VERIFICATION MIRAGE: MAPPING AND REPAIRING THE RELIABILITY BOUNDARY OF SELF-VERIFICATION IN MEDICAL VQA

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-verification is increasingly used to improve the reliability of vision-language models (VLMs) and provide feedback for iterative self-correction, yet it is useful only when checking an answer provides information that discriminates correct from incorrect predictions rather than simply re-solving the same visual problem. We introduce `VeriMap`, an auditing framework for mapping this reliability boundary in medical visual question answering, where many answers hinge on fine-grained but auditable visual distinctions. Across six VLMs, five datasets, and seven task types, `VeriMap` reveals a *verification mirage*: vanilla verifiers accept 71.0% of incorrect candidates on average, and all six models are more likely to accept their own errors than errors from other models. When reused as feedback, these false acceptances can persist across self-correction rounds. We find that stronger access to the correct image or candidate-hidden re-solving is insufficient. Instead, verification becomes more selective when the candidate is reduced to a falsifiable visual distinction and the resulting observation governs the verdict. Guided by this finding, we introduce *`MirageCheck`*, which separates test construction, candidate-hidden visual measurement, and evidence-to-verdict binding through CONTRAST, MEASURE, and BIND. Across models and datasets, `MirageCheck` reduces mean false acceptance by 14.6 percentage relative to the strongest candidate-independent baseline while improving balanced accuracy, and substantially reduces false acceptance during iterative correction. Together, our results show that reliable visual self-verification requires turning acceptance to discriminative evidence from the current image rather than another judgment of the same problem.

1 INTRODUCTION

Self-verification is often motivated by a familiar asymmetry: *a proposed answer may be easier to check than to produce from scratch* (Cobbe et al., 2021; Kadavath et al., 2022; Lightman et al., 2024). For an equation, deriving a solution may require several steps, while checking a proposed solution can require only substituting it back into the equation. Visual questions, however, do not automatically inherit this advantage. A candidate answer provides a hypothesis, but not necessarily the visual evidence needed to test it. When verification does not introduce a more discriminative check, the verifier may still need to recover the same finding, anatomy, count, or spatial relation from the image that was required during generation. Self-verification can therefore collapse into another judgment of nearly the same problem, now conditioned on a plausible answer. This limitation matters beyond a single verdict. In actor-verifier (Gou et al., 2024), self-correction (Madaan et al., 2023), and self-improving systems (Song et al., 2025), verification is increasingly reused as feedback to revise an answer or guide subsequent behavior. If the verifier reproduces the actor's visual error, a local mistake can instead become a persistent positive signal.

This concern is especially sharp in medical VQA, where correctness often hinges on fine-grained distinctions in findings, anatomy, laterality, counts, or spatial relations (Meddeb et al., 2025). In our analysis in Appendix B.2, false acceptance is consistently higher on sampled medical benchmarks than on general-domain benchmarks, by 22-34 percentage points across five VLMs. Many medical questions expose concrete, auditable answer-critical visual evidence. This combination motivates our

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.

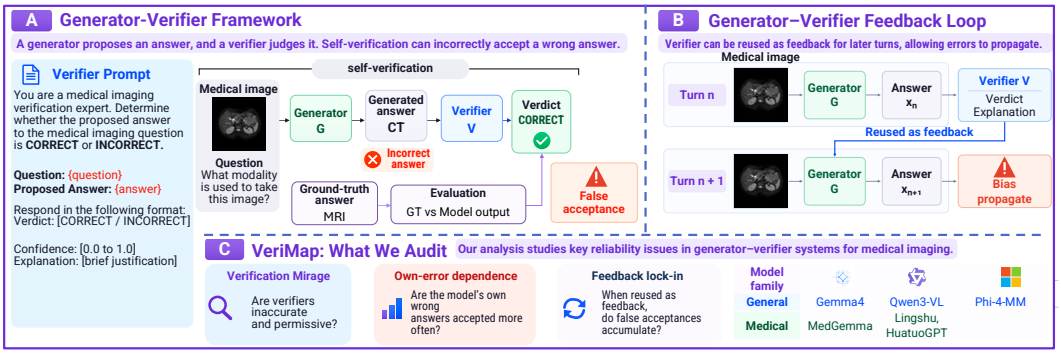


Figure 1: **Generator-verifier self-verification, feedback reuse, and the `VeriMap` audit. (A)** A generator answers a medical image question, and a verifier judges the generated answer using the same visual input and the verification prompt shown at left. Verifier decisions are evaluated against the ground-truth answer. **(B)** When the verifier's verdict and explanation are reused as feedback for a later turn, a false acceptance can influence subsequent generation and allow the original error to propagate. **(C)** `VeriMap` audits this reliability pathway at three levels: the overall *verification mirage*, preferential acceptance of a model's own errors, and feedback lock-in across turns, evaluated across general and medical VLMs from multiple model families.

central question: *when does medical visual self-verification provide a genuinely discriminative check rather than reproduce the original error?*

Prior work suggests two natural explanations for failures of visual verification. *First*, conditioning on a proposed answer can induce self-preference and agreement bias, motivating *candidate-independent verification strategies* (Xu et al., 2024; Panickssery et al., 2024; Wataoka et al., 2024; Andrade et al., 2025). *Second*, multimodal models can produce judgments that are insufficiently grounded in the current image, motivating methods that *stronger visual grounding* (e.g., hallucination) (Lee et al., 2024; Liao et al., 2025a). These two explanations lead to a natural question: *are reducing candidate dependence and strengthening image dependence sufficient to make verification genuinely discriminative between correct and incorrect answers?*

We study this question with **`VeriMap`**, an auditing framework for mapping the reliability boundary of visual self-verification (Fig. 1). Across six VLMs, five medical VQA datasets, and seven task types, `VeriMap` reveals a **verification mirage**: verifiers frequently accept incorrect candidates, and their failures are disproportionately concentrated on cases where generation has already gone wrong. Across the unified suite, verifiers accept 71.0% of incorrect candidates on average. More importantly, this failure is not explained by general verifier permissiveness alone. In a complete cross-generator audit over wrong answers, all six verifiers preferentially accept errors produced by themselves even after controlling for verifier- and generator-specific effects. This matters when verifier outputs are reused as feedback, where the bias can propagate. In a five-turn actor-verifier self-improvement, false acceptance accumulates among initially wrong trajectories. We then test two natural explanations for the mirage (candidate independence and stronger visual grounding). Neither reliably separates correct from incorrect answers.

