PAVIR: WHEN SHOULD MULTIMODAL CASCADES REVISE, NOT ADOPT?

Anonymous authors
Paper under double-blind review

ABSTRACT

What cascades treat as progress, a stronger visual answer, rests on an untested assumption: that the expert's answer is the decision the task scores, even when a required claim can make Conflict correct. We test that assumption on public ScienceQA, MMStar, and MMMU-Pro items with aligned, conflicting, and missing claims, separating who sees the text from whether the expert writes a full answer. PAVIR keeps the claim and the source requirements, calls an image-only expert only when the sources already disagree, and recomputes Answer, Conflict, or RequestEvidence with one rule. With the rule fixed, isolating the observer raises ScienceQA accuracy by 9.66 points, and the gain sits in the conflict cells; the same isolation lowers accuracy when every output must be an answer. On MMStar, disagreement-gated revision reaches 88.67% at 399/1,200 calls (61 repairs, 15 harms), from 84.83% with no calls; a 256-token 32B model that sees the claim reaches 87.25%, and 88.50% only after the same rule is applied (65 repairs, 21 harms; paired interval $[-1.16, 1.50]$), while calling all 494 eligible disagreements yields 88.58%. The comparison rule, not the ranking score or the size of the generator, is what makes the call useful, and only when Conflict is a legal output.

1 INTRODUCTION

A visual expert can answer an image question correctly and still leave the task unfinished. In Figure 2(a) the chart supports B and a required claim names A, so the complete decision is Conflict. Calling a stronger image model may improve the visual observation, but returning B answers a different question. Cascades that route by expected quality (Chen et al., 2024b; Ding et al., 2024; Ong et al., 2025) and methods that revise a single string (Aggarwal et al., 2024; Gou et al., 2024; Asai et al., 2024) do not keep this distinction: once the claim is discarded, a later visual prediction cannot be compared with it.

PAVIR keeps the claim, the source requirements, and the current decision, and lets the expert update only the visual observation (Figure 1). The same comparison then maps the original or the updated observation to Answer, Conflict, or RequestEvidence. The default caller spends an expert only when the sources already disagree. We test three consequences. Isolating the observer should help only when Conflict is a legal output. An image-only observation plus that comparison should match, rather than lose to, a model that writes the full decision. After the disagreement gate is shared, the ranking score should matter less than the gate.

2 RELATED WORK

Image–text cascades decide when to call a stronger visual model and what that call may change. The scored object is not a native answer: a required claim can make Conflict correct. Recent work falls into three lines (Table 1).

Routing decides when to call, then adopts the answer. FrugalGPT, Hybrid LLM, RouteLLM, and cascade routing spend a larger model only when quality justifies the cost (Chen et al., 2024b; Ding et al., 2024; Ong et al., 2025; Dekoninck et al., 2025). A call helps when the two decisions differ in correctness, not merely when the first is unconfident (Jitkrittum et al., 2023; Narasimhan

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

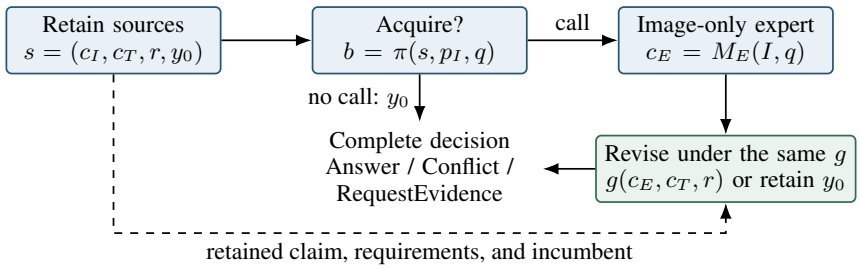

Figure 1: PAVIR retains the sources the task still needs and uses an expert to update the visual observation only. The acquisition policy can change while the revision rule stays fixed.

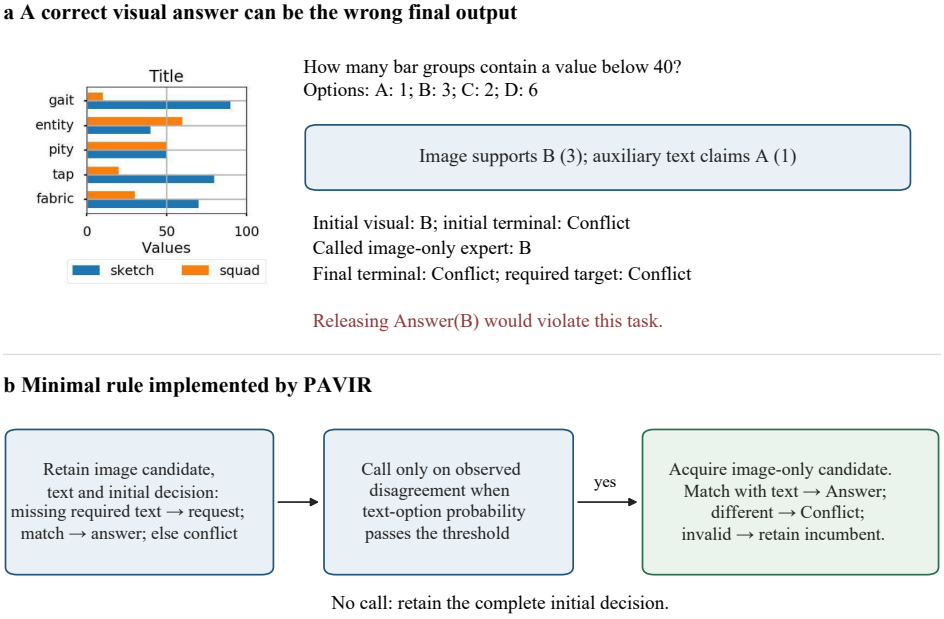

Figure 2: A correct visual answer can still be the wrong complete decision. (a) An MMStar conflict episode (unaltered cached image): the chart supports B, the required claim names A, and both visual models predict B. The overlay target is Conflict; releasing Answer(B) would solve a different task. (b) The implemented rule: retain sources, call an image-only expert only on observed disagreement, and compare the new observation with the retained claim. Match uses option identifiers, not string similarity.

et al., 2025; Cheng & Dong, 2026). VL-RouterBench and MMR-Bench score that choice on native answers (Huang et al., 2026; Ma et al., 2026). Despite the cost–quality gains, treating the expert's answer as the terminal fails when Conflict is legal.

