strict errors have $g_t < 0$ and exact ties have $g_t = 0$. All 1,587 full-vocabulary proposals are Yes or No.

All 33 initially strict errors, from 25 images, persist through transfer across three to eight readouts each. Of the other 267 questions, 266 remain strictly correct. Question 1697 moves from $+0.125$ to an exact tie, resolved to No by the native token-order rule, yielding a false negative. There are no strict sign reversals. Final counts are 128 true positives, 138 true negatives, 12 false positives, and 22 false negatives (Figure 9).

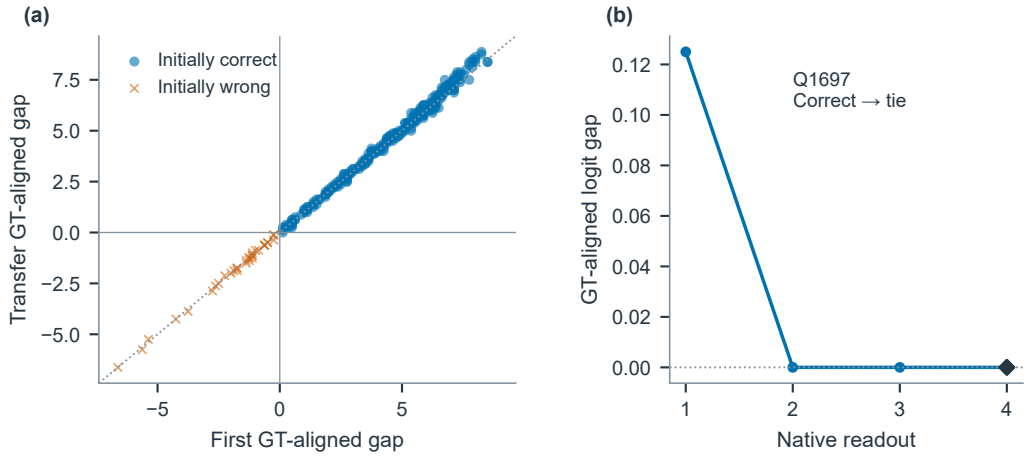


Figure 9: **Complete-cohort check of native preference persistence.** (a) First-readout versus transfer gaps for all 300 questions, aligned with the official answer; negative values favor the incorrect candidate. The dotted diagonal indicates an unchanged gap. (b) The sole change in binary outcome: question 1697 moves from a strictly correct preference to tied logits, resolved to an incorrect No at transfer. No strict gap changes sign in the cohort.

Illustrative trajectory. Question 1124 on COCO image 011241 asks about a dining table. Among the 12 native false positives, its margin is closest to their median of 1.1875, with ties broken by question ID. The Yes-minus-No gaps are $1.25, 1.25, 1.375, 1.25$; transfer occurs at readout four. The Yes/No logits are $(21.625, 20.375)$, $(21.25, 20.0)$, $(21.25, 19.875)$, and $(21.25, 20.0)$. Masked local positions are successively $\{4, 5, 6, 7\}$, $\{4, 6, 7\}$, $\{4, 7\}$, and $\{4\}$, with answer position 4. Figure 2 shows these mask states and the transfer. The official POPE annotation is No for dining table in this buffet-counter scene.

The six-question pilot verified identical outputs to the uninstrumented generator and bitwise preservation of the prefix cache. All 300 trajectories are complete; native outputs and transfer logits exactly match the earlier 3,000-question records. Transfer selects still-masked positions, and the current block reuses the prefetched visual-prefix key/value cache.

D.4.2 Paired Visual-Access Probe

The right example in Figure 2 uses question 1027 on COCO image 153300: the spoon label is Yes and the native answer is No. A fixed-case run uses the same checkpoint and decoding settings. Pilot and formal traces agree, including the native transfer and auxiliary logits from the original diagnostic. This saved-state probe is implemented separately from the benchmark controller.

Common state and matched branches. Both branches copy the same pre-transfer block embeddings, attention mask, position IDs, and random state, sharing a read-only prefix key/value cache. The answer remains masked, and c_0 serves as the comparison reference. At every decoder layer and head, the visual branch modifies the additive attention mask B by

$$B_{ij}^{(\beta)} = B_{ij} + \beta\, \mathbf{1}[i = a,\ j \in \mathcal{V}], \qquad \beta = 2. \tag{15}$$

Here $\mathcal{V}$ is the processor's image-token expansion; the matched branch uses $\beta = 0$. Image pixels, prompt, cached visual keys/values, model weights, and all other mask entries remain fixed. Each

branch executes $K = 2$ current-block forwards with identical tensor shapes. These are auxiliary decoder readouts; the image encoder is unchanged. The model has no explicit diffusion-timestep argument.