These results point to a more specific failure: the verifier may use the image and reason without seeing the candidate, yet still reproduce the same error. The missing part is answer-discriminative visual evidence. Instead, verification becomes more selective when it tests a concrete visual distinction between the candidate and an alternative, then bases the verdict on what the image shows. In our frozen probe, this catches false acceptances more often than it overturns correct ones, motivating `MirageCheck`.

Having localized this failure, we intervene directly on the missing visual check with **`MirageCheck`**, a same-model protocol that replaces holistic re-judgment with an evidence-bound falsification test. `MirageCheck` decomposes verification into three stages: CONTRAST identifies the answer-critical visual distinction that separates the candidate from plausible alternatives. MEASURE obtains the corresponding observation without revealing which hypothesis is the original answer, and BIND reconstructs the supported hypothesis from randomized alternatives using only the sealed evidence. The original acceptance is retained only when this reconstruction succeeds. `MirageCheck` uses the

same VLM checkpoint throughout and requires neither reference answers nor an external critic at inference time.

Our work makes three contributions. **(i) We introduce `VeriMap`, a diagnostic framework for mapping the reliability boundary of medical visual self-verification.** Beyond aggregate verifier accuracy, `VeriMap` separates false acceptance from discrimination, identifies where verifier failures concentrate relative to generation errors, delineates the boundary of this dependence, and exposes its consequence when verification is reused as feedback. **(ii) We localize a missing discriminative visual check.** Through controlled interventions and a falsification probe, we show that visual access and candidate independence alone are insufficient, and provide evidence that verification becomes more selective when the verdict is tied to answer-discriminative visual evidence. **(iii) We translate this diagnosis into `MirageCheck`, a targeted intervention for evidence-bound visual falsification.** `MirageCheck` replaces holistic re-judgment with CONTRAST, MEASURE, and BIND, improves verification across the five-dataset evaluation suite, and is further tested when its verdicts are reused for iterative self-correction.

2 RELATED WORK

Self-verification, verifier reliability, and self-improvement. Self-verification supports selective generation, self-correction, and verifier-driven self-improvement across reasoning, code, and multimodal tasks (Kadavath et al., 2022; Ren et al., 2023; Ni et al., 2023; Chen et al., 2023; Madaan et al., 2023; Song et al., 2025). Its usefulness, however, depends on the verifier providing information beyond the generator's original failure. Prior work documents self-preference in LLM judges (Xu et al., 2024; Panickssery et al., 2024; Wataoka et al., 2024), bias in multimodal verification (Andrade et al., 2025), and limits of same-model verification for self-improvement (Song et al., 2025). We study a complementary question in visual verification. Rather than asking only whether verification is accurate or improves a solver, `VeriMap` tests whether errors are preferentially accepted when produced by the verifier itself, what makes verification discriminative, and how false acceptance propagates when verdicts are reused as feedback.

Multimodal verification and visual grounding. Multimodal self-correction and hallucination mitigation often improve reliability by eliciting candidate-independent judgments or strengthening visual grounding (Lee et al., 2024; Andrade et al., 2025; Sun et al., 2025; Wu et al., 2025; Ji et al., 2026; Yin et al., 2025). These methods motivate two natural explanations for visual verification failure: candidate independence and stronger visual grounding. Our analysis separates these from a stricter requirement. A verifier can depend on the correct image and reason independently of the candidate while still reproducing the same error. `VeriMap` localizes the missing signal to answer-discriminative visual evidence, and `MirageCheck` operationalizes this diagnosis by separating contrast construction, candidate-hidden measurement, and evidence-to-verdict binding. We use medical VQA as an evidence-explicit testbed for this reliability boundary as detailed in Appendix B.2.

3 VERIMAP: DIAGNOSING FAILURES IN SELF-VERIFICATION

`VeriMap` maps the reliability boundary of self-verification by asking whether verification discriminates correct from incorrect candidates, where its errors concentrate, and what happens when its verdicts are reused as feedback. We then test whether two prominent remedies, *candidate independence* and *stronger visual grounding*, are sufficient to resolve the failure and, if not, localize what additional signal is needed for a discriminative verdict.

3.1 EVALUATION SETUP

For each image-question pair (I_i, q_i) with reference answer a_i^*, a VLM first generates a candidate answer $\hat{a}_i = G(I_i, q_i)$. Comparing $\hat{a}_i$ with a_i^* determines the candidate-correctness label $y_i \in \{0, 1\}$, where $y_i = 1$ indicates a correct candidate and $y_i = 0$ an incorrect one. We then freeze $\hat{a}_i$ and ask a verifier (the model itself), in a fresh context, to judge whether the candidate is correct from $(I_i, q_i, \hat{a}_i)$. Its decision is denoted by $\hat{y}_i \in \{0, 1\}$, where $\hat{y}_i = 1$ denotes acceptance. The reference answer is used only to determine y_i and is never provided to the verifier. Verification is therefore evaluated

as binary discrimination between correct and incorrect candidates. We report Accuracy, Balanced Accuracy (BAcc.), and $\text{FPR} = \Pr(\hat{y} = 1 \mid y = 0)$. Directionality is summarized by

$$\text{Bias} = \frac{\text{FP} - \text{FN}}{N}, \qquad \text{D-Skew} = \frac{(\text{FP} - \text{FN})^2}{\text{FP}^2 + \text{FN}^2 + (\text{TP} + \text{TN})(\text{FP} + \text{FN})}.$$

The `VeriMap` covers six open-weight VLMs: Qwen3-VL-8B (Bai et al., 2025), Gemma-4-E4B (Team et al., 2026), Phi-4-Multimodal (Abouelenin et al., 2025), MedGemma-1.5-4B (Sellergren et al., 2025), Lingshu-7B (Xu et al., 2026), and HuatuoGPT-Vision-7B (Chen et al., 2024). We evaluate them on five datasets: VQA-RAD (Lau et al., 2018), SLAKE (Liu et al., 2021), PathVQA (He et al., 2020), MedXpert-QA (Zuo et al., 2025), and PMC-VQA (Zhang et al., 2023b). The details are provided in Appendix B.