Revision improves one string, not a source relation. CRITIC, Self-RAG, and AutoMix revise one output with feedback, critique, or routing (Gou et al., 2024; Asai et al., 2024; Aggarwal et al., 2024). Without an external signal, revision often replaces a correct answer with a wrong one (Huang et al., 2024). They do not keep a second source whose disagreement is itself the correct decision.

Conflict is removed or forced into an answer. IR-VQA, RGMR, and irrelevant-passage filters drop distracting context so the model can still answer (Yang et al., 2026; Yoran et al., 2024). CC-VQA analyzes a clash in order to emit one answer (Hong et al., 2026); HallusionBench and MC2 test whether misleading context pulls a model off a visual answer (Guan et al., 2024; Zhang et al., 2026). Despite measuring conflict, they do not score it.

Table 1: Evidence used and decision scored by representative routing, conflict-handling, and revision methods, compared with source-preserving revision.

Line of work	Evidence used	Scored decision
Routers, VL-RouterBench	query and model features	native answer vs. cost
IR-VQA, RGMR, CC-VQA	context to drop or to side with	an answer, not Conflict
CRITIC, Self-RAG, AutoMix	one string plus feedback	revised answer
PAVIR	both sources and requirements	Answer, Conflict, or RequestEvidence

None of the three keeps a required claim in the state and scores Conflict when that claim disagrees with the image.

3 SOURCE-PRESERVING REVISION

3.1 TASK AND RETAINED STATE

An input contains an image I, a question q, answer options $\mathcal{C}_q$, auxiliary text T, and source requirements r.

We first consider a mutually exclusive option domain: the visual observation and a valid textual claim each name one option, and two different options are incompatible answers to the same question. The complete output space is

$$\mathcal{Y}_q = \{\text{Answer}(c) : c \in \mathcal{C}_q\} \cup \{\text{Conflict}, \text{RequestEvidence}\}. \tag{1}$$

An answer is correct only when both its type and option match the task target. Conflict is a positive label for incompatible sources. RequestEvidence denotes a missing required source. Neither is an error by default, and neither is a generic refusal.

The initial image-only model produces a candidate c_I and option scores p_I. A parser extracts c_T from the auxiliary claim. For valid option observations,

$$g(c, c_T, r) = \begin{cases} \text{RequestEvidence}, & \text{a required source is absent}, \\ \text{Answer}(c), & \text{text is optional and absent}, \\ \text{Answer}(c), & \text{text is present and } c = c_T, \\ \text{Conflict}, & \text{text is present and } c \neq c_T. \end{cases} \tag{2}$$

The first case takes precedence. The last two cases require valid candidates in the declared mutually exclusive domain. Matching compares option identifiers in $\mathcal{C}_q$, not surface forms. A free-form claim needs a relation adapter before g applies; Section 4 evaluates that case separately on M2CQA.

The retained state is $s = (c_I, c_T, r, y_0)$ with $y_0 = g(c_I, c_T, r)$ for valid observations. Keeping c_T and r makes a new visual observation evaluable under the same task; y_0 supplies the no-call decision and invalid-expert fallback. The implementation also forms a RequestEvidence incumbent when the initial visual observation is unavailable or invalid. That separate branch is described in Appendix A; the simplified equations and main comparisons concern valid option observations.

3.2 A SHARED RULE WITH A REPLACEABLE CALLER

Let $b = \pi(s, p_I, q)$ be a pre-call acquisition decision. If $b = 1$, the expert receives the image, question, and options, without the auxiliary claim, and returns c_E. The final decision is

$$F_\pi(x) = \begin{cases} g(c_E, c_T, r), & b = 1 \text{ and } c_E \text{ is valid}, \\ y_0, & \text{otherwise}. \end{cases} \tag{3}$$

The acquired observation updates the visual component of the state. It does not erase the text or change which sources are required. The rule is deterministic given those observations. Observation models are not fine-tuned. The caller and the relation adapter may require fitting or calibration.

Disagreement-gated implementation. The default policy considers only complete, valid inputs on which the initial sources already disagree:

$$b = \mathbf{1}[\text{required sources present and valid},\ c_I \neq c_T]\,\mathbf{1}[p_I(c_T) \geq \tau]. \tag{4}$$

$p_I(c_T)$ is the initial visual score of the text-named option, used only to rank eligible disagreements. We choose τ on source-development data, then freeze it. On MMStar it keeps 399 of 494 disagreements. Missing required text is ineligible, and so is an observed agreement, including one in which both sources are wrong. Appendix A records the thresholds.

Why a better visual prediction can help or harm. On an eligible episode, $y_0 = \text{Conflict}$. Let A denote that the text-named option is the correct visual answer, and let z be the available pre-call information. A valid expert match $c_E = c_T$ changes Conflict to an answer. It repairs the decision when A holds and harms it otherwise.

For this Conflict incumbent, the expected change in complete-decision accuracy is

$$\Delta(z) = \Pr(c_E = c_T, A \mid z) - \Pr(c_E = c_T, A^c \mid z). \tag{5}$$

Invalid expert observations contribute zero because y_0 is retained. An expert mismatch leaves Conflict unchanged.

Equation 5 specializes the relative-value structure of cascade deferral (Jitkrittum et al., 2023) to source reconciliation. It identifies the quantity a useful acquisition score should predict: repairs minus harms, rather than expert agreement alone. We evaluate inexpensive proxies for this quantity.

4 Experiments

4.1 Experimental setup

Each comparison below changes one factor: who sees the text, whether the expert writes a full decision or only an image observation, or which score ranks a shared set of disagreements. The no-call policies already emit Answer, Conflict, or RequestEvidence, so a gain is not an artifact of a larger label set.