Between reads, the update rule transfers the highest-confidence other masked site, if present, using the native greedy sampler. Only the audited answer remains masked in this case, so both branches retain their token states and repeat the first readout. Identical-forward replay and cache checks verify numerical consistency and preservation of the saved prefix.

Candidate-aligned trace. All eight native proposals are No. Reads 1–7 have exact BF16 ties, resolved to No by token order. At transfer (read 8), $(z(\mathrm{Yes}), z(\mathrm{No})) = (20.000, 20.125)$. Both matched reads reproduce this pair; both visual reads give $(20.125, 20.000)$. The native-candidate margin changes from $+0.125$ to -0.125, favoring the correct Yes candidate.

Selection and aggregate outcomes. This example was selected from the 3,000-question run as the beneficial candidate change with a strictly positive native margin. The probe produced 11 beneficial and 3 harmful candidate changes overall; the matched branch preserved every native candidate.

The controlled factor is a global bias to visual keys, with equal forward counts and a common saved state. The 0.125 BF16 gap is one representable increment at this logit magnitude. Target-pixel attribution and full-precision behavior remain untested. The probe omits the maturity gate and target-support policy, so these candidate changes are distinct from full-controller benchmark edits.

D.4.3 Native Target-Attention Readout

Figure 3 pairs fork/knife on COCO image 003845 and suitcase/backpack on COCO image 120648. At native commitment, the readout averages target-to-visual attention over all target tokenizer pieces, heads, and 28 decoder layers. This differs from the controller's answer-state cosine relevance ρ_i and region–target score e.

Overlays use Gaussian smoothing ($\sigma = 0.6$) on the visual-token grid, then bilinear interpolation to image resolution. Each TP/FP pair shares a ceiling at the joint 99th percentile of its smoothed grids. For interpolated response v and ceiling v_{99}, $u = \mathrm{clip}(v/v_{99}, 0, 1)$ maps through inferno with opacity $0.08 + 0.57u$. No spatial mask or target-dependent enhancement is applied. Scales are shared within pairs only. Red rectangles are editorial guides, separate from ground-truth boxes, model inputs, and V_r.

Both queries in each pair receive Yes but have different POPE labels, giving the grounded/hallucinated outcome labels. The plots show native attention and contain no pixel intervention or controller correction.

E Edit and Compute Accounting

The accuracy effect of binary revision depends on the frequency and correctness of its edits.

Proposition 1 (Net accuracy change)**.** *Consider N examples with ground-truth binary labels y_i, native predictions c_i, and revised predictions $\hat{y}_i$. Let B count edits from an incorrect to a correct answer, and H count edits from a correct to an incorrect answer. Then*

$$\mathrm{Acc}(\hat{y}) - \mathrm{Acc}(c) = \frac{B - H}{N}. \tag{16}$$

If every edit changes a binary answer and $M = B + H > 0$, this is also

$$\mathrm{Acc}(\hat{y}) - \mathrm{Acc}(c) = \frac{M}{N}\left(2\frac{B}{M} - 1\right). \tag{17}$$

Proof. For each example, subtract $\mathbf{1}[c_i = y_i]$ from $\mathbf{1}[\hat{y}_i = y_i]$. The difference is $+1$ for a beneficial edit, -1 for a harmful edit, and zero for an unchanged answer. Summing and dividing by N gives the first identity. In binary classification, changing the answer reverses correctness, so every changed example belongs to one of the two edit sets and $M = B + H$. Substituting $H = M - B$ gives the second identity. □

With more than two labels, an edit can change one incorrect answer to another. Equation (16) still holds, but $M = B + H$ need not. F1 additionally depends on the direction of edits through precision and recall.

Gate-conditioned form. Let G mark an audit, D mark a changed answer, and suppose the controller only changes gated examples. Define $r = \Pr(G = 1)$, $u = \Pr(D = 1 \mid G = 1)$, and $p = \Pr(c_0 \neq y, \hat{y} = y \mid D = 1)$. For binary decisions and well-defined conditional probabilities,

$$\mathrm{Acc}(\hat{y}) - \mathrm{Acc}(c_0) = ru(2p - 1). \tag{18}$$

This follows from $\Pr(D = 1) = ru$ and Eq. (17). With no edits, accuracy is unchanged and p may be undefined. The factors measure audit rate (r), conditional edit frequency (u), and edit precision (p). A stricter gate can reduce both beneficial and harmful edits; its accuracy effect depends on the resulting p.

The diagnostic's $B = 107$ and $H = 90$ give 17 net corrected decisions and $p \approx 0.543$. Its population size is unspecified in the source record, precluding an independently verified count-to-rate conversion or identification with another evaluation split.

Compute cost. Equation (14) gives the approximate extra cost. Increasing K changes per-audit cost and may change edit precision; a fixed gate retains the same audited examples. Thus the one-factor length sweep has a constant reread rate, while its F1 and latency vary.