3.2 The Verification Mirage

Self-verification frequently accepts incorrect answers. Fig. 2 shows that verification is both inaccurate and permissive across 6 models and 5 datasets (30 settings). The star in panel (a) marks the macro average over these settings: verifiers are wrong about half the time (48.5% error), and even among incorrect candidates they return Correct 71.0% of the time. Across settings, verifier error tends to be positively correlated with acceptance-side failure in Bias and D-Skew. To quantify these trends while accounting for systematic differences across models and datasets, we fit separate linear mixed-effects models with verifier error as the fixed-effect predictor and random intercepts for model and dataset. The fitted regression coefficients of the fixed effect in panels (a)-(c) show that each 10 percentage-point increase in verifier error is associated with increases of 3.3, 12.3, and 12.9 points in FPR, Bias, and D-Skew, respectively (the FPR slope is has $p = .116$, while both asymmetry slopes have $p < .001$ with details in Appendix C.2). Panel (d) makes the directional asymmetry explicit: FPR exceeds FNR in 27 of 30 settings. Verification is therefore not only inaccurate, but its errors are also usually skewed toward accepting incorrect candidates. We call this combination of substantial verifier error and directional false acceptance the *Verification Mirage*.

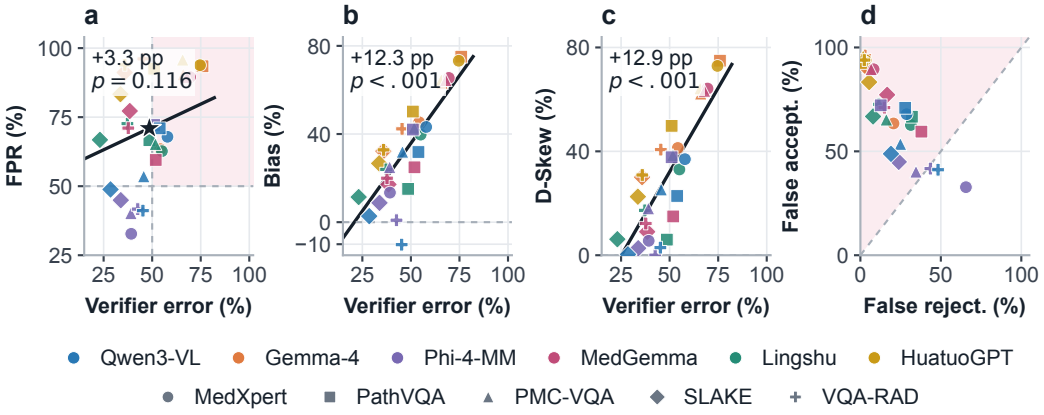


Figure 2: **Analysis of the Verification Mirage across model-dataset settings. (a-c)** Verifier error versus FPR, Bias, and D-Skew across the 30 settings. **(d)** False acceptance versus false rejection. Solid lines in **(a-c)** show fitted trends.

3.3 Generator-Verifier Error Coupling

Models are more likely to accept their own wrong answers. The acceptance asymmetry in Fig. 2 could arise from a generally permissive verifier, without any dependence on who produced the error. We therefore ask a stricter question: *among candidates already known to be wrong, does a verifier accept its own errors more often than errors from other models?* To isolate this dependence, every verifier judges frozen wrong candidates from all six generators across five datasets, under the same prompt and in a fresh context, with generator identity hidden.

We fit a mixed-effects model that separates this own-error effect from two simpler sources of variation: some verifiers may be more permissive overall, and some generators may produce errors that are easier to accept regardless of who verifies them. The resulting self-effect odds ratio therefore measures the additional odds of accepting a wrong candidate when generator and verifier are the same model. An odds ratio above one indicates preferential acceptance of the verifier's own errors. Appendix C.2 gives the full specification. All six verifiers show this own-error dependence, with adjusted ORs from 2.32 to 8.53 and all Holm-adjusted $p < .001$ (Fig. 3). In other words, after accounting for general verifier permissiveness and differences among generators' errors, the odds of accepting a wrong answer are consistently higher when that answer was produced by the verifier itself.

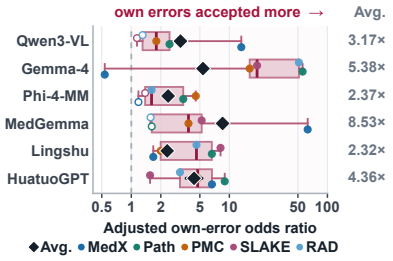


Figure 3: **Own-error dependence across verifiers and datasets.** Points show dataset-specific adjusted odds ratios, and black diamonds show five-dataset pooled estimates with 95% CIs.

3.4 From Existing Explanations to a Discriminative Check

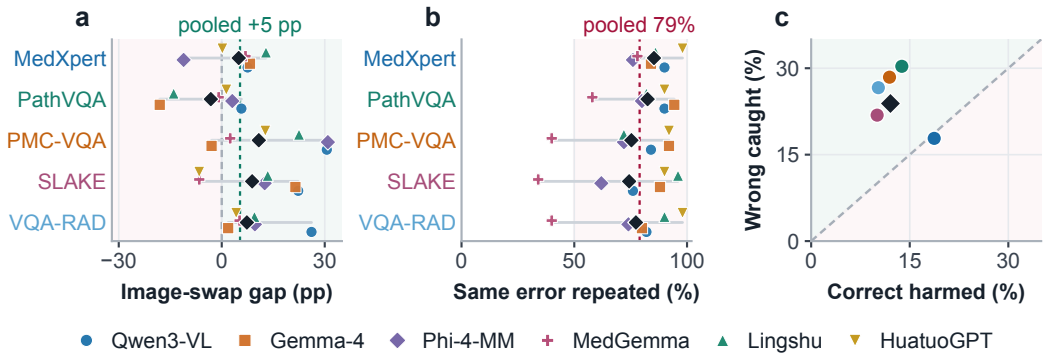


Figure 4: **Analysis of existing explanations and a discriminative visual check. (a)** Image-swap gap. **(b)** Same error repeated under candidate-hidden re-solving. **(c)** Wrong caught versus correct harmed under the falsifiable visual-check probe.