Hosts. ScienceQA, MMStar, and MMMU-Pro supply the images, questions, and options (Lu et al., 2022; Chen et al., 2024a; Yue et al., 2025). On each public item we attach an aligned claim, a conflicting claim, or no claim, as in MC2 (Zhang et al., 2026). Aligned claims name the reference option; conflicting claims name a wrong one. ScienceQA treats absent text as optional, so the target stays a visual answer. MMStar and MMMU-Pro require the claim and use RequestEvidence when it is missing. This is a reconciliation task. Native VQA routers and irrelevance filters score a different output (Huang et al., 2026; Yang et al., 2026).

ScienceQA has 169 semantic groups from 166 images and two endpoint pairs: PaliGemma2-3B $\rightarrow$ Qwen3-VL-8B and Qwen3-VL-8B $\rightarrow$ Qwen3-VL-32B (Steiner et al., 2024; Bai et al., 2025a). Each factorial cell contains $169 \times 3 \times 2 = 1{,}014$ observations. MMStar has 400 evaluation items from 397 images (1,200 episodes) and 400 image-disjoint development items. Its pair is Qwen3-VL-8B $\rightarrow$ Qwen3-VL-32B, with visual-option accuracies 65.50% and 75.00% on the 400 evaluation questions. Primary MMMU-Pro has 860 image groups and 2,580 episodes, with Qwen3-VL-8B $\rightarrow$ Qwen3.5-27B; a second pair uses Gemma-3-4B $\rightarrow$ Gemma-3-12B. Visual-option accuracy and reconciliation accuracy are different targets.

Comparisons and metrics. Fixed-observation replays change a caller or a post-call rule while reusing cached candidates. Caller comparisons share revision and budget, and the gated comparisons share eligibility as well. Task-adapted LatentRouter (Cheng & Dong, 2026) uses the released learner with either question-hash features or OpenCLIP image/text embeddings on both hosts; both include the same source-state features (Appendix A). Full-task experts receive the image, claim, initial candidate, and complete output space. A separate same-backbone comparison evaluates CC-VQA task adaptations and their component ablations. We report complete-decision accuracy, calls, repairs, harms, answer coverage, and error conditional on emitting an answer (Whitehead et al., 2022).

Table 2: ScienceQA factorial, 1,014 observations per row. Percentages; absent text is optional. Both final rules use the same two-call semantic consensus. Adoption cannot emit Conflict. This is a component study of visibility and the final rule, not the MMStar disagreement-gated cascade.

Observer input	Final rule	Overall	Aligned	Conflict	Absent
Mixed image–text	Consensus adoption	63.41	99.70	0.00	90.53
Image only	Consensus adoption	60.36	90.53	0.00	90.53
Mixed image–text	Source comparison	82.15	99.70	56.21	90.53
Image only	Source comparison	**91.81**	90.53	94.38	90.53

Table 3: Same-output-space controls. MMStar: 1,200 episodes and 399 calls; primary MMMU-Pro: 2,580 episodes and 1,232 calls. Rows marked no calls use zero acquisitions. Gated callers share eligibility and revision. LR denotes five-seed LatentRouter task adaptations; repair/harm counts are relative to initial.

	MMStar		MMMU-Pro primary	
Acquisition / update	Acc.	Repair/harm	Acc.	Repair/harm
Initial / recompute (no calls)	84.83	0/0	75.89	0/0
Calibrated text score (no calls)	84.67	21/23	77.25	97/62
Calibrated relative score (no calls)	84.58	29/32	77.71	190/143
Global uncertainty	87.92	65/28	77.60	166/122
Gate + uncertainty	88.42	59/16	79.69	145/47
Gate + LR (hash)	87.78	49/13.6	79.45	137.2/45.4
Gate + LR (OpenCLIP)	87.90	49.4/12.6	79.43	138.6/47.2
Gate + text-option score	**88.67**	61/15	**79.73**	148/49

Confidence intervals use paired group resampling, keeping associated source conditions and observations together. ScienceQA clusters 169 semantic groups by 166 original images; MMStar uses 397 decoded-image groups. The MMMU-Pro eight-comparison family uses Holm correction (Appendix D); shared-gate, full-task, and same-backbone method contrasts are exploratory. Aggregate scores use the constructed episode mixture and are reported with state-specific counts.

4.2 Source visibility changes the observation used for revision

With the rule fixed, isolating the observer raises accuracy from 82.15% to 91.81% ($+9.66$ $[7.35, 11.93]$; Table 2). The gain is in the conflict cells, 56.21% to 94.38%; aligned accuracy falls, and optional-absent accuracy is unchanged. If the rule cannot emit Conflict, the same isolation lowers accuracy by 3.05 points. The interaction is 12.72 points $[10.62, 14.87]$. On released MC2 items the same visibility effect appears without a cascade gain: image-only Qwen3-VL-4B is correct on 997/1,000 records versus 809/1,000 with mixed input ($+18.80$ $[15.60, 22.10]$), while a 40% cascade retains the initial decision on every call (Zhang et al., 2026).

4.3 A shared rule improves a full-output initial policy

We measure expert value against an initial policy that already applies Equation 2 and can emit every task terminal. Appendix C reports the same calls under answer-only adoption, which cannot emit Conflict.

Recomputing g with no new observation leaves every MMStar terminal unchanged at 1,018/1,200 (84.83%). Disagreement-gated revision reaches 1,064/1,200 (88.67%) in 399 calls: 61 repairs and 15 harms (Table 3). Of those calls, 323 stay Conflict and 76 become an answer. Updating to Answer only when the expert matches the claim reproduces every PAVIR terminal, because a called episode is already Conflict. Primary MMMU-Pro moves from 75.89% to 79.73% ($+3.84$ $[2.83, 4.88]$; 148 repairs, 49 harms), and the Gemma pair from 68.99% to 71.01%. Thresholds fit on development data that emit Answer or Conflict with no expert reach 84.67% and 84.58% on MMStar, below both gated callers.