Prior work suggests two potential explanations for the *Verification Mirage*. On the *evidence side*, insufficient visual grounding can produce hallucinated judgments (Lee et al., 2024; Liao et al., 2025a). On the *decision side*, conditioning on a proposed answer can induce agreement bias and over-validation (Andrade et al., 2025). We probe these explanations with two interventions. Image replacement tests whether verifier decisions are selectively tied to the current image, while candidate hiding tests whether removing direct access to the proposed answer breaks the original error.

Using the correct image does not reliably separate correct from incorrect answers. We replace the original image with its nearest non-identical neighbor in BioMedCLIP image-embedding space, restricting the search to the same dataset and task stratum while holding all other inputs fixed. Across the full model-dataset matrix, 66.4% of false acceptances persist after replacement, compared with 70.4% of true acceptances (Fig. 4a). Thus, changing the image can alter the verdict without preferentially removing incorrect acceptances. PMC-VQA makes this distinction especially clear: image replacement preserves 35.3% of false acceptances and 38.3% of true acceptances.

Re-solving without the candidate often repeats the same error. Among known-wrong candidates, candidate-hidden re-solving produces an answer judged identical to or compatible with the original error in 79.0% of cases (Fig. 4b). Removing direct access to the candidate therefore reduces one source of anchoring, but does not reliably produce a competing hypothesis that can falsify the original answer.

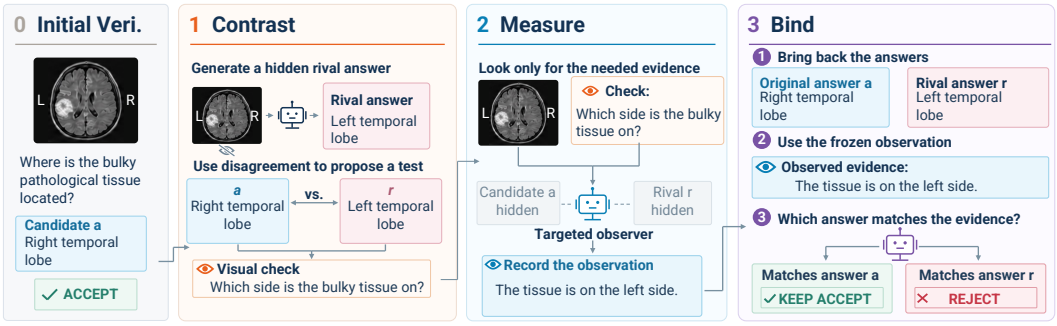


Figure 5: **Overview of `MirageCheck`.** `MirageCheck` rechecks initially accepted candidates through CONTRAST, MEASURE, and BIND, which define a visual distinction, observe candidate-hidden evidence, and bind that evidence back to the competing hypotheses.

Together, the two interventions isolate a narrower failure mode. Verification can depend on the image and can be made candidate-independent, yet still fail to discriminate correct from incorrect acceptances. What is missing is a candidate-specific visual observation that can falsify the answer and govern the verdict.

A falsifiable visual check recovers selective verification. The preceding results suggest why self-verification fails to obtain the usual advantage of checking over solving. Seeing the image again or removing the candidate still leaves the model with much of the original visual problem. We therefore test a narrower requirement: *can the candidate be converted into a small visual distinction whose observed outcome directly determines whether the candidate should remain accepted?*

We evaluate this requirement with a frozen mechanism probe. When an accepted candidate disagrees with a candidate-hidden hypothesis, the pair is used only to specify the smallest visual distinction between them. A fresh observer then measures that distinction without seeing either hypothesis or its role, and a fixed mapping determines whether the sealed observation supports the candidate. The alternative therefore defines what must be tested, but does not act as a vote.

The diagnostic signature is selective revocation: incorrect acceptances should be caught more often than correct acceptances are harmed. In Fig. 4c, this holds in 25 of 30 model–dataset settings. Pooled across cells, the probe catches 23.9% of false acceptances while harming 12.1% of true acceptances, corresponding to 2,809 rescues and 1,064 harms. The probe therefore provides evidence consistent with a candidate mechanism for selective verification: reducing the candidate to a falsifiable visual test and tying the verdict to the resulting observation.

False acceptance can carry errors across self-improvement rounds. When verification is reused as feedback in a five-turn self-improving actor-verifier loop, a local false acceptance can become a persistent positive signal that prevents correction. As shown in Fig. 7, all five datasets, 66.9-96.0% of initially wrong answers finish both wrong and falsely accepted, while only 1.5-5.9% are corrected. Cumulative false acceptance further increases by 9.1-15.2 percentage points on average across datasets. These results expose a system-level risk for verifier-driven self-improvement: repeated feedback does not necessarily wash out verification errors and can instead lock them into subsequent rounds. Appendix G provides the turn-by-turn stress test. Section 4.4 evaluates whether repairing the verification channel weakens this lock-in without sacrificing useful self-correction.

4 MIRAGECHECK: INTERVENING AND REPAIRING SELF-VERIFICATION

4.1 MIRAGECHECK

Section 3.4 motivates a more specific design principle for addressing the Verification Mirage. Verification needs an answer-discriminative visual test, not another solution of the same VQA. Candidate-independent re-solving can still reproduce the original error. Disagreement becomes useful when it specifies *what the image must distinguish* and the resulting evidence governs the verdict.

Table 1: **Verification performance across five medical VQA datasets.** Entries are unweighted six-model means (%); BAcc is $(\mathrm{TPR}+\mathrm{TNR})/2$. Arrows indicate the preferred direction. †Candidate-independent and ‡visual-grounding baselines.

Method	PathVQA		VQA-RAD		SLAKE		MedXpert-QA		PMC-VQA	
	BAcc ↑	FPR ↓	BAcc ↑	FPR ↓	BAcc ↑	FPR ↓	BAcc ↑	FPR ↓	BAcc ↑	FPR ↓
Vanilla verification	52.61	75.87	54.99	69.43	59.34	68.71	**52.68**	68.34	56.36	72.47
Candidate-Hidden Re-solving†	53.71	83.44	52.26	92.30	51.57	95.07	49.88	92.01	53.73	86.66
SGV†	55.74	58.02	57.43	64.21	60.08	62.26	52.27	56.26	60.87	56.22
CausalLens‡	50.98	72.06	55.81	63.05	57.95	63.39	52.14	56.02	55.97	67.48
ClearSight‡	52.04	74.09	54.16	71.02	57.96	67.20	52.37	63.06	54.93	74.18
`MirageCheck` (ours)	**59.51**	**40.68**	**61.55**	**47.63**	**62.59**	**48.93**	51.43	**46.29**	**64.15**	**40.29**
	|Bias| ↓	D-Skew ↓	|Bias| ↓	D-Skew ↓	|Bias| ↓	D-Skew ↓	|Bias| ↓	D-Skew ↓	|Bias| ↓	D-Skew ↓
Vanilla verification	39.86	34.37	21.69	17.34	16.49	11.80	46.73	42.32	48.66	45.27
Candidate-Hidden Re-solving†	48.44	45.81	37.92	36.20	32.64	31.52	69.17	68.00	61.38	60.24
SGV†	26.12	20.36	20.66	15.75	13.21	9.50	34.49	28.40	34.94	30.17
CausalLens‡	34.35	28.21	25.91	19.96	15.54	10.54	34.16	29.98	43.58	40.58
ClearSight‡	37.48	31.60	21.99	17.43	16.59	11.30	41.26	36.46	49.63	46.93
`MirageCheck` (ours)	**16.39**	**10.38**	**11.75**	**5.49**	**9.35**	**4.05**	**24.15**	**19.65**	**21.57**	**16.75**

`MirageCheck` operationalizes this principle by separating test construction, visual measurement, and evidence-to-verdict binding (Fig. 5). It is applied to an initially accepted candidate a for efficiency. In CONTRAST, the VLM first re-answers the image and question in a fresh context without seeing a, producing a rival r. Their answer-critical disagreement is then converted into a neutral visual check that distinguishes the two hypotheses. In MEASURE, a fresh observer receives only the image and this check, with both answers hidden, and records the relevant visual evidence. This observation is frozen before either hypothesis is revealed. In BIND, the frozen evidence is paired with a and r in randomized order to determine which answer it supports. The original acceptance is kept only when the evidence supports a. Otherwise, it is revoked.

This information separation prevents the rival from acting as a second vote. It can define what should be measured, but the verdict is determined only after the evidence has been formed independently of both answers. Initial rejections are unchanged, making `MirageCheck` a targeted intervention on the false-acceptance failure identified by `VeriMap`. All stages use the same frozen VLM checkpoint in fresh contexts and require no additional training, reference answer, or external critic at inference time. Appendix E gives the full details about it.

4.2 EVALUATION PROTOCOL AND MAIN RESULTS

Evaluation setup and baselines. We evaluate all methods on the same frozen candidates produced by the six VLMs across the five medical VQA datasets in `VeriMap`. For each candidate, the image, question, model checkpoint, and proposed answer are fixed. Ground-truth answers are used only for post-hoc evaluation. We retain the four metrics from Section 3.1. This paired setup isolates changes in the verification procedure without changing the underlying generations. For method comparison, we use BAcc. as the primary performance metric. A method can reduce FPR simply by rejecting more candidates, at the cost of increasing false rejection. Because BAcc. weights correct-candidate preservation and incorrect-candidate rejection equally, it measures whether the reduction in false acceptance yields a favorable overall trade-off.

The baselines target the two categories motivated by Section 3.4. For candidate independence, *Candidate-Hidden Re-solving* hides the candidate and re-solves the original image-question pair in a fresh context before comparison (Manakul et al., 2023), while *SGV* forms candidate-independent visual findings before judging the revealed candidate (Andrade et al., 2025). For stronger visual grounding, we adapt two training-free hallucination mitigation methods to verification: *CausalLens* strengthens visually sensitive attention pathways during decoding (Ji et al., 2026), while *ClearSight* amplifies visual signals during multimodal fusion (Yin et al., 2025). Vanilla self-verification provides the baseline reference. Together, these baselines test whether candidate independence or stronger visual grounding is sufficient without the evidence-bound information flow used by `MirageCheck`. Implementation details are provided in Appendix B.

`MirageCheck` improves verification beyond existing baselines. Table 1 compares `MirageCheck` with Vanilla verification, candidate-independent baselines, and visual-grounding baselines across all five datasets. `MirageCheck` yields the lowest FPR, |Bias|, and D-Skew on every dataset, while achieving the highest balanced accuracy on four of five. Compared with SGV, the strongest competing baseline by mean balanced accuracy, `MirageCheck` improves the 30-setting mean from 57.28% to 59.85% and reduces mean FPR from 59.39% to 44.76%. Balanced accuracy improves over SGV on PathVQA, VQA-RAD, SLAKE, and PMC-VQA, but decreases on MedXpert-QA from 52.27% to 51.43%. On that dataset, FPR still falls from 56.26% to 46.29%.

4.3 Mechanistic Analysis and Ablations

`MirageCheck` reduces false acceptance while improving balanced discrimination. Table 1 establishes the aggregate improvement. We next ask whether `MirageCheck` changes the specific failure isolated in Section 3.4, rather than simply shifting the overall acceptance rate. Fig. 6 examines this question across the 30 model and dataset settings. Under Vanilla verification, errors remain strongly skewed toward false acceptance. Candidate-Hidden Re-solving and the visual-grounding interventions leave substantial directional error, while SGV reduces but does not eliminate it. `MirageCheck` reduces false acceptance further.