Table 4: MMStar accuracy (%) at shared disagreement-pool quotas, over the same 1,200 episodes. All methods share observations and revision. LR columns average five fits; Uniform is the exact expectation under uniform fixed-size sampling. Quotas reuse the earlier budget grid, plus zero and the original 399-call point. These match expert acquisitions, not total pipeline compute.

Calls	Text score	Uncertainty	LR hash	LR visual	Uniform
0	84.83	84.83	84.83	84.83	84.83
49	85.75	85.33	85.10	85.18	85.21
124	87.50	86.17	85.68	85.72	85.77
247	88.33	87.83	86.65	86.55	86.71
370	88.50	88.42	87.57	87.57	87.64
399	**88.67**	88.42	87.78	87.90	87.86
494	88.58	88.58	88.58	88.58	88.58

Table 5: MMStar at the same 399 call locations, retaining the common initial output elsewhere. Procedures share the complete output space and use the same 32B expert snapshot. Bounded analysis allows 256 tokens before a constrained A–F decision; projection adds no model calls. Initial uses zero calls. State columns are correct counts out of 400.

Decision procedure	Acc.	Repair/harm	Aligned	Conflict	Missing
Initial (8B, no calls)	84.83	0/0	262	356	400
Full task (1 token)	83.58	34/49	296	307	400
Full task, strong (1 token)	84.92	15/14	277	342	400
Full task (bounded analysis)	87.25	65/36	327	320	400
+ consistency projection	88.50	65/21	327	335	400
Image-only + revision	**88.67**	61/15	323	341	400

4.4 Eligibility and ranking matter at different budgets

Under the shared gate, text-option ranking and uncertainty differ by three MMStar decisions (88.67% versus 88.42%) and one MMMU-Pro decision (79.73% versus 79.69%). Their call sets differ by 44 and 54 episodes, but almost all of those swaps leave the decision unchanged. Five-seed LatentRouter (Cheng & Dong, 2026), with the released learner and either a question hash or OpenCLIP features (Cherti et al., 2023), reaches 87.78% and 87.90% on MMStar once it may call only disagreements; the hash-to-visual change is $+0.12$ $[-0.20, 0.45]$. At 399 calls the text-option score is 0.81 points above uniform draws from the 494 eligible disagreements (87.86%). Calling all 494 yields 88.58%, one harm worse. From 49 to 247 calls the text-option score leads uncertainty by 0.42–1.33 points (Table 4). The gate decides which episodes can change. The score matters when the budget cannot cover that pool, and little once it can. Details are in Appendix F.

4.5 Full-task generation benefits from explicit source consistency

The 32B expert is the same snapshot used for image-only calls. One-token full-task prompts at the same 399 locations reach 83.58% and 84.92%. Allowing 256 analysis tokens before a legal label reaches 87.25% (65 repairs, 36 harms; Table 5). We then apply g with no new call: an Answer that contradicts the retained claim becomes Conflict. This changes 19 outputs, corrects 15, and reaches 88.50%. Image-only revision is two decisions higher, $+0.17$ $[-1.16, 1.50]$, with fewer harms (15 versus 21). The correct sets still differ, by 30 and 28 episodes. Unbounded decoding is not evaluated. Appendix A records the prompts.

4.6 Conflict analysis and decision consistency play different roles

We compare CC-VQA (Hong et al., 2026) with full-task generation and image-only revision using the same Qwen2.5-VL-7B-Instruct backbone (Bai et al., 2025b), the same 399 call locations, and the common initial decisions elsewhere. The single-claim adaptation retains released conflict-analysis prompts, correlation-guided encoding and decoding (CGED), and question rewriting. The given claim replaces retrieved passages; final decisions use the reconciliation output space. With two available reasons, we enable the conflict summary at two rather than the release's four-reason

Table 6: Same-backbone MMStar comparison: 1,200 episodes, 399 fixed acquisitions, Qwen2.5-VL-7B. Accuracy (%); repair/harm relative to the common 84.83% initial policy. The same zero-call consistency projection is applied to every full-output generative procedure. Image-only revision already enforces source comparison, so no additional projection is needed. Generated tokens are means per acquisition, counting each procedure's shared intermediates once.

	Procedure output		+ consistency		
Procedure	Acc.	Repair/harm	Acc.	Repair/harm	Gen. tokens
Full task (bounded)	79.42	46/111	84.92	46/45	99.8
CC-VQA, single claim	76.42	37/138	84.50	37/41	407.4
w/o CGED	77.92	39/122	84.83	39/39	407.4
original summary gate	74.83	35/155	85.00	35/33	266.8
Image-only + revision	**86.42**	46/27	—	—	1.0

Table 7: Where complete-decision gains arise. Entries are correct decisions before → after acquisition. Conditional answer error is computed only over emitted answers. Initial wrong answers persist because the gate does not call observed agreements.

Quantity	MMStar	MMMU-Pro primary
Aligned	$262 \rightarrow 323$ / 400	$314 \rightarrow 462$ / 860
Conflict	$356 \rightarrow 341$ / 400	$784 \rightarrow 735$ / 860
Missing required text	$400 \rightarrow 400$ / 400	$860 \rightarrow 860$ / 860
Answer coverage (%)	$25.50 \rightarrow 31.83$	$15.12 \rightarrow 22.75$
Conditional answer error (%)	$14.38 \rightarrow 15.45$	$19.49 \rightarrow 21.29$
Wrong answers (count)	$44 \rightarrow 59$	$76 \rightarrow 125$

threshold. We also evaluate the original, inactive summary gate and a variant without CGED (Appendix A).

Single-claim CC-VQA reaches 76.42%. Applying g with no new call raises it to 84.50%, cutting harms from 138 to 41 while leaving its 37 repairs unchanged. The same projection lifts ordinary 7B generation from 79.42% to 84.92%. After projection the CC-VQA variants span 84.50–85.00% (Table 6). Image-only revision reaches 86.42% (46 repairs, 27 harms): $+1.92$ $[0.42, 3.43]$ over projected single-claim CC-VQA, with one generated token against 407. Writing an analysis does not compare the sources. The rule does, and it is cheaper.