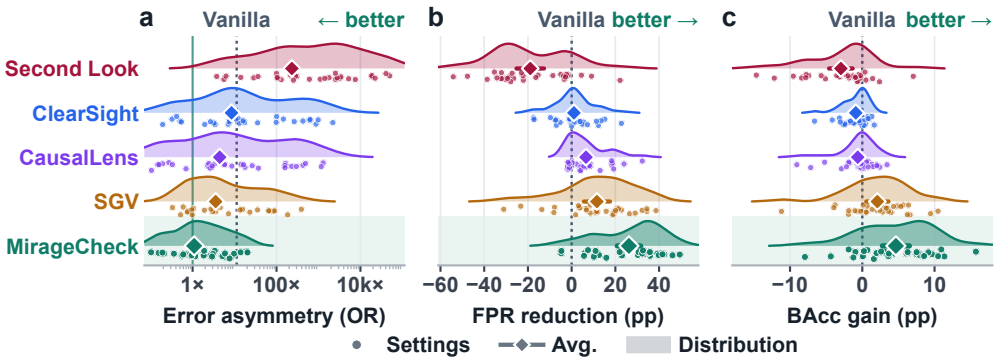


Figure 6: **Comparison of verification methods across 30 model-dataset settings.** **(a)** False-acceptance asymmetry, measured by $\mathrm{odds(FPR)}/\mathrm{odds(FNR)}$. Diamonds show adjusted pooled estimates with 95% confidence intervals. **(b)** FPR reduction relative to Vanilla. **(c)** BAcc gain relative to Vanilla. Points show individual settings, curves show their distributions, and diamonds in **(b,c)** show means with ± 1.96 standard errors.

As presented in Fig. 6, this reduction is accompanied by higher balanced accuracy. Among all settings, `MirageCheck` lowers FPR by 14.63 points and raises BAcc by 2.57 points relative to SGV. Relative to Vanilla, both metrics also improve, as shown in Fig. 6. The joint reduction in false acceptance and increase in BAcc is important because a uniformly more conservative verifier could lower FPR simply by rejecting more candidates. Instead, `MirageCheck` substantially reduces the acceptance-side asymmetry while preserving useful acceptances across most settings.

Removing any `MirageCheck` stage weakens performance. We ablate one stage of `MirageCheck` at a time on a PathVQA subset while keeping the remaining procedure fixed. Each variant removes only the designated stage and introduces no substitute inference.

Table 2: **Ablation of `MirageCheck`.**

Variant	BAcc ↑	FPR ↓	|Bias| ↓	D-Skew ↓
Full `MirageCheck`	**58.28**	**52.98**	**20.08**	**14.72**
w/o Contrast	53.49	72.42	35.27	29.67
w/o Measure	54.61	68.80	32.60	26.46
w/o Bind	45.38	68.80	25.15	17.63

Table 2 shows that the full pipeline performs best on all four metrics. Removing Contrast produces the largest drop, reducing BAcc. from 58.28% to 53.49% and increasing FPR from 52.98% to 72.42%. Removing Measure or replacing Bind with the baseline

also degrades performance, reducing BAcc to 54.61% and increasing FPR to 68.80%. The same pattern holds for |Bias| and D-Skew. These results suggest that the gain is not obtained from contrast construction alone. It depends on carrying the contrast through visual measurement and into the final decision.

4.4 MIRAGECHECK IN THE SELF-IMPROVING ACTOR-VERIFIER LOOP

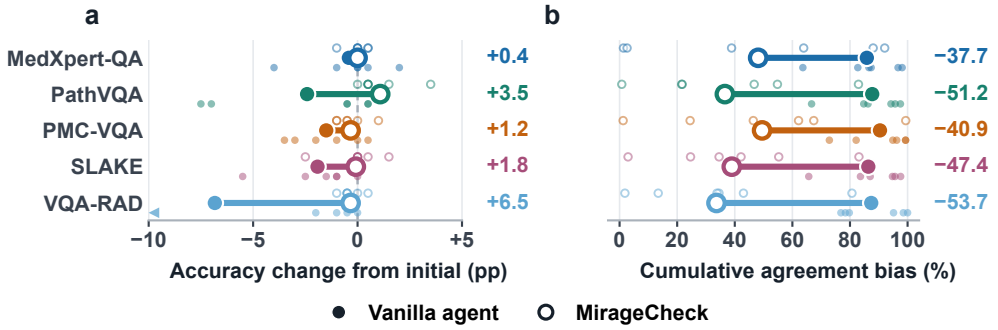


Figure 7: **`MirageCheck` in the self-improving actor-verifier loop.** Matched comparison of Vanilla verification and `MirageCheck` in the same five-turn feedback loop, averaged over six models within each dataset. Filled and open markers denote Vanilla and `MirageCheck`, respectively. **(a)** Final answer accuracy relative to the initial answer. **(b)** Cumulative false acceptance among initially incorrect trajectories. Labels show the paired `MirageCheck` minus Vanilla difference in percentage points.

`MirageCheck` reduces false feedback during iterative self-correction. Section 3.4 showed that false acceptance can become a persistent feedback signal when verification is reused for self-improvement. We therefore rerun the same five-turn self-improving actor-verifier loop and replace only Vanilla verification with `MirageCheck`. Fig. 7 reports the matched comparison across all six models and five datasets. Panel (a) measures final accuracy relative to the initial answer, while panel (b) measures cumulative false acceptance among trajectories that were initially incorrect.

With Vanilla feedback, average accuracy falls from 41.20% initially to 38.58% after five turns. With `MirageCheck`, it instead remains at 41.27%. The same pattern holds at the dataset-average level, with `MirageCheck` achieving higher final accuracy than Vanilla on all five datasets in Fig. 7a. More importantly, cumulative false acceptance falls from 87.67% to 42.72%, with reductions of 37.7 to 53.7 percentage points across datasets (panel b). These results highlight a key limitation of vanilla verifier-driven self-improvement. Repeated feedback is useful only when the verification signal itself is reliable. `MirageCheck` does not produce large gains through repeated correction. Instead, it prevents false verification signals from locking initial errors into subsequent self-improvement rounds.