4.7 Transfer results and task boundaries

On native MMMU-Pro, absent text is optional, so g returns the visual answer and revision equals adoption. The primary controlled contrasts with confidence and restricted-feature LatentRouter survive Holm correction; the Gemma contrasts do not (Appendix D). On M2CQA the claims are free-form, so option identity does not apply. At 480 calls, source comparison moves task-fitted callers from 77.92% and 78.08% only to 78.08% and 78.17% (Appendix B). The adapter and the caller set the accuracy; g does not transfer unchanged.

4.8 Repairs trade against harms; joint errors are untouched

The calls repair aligned disagreements and harm some conflicts (Table 7). They never touch $c_I = c_T$, so 44 MMStar and 76 MMMU-Pro joint errors remain; final wrong answers are those misses plus the harms. Source-complete accuracy is 83.00% versus 82.75% for the projected 32B expert. Appendix E reweights the mixture; Appendix G perturbs the rule.

5 Discussion and Conclusion

We asked whether a stronger visual model should replace the cascade's decision. When the task still depends on a textual claim, it should not. Keeping the claim, reading the image without it, and comparing the two sources explains the gains we measure. On ScienceQA, hiding the text raises

accuracy by 9.66 points if a conflict may be reported, and lowers it if the model must answer. On MMStar, enforcing the comparison on a 256-token 32B answer reaches 88.50%, against 88.67% when the expert never sees the claim. The interval includes zero, and the image-only procedure introduces fewer wrong revisions (15 versus 21).

Which score selects the call is secondary. At the operating point chosen on development data, ranking disagreements by uncertainty is nearly tied with the text-option score, and calling every disagreement is slightly worse than stopping at 399 calls. The comparison assumes mutually exclusive option labels. It misses agreements that are both wrong, and it does not cover ordinary visual QA, irrelevant-context filtering, or free-form claims. When the task must reconcile an image with a claim, keep both and report repairs with harms.

AI USE STATEMENT

Generative AI was used to polish the manuscript text.

REPRODUCIBILITY STATEMENT

Section 3 states the decision rule. Section 4 and Appendix A state the data, calibration, models, and budgets; Appendix D states the multiplicity correction. Replayed observations and new generations are marked separately.

REFERENCES

Pranjal Aggarwal, Aman Madaan, Ankit Anand, Srividya Pranavi Potharaju, Swaroop Mishra, Pei Zhou, Aditya Gupta, Dheeraj Rajagopal, Karthik Kappaganthu, Yiming Yang, Shyam Upadhyay, Manaal Faruqui, and Mausam. AutoMix: Automatically mixing language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 131000–131034, 2024. doi: 10.52202/079017-4164.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/25f7be9694d7b32d5cc670927b8091e1-Paper-Conference.pdf.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025a. doi: 10.48550/arXiv.2511.21631. Preprint.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025b. URL https://arxiv.org/abs/2502.13923. Preprint.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision–language models? In *Advances in Neural Information Processing Systems*, volume 37, pp. 27056–27087, 2024a. doi: 10.52202/079017-0850.

Lingjiao Chen, Matei Zaharia, and James Zou. FrugalGPT: How to use large language models while reducing cost and improving performance. *Transactions on Machine Learning Research*, 2024b. URL https://openreview.net/forum?id=cSimKw5p6R.

Xueqi Cheng and Yushun Dong. LatentRouter: Can we choose the right multimodal model before seeing its answer? *arXiv preprint arXiv:2605.11301*, 2026. doi: 10.48550/arXiv.2605.11301. Preprint.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language–image learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2818–2829, 2023.

Jasper Dekoninck, Maximilian Baader, and Martin Vechev. A unified approach to routing and cascading for LLMs. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 12987–13010, 2025. URL https://proceedings.mlr.press/v267/dekoninck25a.html.

Dujian Ding, Ankur Mallick, Chi Wang, Robert Sim, Subhabrata Mukherjee, Victor Ruhle, Laks V. S. Lakshmanan, and Ahmed Hassan Awadallah. Hybrid LLM: Cost-efficient and quality-aware query routing. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/b47d93c99fa22ac0b377578af0a1f63a-Abstract-Conference.html.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. CRITIC: Large language models can self-correct with tool-interactive critiquing. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/fef126561bbf9d4467dbb8d27334b8fe-Paper-Conference.pdf.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision–language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14375–14385, 2024.

Yuyang Hong, Jiaqi Gu, Yujing Lou, Lubin Fan, Qi Yang, Ying Wang, Kun Ding, Yue Wu, Shiming Xiang, and Jieping Ye. CC-VQA: Conflict- and correlation-aware method for mitigating knowledge conflict in knowledge-based visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5232–5241, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Hong_CC-VQA_Conflict-_and_Correlation-Aware_Method_for_Mitigating_Knowledge_Conflict_in_CVPR_2026_paper.html.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/8b4add8b0aa8749d80a34ca5d941c355-Paper-Conference.pdf.

Zhehao Huang, Baijiong Lin, Jingyuan Zhang, Jingying Wang, Yuhang Liu, Ning Lu, Tao Li, and Xiaolin Huang. VL-RouterBench: A benchmark for vision–language model routing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://arxiv.org/abs/2512.23562.

Wittawat Jitkrittum, Neha Gupta, Aditya Krishna Menon, Harikrishna Narasimhan, Ankit Singh Rawat, and Sanjiv Kumar. When does confidence-based cascade deferral suffice? In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1f09e1ee5035a4c3fe38a5681cae5815-Abstract-Conference.html.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/11332b6b6cf4485b84afadb1352d3a9a-Abstract-Conference.html.

Haoxuan Ma, Guannan Lai, and Han-Jia Ye. MMR-Bench: A comprehensive benchmark for multimodal LLM routing. *arXiv preprint arXiv:2601.17814*, 2026. URL https://arxiv.org/abs/2601.17814. Preprint.

Atsuyuki Miyai, Jingkang Yang, Jingyang Zhang, Yifei Ming, Qing Yu, Go Irie, Yixuan Li, Hai Helen Li, Ziwei Liu, and Kiyoharu Aizawa. Unsolvable problem detection: Robust understanding evaluation for large multimodal models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6497–6540, 2025. doi: 10.18653/v1/2025.acl-long.326.