5 CONCLUSION

We study when visual self-verification provides a discriminative check rather than another judgment of the same problem. `VeriMap` reveals a *Verification Mirage*: medical VLMs frequently accept incorrect candidates, with substantial generator–verifier error coupling when the same model fills both roles. Our analyses show that candidate independence and stronger visual grounding alone are insufficient, and support a design principle in which verification tests an answer-discriminative visual distinction and ties the verdict to the resulting evidence. `MirageCheck` operationalizes this principle through CONTRAST, MEASURE, and BIND, substantially reducing false acceptance while improving balanced accuracy on average across the evaluation suite. Overall, reliable visual self-verification requires making a candidate falsifiable from the current image rather than simply asking the model to judge again.

AI USE STATEMENT

Language models assisted with drafting and figure prototyping. Authors inspected the outputs and retain responsibility for the design, analysis, interpretation, and claims.

REPRODUCIBILITY STATEMENT

To support reproducibility, we provide detailed experimental specifications throughout the paper and appendix. Appendix B reports the evaluation suite, exact model checkpoints, metrics and coverage handling, correctness label validation, and inference settings. Appendices C and D provide the complete statistical analyses and controlled diagnostic experiments. Appendix E specifies the full `MirageCheck` protocol and its information-access contract. Appendices F and G report supplementary evaluations, controls, ablations, and the iterative actor-verifier experiments. Appendix H provides all prompts, parsing rules, and reporting procedures. All experiments use open-source models and fixed evaluation procedures to support matched comparisons. We will publicly release the implementation and evaluation code upon publication.

REFERENCES

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, et al. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*, 2025.

Moises Andrade, Joonhyuk Cha, Brandon Ho, Vriksha Srihari, Karmesh Yadav, and Zsolt Kira. Let's think in two steps: Mitigating agreement bias in mllms with self-grounded verification. *arXiv preprint arXiv:2507.11662*, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Zhenyang Cai, Ke Ji, Xiang Wan, et al. Towards injecting medical visual knowledge into multimodal llms at scale. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 7346–7370, 2024.

Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. Teaching large language models to self-debug. *arXiv preprint arXiv:2304.05128*, 2023.

Philip Chung, Akshay Swaminathan, Alex J Goodell, Yeasul Kim, S Momsen Reincke, Lichy Han, Ben Deverett, Mohammad Amin Sadeghi, Abdel-Badih Ariss, Marc Ghanem, et al. Verifact: Verifying facts in llm-generated clinical text with electronic health records. *arXiv preprint arXiv:2501.16672*, 2025.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. Chain-of-verification reduces hallucination in large language models. In *Findings of the association for computational linguistics: ACL 2024*, pp. 3563–3578, 2024.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yujiu Yang, Nan Duan, Weizhu Chen, et al. Critic: Large language models can self-correct with tool-interactive critiquing. In *International Conference on Learning Representations*, volume 2024, pp. 57734–57811, 2024.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.

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*, volume 2024, pp. 32808–32824, 2024.

Junyang Ji, Qifan Liu, Wenming Yang, and Zhihai He. Causallens: Sensitivity-guided multi-head causal intervention for hallucination mitigation in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4199–4209, 2026.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):180251, 2018.

Seongyun Lee, Sue Hyun Park, Yongrae Jo, and Minjoon Seo. Volcano: mitigating multimodal hallucination through self-feedback guided revision. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 391–404, 2024.

Yuan-Hong Liao, Rafid Mahmood, Sanja Fidler, and David Acuna. Can large vision-language models correct semantic grounding errors by themselves? In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14667–14678, 2025a.

Zehui Liao, Shishuai Hu, Ke Zou, Mengyuan Jin, Yanning Zhang, Huazhu Fu, Liangli Zhen, and Yong Xia. Univrse: Unified vision-conditioned response semantic entropy for hallucination detection in medical vision-language models. *arXiv preprint arXiv:2503.20504*, 2025b.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, volume 2024, pp. 39578–39601, 2024.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pp. 1650–1654. IEEE, 2021.

Hao Liu, Ali Soroush, Jordan G Nestor, Elizabeth Park, Betina Idnay, Yilu Fang, Jane Pan, Stan Liao, Marguerite Bernard, Yifan Peng, et al. Retrieval augmented scientific claim verification. *JAMIA open*, 7(1):ooae021, 2024.

Jiaqi Liu, Kaiwen Xiong, Peng Xia, Yiyang Zhou, Haonian Ji, Lu Feng, Siwei Han, Mingyu Ding, and Huaxiu Yao. Agent0-vl: Exploring self-evolving agent for tool-integrated vision-language reasoning. *arXiv preprint arXiv:2511.19900*, 2025.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems*, 36:46534–46594, 2023.

Razi Mahmood, Diego Machado-Reyes, Joy Wu, Parisa Kaviani, Ken CL Wong, Niharika D'Souza, Mannudeep Kalra, Ge Wang, Pingkun Yan, and Tanveer Syeda-Mahmood. Phrase-grounded fact-checking for automatically generated chest x-ray reports. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 441–452. Springer, 2025.

Potsawee Manakul, Adian Liusie, and Mark Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 9004–9017, 2023.

Aymen Meddeb, Ida Rangus, Paolo Pagano, Insaf Dkhil, Soumaya Jelassi, Keno Bressem, Michael Scheel, Mike P Wattjes, Sonia Nagi, Laurent Pierot, et al. Evaluating the diagnostic accuracy of vision language models for neuroradiological image interpretation. *NPJ Digital Medicine*, 8(1): 666, 2025.

Ansong Ni, Srini Iyer, Dragomir Radev, Veselin Stoyanov, Wen-tau Yih, Sida Wang, and Xi Victoria Lin. Lever: Learning to verify language-to-code generation with execution. In *International Conference on Machine Learning*, pp. 26106–26128. PMLR, 2023.

Arjun Panickssery, Samuel R Bowman, and Shi Feng. Llm evaluators recognize and favor their own generations. *Advances in Neural Information Processing Systems*, 37:68772–68802, 2024.