Basel Mousi, Fahim Dalvi, Shammur Absar Chowdhury, Firoj Alam, and Nadir Durrani. Once correct, still wrong: Counterfactual hallucination in multilingual vision–language models. In

Findings of the Association for Computational Linguistics: ACL 2026, pp. 4763–4788, 2026. doi: 10.18653/v1/2026.findings-acl.234.

Harikrishna Narasimhan, Wittawat Jitkrittum, Ankit Singh Rawat, Seungyeon Kim, Neha Gupta, Aditya Krishna Menon, and Sanjiv Kumar. Faster cascades via speculative decoding. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=vo9t20wsmd.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M. Waleed Kadous, and Ion Stoica. RouteLLM: Learning to route LLMs with preference data. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/5503a7c69d48a2f86fc00b3dc09de686-Abstract-Conference.html.

Teodora Popordanoska, Jiameng Li, and Matthew B. Blaschko. CLASH: A benchmark for cross-modal contradiction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Findings*, pp. 6051–6061, June 2026.

Qwen Team and MLX Community. Qwen3.5-27b: Mlx 4-bit model card. Hugging Face, 2026. URL https://huggingface.co/mlx-community/Qwen3.5-27B-4bit. Pinned revision 45797d2985a12c55e6473686e9ea91b95e959553.

Andreas Steiner, André Susano Pinto, Michael Tschannen, et al. PaliGemma 2: A family of versatile VLMs for transfer. *arXiv preprint arXiv:2412.03555*, 2024. doi: 10.48550/arXiv.2412.03555. Preprint.

Spencer Whitehead, Suzanne Petryk, Vedaad Shakib, Joseph Gonzalez, Trevor Darrell, Anna Rohrbach, and Marcus Rohrbach. Reliable visual question answering: Abstain rather than answer incorrectly. In *European Conference on Computer Vision*, 2022. URL https://www.ecva.net/papers/eccv_2022/papers_ECCV/papers/136960146.pdf.

Jinhui Yang, Ming Jiang, and Qi Zhao. Defying distractions in multimodal tasks: A novel benchmark for large vision-language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(6):6314–6331, 2026. doi: 10.1109/TPAMI.2026.3655641.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. Making retrieval-augmented language models robust to irrelevant context. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/8011b23e1dc3f57e1b6211ccad498919-Paper-Conference.pdf.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. MMMU-Pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15134–15186. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.736.

Yu Zhang, Jinlong Ma, Yongshuai Hou, Xuefeng Bai, Kehai Chen, Yang Xiang, Jun Yu, and Min Zhang. Evaluating and steering modality preferences in multi-modal LLMs. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/61222.

A Implementation and Evaluation Details

Observation models. MMStar uses MLX 4-bit Qwen3-VL-8B/32B Instruct snapshots. The primary MMMU-Pro expert is MLX 4-bit Qwen3.5-27B with thinking disabled (Qwen Team & MLX Community, 2026). Option observations use greedy one-token prediction constrained to legal option tokens. MC2's image-only comparison uses Qwen3-VL-4B. ScienceQA forms a two-call semantic consensus before either final rule; unresolved consensus yields RequestEvidence under both rules. It is a component study, distinct from the disagreement-gated caller on MMStar and MMMU-Pro.

Splits, grouping, and thresholds. MMStar development contains 400 items and 393 decoded images, disjoint from the 397 evaluation images. The restricted-feature router uses 315 training and 78 validation images. MMMU-Pro has 860 evaluation image groups; option counts range from two to ten. ScienceQA has 169 semantic groups from 166 original images; its intervals use 10,000 paired bootstrap replicates clustered by those images, seed 20260908. Source-development selection maximizes complete-decision accuracy, retains a no-call candidate, and breaks ties toward fewer calls. Recorded score thresholds are 3.06×10^{-7} on MMStar (399 of 494 eligible disagreements) and 5.04×10^{-7} on primary MMMU-Pro. The 494-call configuration is used only for released-router comparison, latency, and interface perturbations.

Calibrated zero-call controls. Source-only decision thresholds 0.1186519731 for $p_I(c_T)$ and 0.0820849986 for $p_I(c_T)/\max p_I$ are fitted on 800 source-complete development episodes from 393 images, maximizing decision accuracy with ties preferring fewer answers, and reused on MMStar and primary MMMU-Pro.

LatentRouter. The option-host adaptation uses a 192-dimensional question hash and 17 numerical/source-state features, without direct image embeddings. Seeds are 20260308–20260312; hidden size 48, two layers, four heads, seven capsules, four query tokens. No target example is used for fitting. Shared-gate comparisons retain all five fitted score vectors, drop ineligible episodes, and rank the rest under one quota. Ties use the same public-input-content key as uncertainty. The five restricted-feature MMStar fits allocate 118, 14, 104, 28, and 6 calls to missing-source episodes when eligibility is not shared.

The vision-enabled versions replace the hash with 512-dimensional image and 512-dimensional text features from OpenCLIP ViT-B-32 (Cherti et al., 2023), checkpoint `laion2b_s34b_b79k`, keeping the 17 scalars, 315/78 split, seeds, and learner. Multi-image sets use the released encoder's normalized pooling. Target correctness fields are zeroed before utility prediction; expert outputs are excluded from query features.

Full-task generation. One-token MMStar prompts map A–D to answers, E to Conflict, and F to RequestEvidence with greedy decoding. They exclude reference labels and cached expert answers. Applied to all 1,200 episodes they reach 74.50% and 82.75% (aligned/conflict/missing 246/349/398 for the stronger prefix). Table 5 instead uses only the 399 cascade locations.

The bounded variant allows at most 256 analysis tokens, then a constrained A–F decision; temperature is zero. Format trials on 12 development episodes compared 128/256/512-token caps without scoring accuracy. These 399 calls produce 59,454 tokens in 798 passes; 33 analyses reach the cap. Mean/P95 generation latency is 16.65/24.86 seconds on an Apple M3 Max. The same projection applied to the two one-token prompts yields 85.67% (34/24) and 85.92% (15/2).

CC-VQA adaptation. The comparison in Section 4.6 uses Qwen2.5-VL-7B-Instruct in BF16 with PyTorch 2.13.0 and Transformers 4.57.6. It retains CC-VQA release `28c4da8` prompts, similarity-modified rotary positions, and dual-stream fusion over the full vocabulary before restricting to A–F. The given claim replaces retrieved passages. The single-claim arm enables the conflict summary at two reasons rather than four; the original-gate control keeps rewriting, CGED, and an inactive summary; the no-CGED arm uses single-stream final decoding. EVA-CLIP-8B text projection is used in FP16 with 77-token truncation. Generation is greedy: 512 tokens for analysis/reasons/summaries and 128 for rewriting. Resource use is in Table 8.

Enabling the two-reason summary changes raw accuracy by $+1.58$ points $[-0.25, 3.42]$ relative to the original gate. Adding CGED to the shared summary/rewrite procedure changes it by -1.50

Table 8: Resource use per expert acquisition for Table 6. Generative stages are autoregressive analysis/reason/summary/rewrite calls; final passes count decision-logit forward passes. EVA counts encoded question/sentence sequences. Warm time excludes loading, record I/O, and orchestration.

Procedure	Gen. stages	Final passes	EVA sequences	Warm mean/P95 (s)
Full task (bounded)	1	1	0.00	9.37/20.16
CC-VQA, single claim	4	2	11.37	34.22/67.41
w/o CGED	4	1	0.00	33.13/66.40
original summary gate	3	2	3.00	23.88/50.13
Image-only + revision	0	1	0.00	0.89/2.32

$[-3.10, 0.08]$. Relative to ordinary full-task generation, single-claim CC-VQA differs by -3.00 $[-5.12, -0.91]$ before projection and -0.42 $[-2.01, 1.17]$ after it.

Released-suite units and M2CQA caller. MC2 uses 100 fitting and 500 test COCO groups (both A/B orders). M2CQA's held-out confirmation uses 400 groups; a separate large arm uses 1,000 groups and 3,000 episodes. CLASH/MM-UPD contributes 2,626 groups, 11,681 instances, and 12,869 rows. M2CQA pre-call features are question TF–IDF, small-model option scores, confidence, margin, entropy, candidate identity, validity indicators, and length/polarity. Two logistic heads estimate rescue and harm on 500 development groups with five-fold image-group out-of-fold estimation.

Latency. Offline quality comparisons reuse cached observations. Timing on an Apple M3 Max at the 494-call MMStar point is mean/P95 1479/3325 ms versus 2761/6225 ms for source-resolved all-expert (46.43% mean reduction). That operating point is not the 399-call versus 32B quality comparison.

B Released-Suite Comparisons

M2CQA tests counterfactual hallucination, CLASH tests cross-modal contradiction, and MM-UPD tests unsolvable problems (Mousi et al., 2026; Popordanoska et al., 2026; Miyai et al., 2025). We do not pool them.

Table 9: M2CQA caller-by-update comparison, accuracy (%). Every cell uses 480 acquisitions over the same 1,200 episodes. Columns share calls within a row.

Caller	Adoption	Source comparison	+ risk filter
Margin	72.58	73.58	73.67
HybridLLM-style	77.92	78.08	78.17
RouteLLM-style	78.08	78.17	78.17

On margin calls, source comparison raises the correct count from 871 to 883 ($+1.00$ point); the risk filter adds one further correct decision (884, 73.67%). Differences from the two learned adoption callers are -4.25 points $[-5.75, -2.83]$ and -4.42 $[-5.83, -3.08]$. Applying source comparison to those learned call sets adds two and one correct decisions.

Table 10: Fixed-observation revision minus direct adoption on released suites. These contrasts include terminal-space effects. Conflict-only evaluation can favor a constant Conflict output.

Setting	Fixed acquisitions	Revision − adoption (pp)
M2CQA held-out, 400 groups	480 calls	$+1.00$ $[0.42, 1.67]$
CLASH (contradiction detection)	Same component calls	$+35.99$
MM-UPD (unsolvable problems)	Same component calls	$+0.05$
CLASH/MM-UPD, LatentRouter calls	Fixed LR calls	$+8.29$ $[7.35, 9.24]$

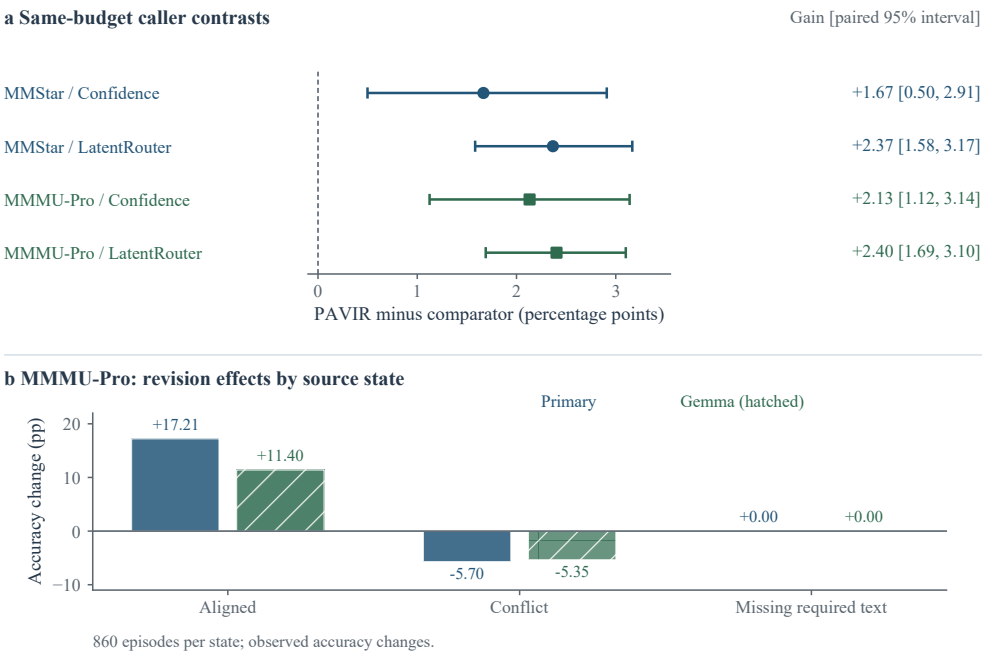

Figure 3: Matched-budget gaps and MMMU-Pro state-wise revision effects on cached candidates. (a) PAVIR minus each comparator under identical observations and revision; MMStar 399/1,200, primary MMMU-Pro 1,232/2,580. These include terminal-space effects and are not Table 3. (b) Accuracy change versus the initial policy on MMMU-Pro, 860 public items per source-visibility condition.

Most of the previously pooled $+16.21$ on CLASH/MM-UPD is CLASH. A constant Conflict rule can succeed on conflict-only data; the mixed-condition comparisons in the main text test overall decision quality. A separate 1,000-group M2CQA arm (600 calls) scores 74.90% versus 72.50% for a task-adapted LatentRouter and 61.53% for always-small. In that arm, adoption and source-aware revision coincide, so the caller gain is not evidence for the update.

C SHARED-RULE REPLAY ACROSS CALLERS

Table 11: MMStar post-call replay at 399 calls per caller over 1,200 episodes. Each row keeps that caller's assignments and candidates. Differences from direct adoption include a change in representable terminals. The main-text test of expert value is Table 3.

Acquisition policy	Direct adoption	Shared revision	Difference (pp)
Random	64.50	85.83	+21.33
Partitioned confidence	66.33	87.00	+20.67
LatentRouter adaptation	68.20 ± 2.07	86.30 ± 0.83	+18.10
Disagreement gate + text-option score	67.67	88.67	+21.00

Shared revision adds 18.10 points $[16.98, 19.22]$ for LatentRouter and 21.00 $[19.40, 22.56]$ for the disagreement-gated policy. The caller-by-revision interaction is 2.90 points $[1.29, 4.47]$.

D COMPLETE MMMU-PRO COMPARISON FAMILY

Budgets are 1,232 and 894 calls over 2,580 episodes (primary and Gemma, controlled overlay) and 832 and 50 calls over the same 860 public questions (native visual QA). Confidence denotes the

comparator in the original eight-contrast family; gated-uncertainty comparisons are not added to that family.

Table 12: PAVIR minus each comparator. Intervals are unadjusted paired image-group 95% intervals; p-values are Holm-adjusted over these eight contrasts. LR is the restricted-feature LatentRouter adaptation.

Pair	Target	Comparator	Difference (pp)	95% CI	Holm p
Primary	Controlled	Confidence	+2.13	[1.12,3.14]	.000160
Primary	Controlled	LR	+2.40	[1.69,3.10]	.000160
Primary	Native	Confidence	+0.00	[0.00,0.00]	1.000
Primary	Native	LR	+0.16	[-0.09,0.42]	1.000
Gemma	Controlled	Confidence	+0.58	[-0.35,1.51]	1.000
Gemma	Controlled	LR	+1.01	[0.14,1.87]	.154
Gemma	Native	Confidence	+0.00	[0.00,0.00]	1.000
Gemma	Native	LR	-0.16	[-1.05,0.70]	1.000

Only the two primary controlled comparisons survive correction. Native-task rows change the evaluation target: a controller that can emit Conflict does not automatically improve visual-option prediction.

E DECISION VALUE AND MIXTURE SENSITIVITY

Equation 5 enumerates transitions from an eligible Conflict incumbent. A valid expert match repairs the decision when the text-named option is the correct visual answer and harms it otherwise. Mismatches and invalid observations leave Conflict unchanged. The identity assumes the mutually exclusive option domain; it is not a derivation for free-form entailment. The main-text reweightings hold recorded state-conditional accuracies fixed and are not a fitted policy.

F BUDGET SENSITIVITY

Table 13 is the source-development threshold sweep on cached MMStar observations. No target accuracy is used to pick a new threshold. Duplicate zero-call settings are shown once. The selected threshold is 3.06×10^{-7} (399 calls); 399–444 calls all score 88.67%, and calling all 494 eligible disagreements scores 88.58%.

Table 13: Source-development-threshold sweep on MMStar, 1,200 episodes, same observations.

Source threshold	Calls	Acc.	Repair/harm
2 (no call)	0	84.83	0/0
8.69×10^{-2}	54	85.75	16/5
1.75×10^{-2}	98	86.75	30/7
3.53×10^{-3}	132	87.67	41/7
4.96×10^{-4}	190	88.00	49/11
7.48×10^{-5}	239	88.42	56/13
1.10×10^{-5}	290	88.42	57/14
1.55×10^{-6}	355	88.50	59/15
3.06×10^{-7}	399	**88.67**	61/15
2.38×10^{-7}	404	88.67	61/15
2.77×10^{-8}	444	88.67	61/15
1.39×10^{-11}	494	88.58	61/16

The shared-pool quota grid in Table 4 remains the ranking diagnostic; this table answers whether the development-chosen cutoff sits on a plateau.

G INTERFACE PERTURBATIONS

The study fixes the 494 MMStar calls and all model candidates. Stochastic conditions perturb an exact fraction of called episodes across 50 deterministic placements. Baseline accuracy at this operating point is 88.58%.

Table 14: Complete-decision accuracy (%) under fixed-candidate interface perturbations. Parsing failures occur after a usable initial state exists.

Perturbation	5%	10%	20%
Relation flip	87.03	85.46	82.21
Parsing failure, retain initial	88.38	88.24	87.89
Parsing failure, request evidence	86.90	85.27	81.92

A systematic Cross↔Visual source-requirement swap is not the same as random parse noise. When detected and handled by retaining the initial decision, accuracy is 84.83%; without detection it becomes direct adoption at 60.25%. Detection is an assumption of that fallback, not a demonstrated universal capability.