Jie Ren, Yao Zhao, Tu Vu, Peter J Liu, and Balaji Lakshminarayanan. Self-evaluation improves selective generation in large language models. In *Proceedings on*, pp. 49–64. PMLR, 2023.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

Yuda Song, Hanlin Zhang, Carson Eisenach, Sham Kakade, Dean Foster, and Udaya Ghai. Mind the gap: Examining the self-improvement capabilities of large language models. In *International Conference on Learning Representations*, volume 2025, pp. 39894–39931, 2025.

Linzhuang Sun, Hao Liang, Jingxuan Wei, Bihui Yu, Tianpeng Li, Fan Yang, Zenan Zhou, and Wentao Zhang. Mm-verify: Enhancing multimodal reasoning with chain-of-thought verification. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14100–14115, 2025.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Weiyun Wang, Zhangwei Gao, Lianjie Chen, Zhe Chen, Jinguo Zhu, Xiangyu Zhao, Yangzhou Liu, Yue Cao, Shenglong Ye, Xizhou Zhu, et al. Visualprm: An effective process reward model for multimodal reasoning. *arXiv preprint arXiv:2503.10291*, 2025.

Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. Self-preference bias in llm-as-a-judge. *arXiv preprint arXiv:2410.21819*, 2024.

Tsung-Han Wu, Heekyung Lee, Jiaxin Ge, Joseph E Gonzalez, Trevor Darrell, and David M Chan. Generate, but verify: Reducing hallucination in vision-language models with retrospective resampling. *arXiv preprint arXiv:2504.13169*, 2025.

Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, et al. Lingshu: Generalist foundation model for unified multimodal medical understanding and reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Wenda Xu, Guanglei Zhu, Xuandong Zhao, Liangming Pan, Lei Li, and William Wang. Pride and prejudice: Llm amplifies self-bias in self-refinement. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15474–15492, 2024.

Hao Yin, Guangzong Si, and Zilei Wang. Clearsight: Visual signal enhancement for object hallucination mitigation in multimodal large language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14625–14634, 2025.

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.

Sau Lai Yip, Sunan He, Yuxiang Nie, Shu Pui Chan, Yilin Ye, Sum Ying Lam, and Hao Chen. Medbookvqa: A systematic and comprehensive medical benchmark derived from open-access book. *arXiv preprint arXiv:2506.00855*, 2025.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*, 2024.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023a.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023b.

Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. *arXiv preprint arXiv:2501.18362*, 2025.

Appendix roadmap. Appendix A provides extended related work and positions our study relative to prior work on self-verification, model-based feedback, multimodal verification, and medical VLM reliability. Appendix B describes the evaluation suite, motivates the focus on medical VQA, defines the evaluation metrics, and provides implementation and correctness-labeling details. Appendix C reports additional `VeriMap` analyses, including task-level results and controlled tests of preferential acceptance of a model's own errors. Appendix D tests whether the observed failure can be explained by visual access, candidate conditioning, or additional visual information, and examines the role of discriminative evidence. Appendix E describes the `MirageCheck` protocol and its information-separation design. Appendix F provides additional `MirageCheck` results and controls, including results across models and datasets, matched-rejection controls, and information-hiding ablations. Appendix G studies how verification errors behave when verifier feedback is reused across multiple turns. Finally, Appendix H provides the prompts, semantic-judge instructions, limitations, and broader impact.

A Extended Related Work

Self-verification, self-correction, and verifier reliability. Self-verification and self-evaluation have been used to filter, select, or revise model outputs across reasoning, code, and multimodal tasks (Kadavath et al., 2022; Ren et al., 2023; Ni et al., 2023; Chen et al., 2023; Sun et al., 2025; Wu et al., 2025). However, additional model calls do not necessarily provide a reliable corrective signal: intrinsic self-correction can fail without informative external feedback (Huang et al., 2024), while repeated sampling or self-checking may reproduce rather than resolve an original error (Manakul et al., 2023). Factored Chain-of-Verification (CoVe) similarly plans fact-checking questions from an initial response and answers them in independent contexts before revising the response (Dhuliawala et al., 2024). While it shares the goal of reducing dependence on the original answer, CoVe targets text-only factual correction and does not require verification questions to isolate an answer-discriminative visual distinction or constrain a frozen-candidate verdict by sealed visual evidence. Our diagnostics show that candidate independence alone is insufficient in medical VQA. Model-based judges can also exhibit agreement and source-dependent biases, including preference for their own generations (Xu et al., 2024; Panickssery et al., 2024; Wataoka et al., 2024; Andrade et al., 2025). Our focus is distinct from asking whether another inference pass improves final task accuracy or whether a judge prefers its own outputs in aggregate. `VeriMap` conditions explicitly on candidate correctness and tests whether a verifier is especially likely to accept *its own wrong answers and investigate why*. Its complete cross-generator design separates this own-source excess from general verifier permissiveness and generator-specific error acceptability, while a correctness-balanced source analysis further probes whether the effect is specific to errors rather than general own-source preference. We therefore treat known self-preference as a competing explanation for generator-verifier dependence, rather than equating the two.

Verifier feedback and self-improving systems. Verifier outputs increasingly serve not only as post-hoc judgments but also as feedback for subsequent generation or optimization (Gou et al., 2024; Yuan et al., 2024; Liu et al., 2025). Self-rewarding methods use model-generated judgments as optimization signals, while self-evolving vision-language agents can explicitly couple solver, verifier, and repair roles. This reuse changes the consequence of verification error: a false acceptance is no longer an isolated evaluation mistake, but can become an endogenous positive signal that influences later behavior. Prior work establishes both the promise and the limitations of model-generated feedback, whereas our analysis isolates a complementary reliability condition: whether verification supplies information that can discriminate an actor's error rather than reproduce it. Our multi-turn experiments are therefore a bounded consequence analysis of the failure identified by `VeriMap`, not a claim that all verifier-driven self-improvement degrades. More broadly, the results motivate a reliability requirement for self-improving systems: verification should be evaluated not only by average judgment accuracy, but by whether its errors remain coupled to the failures of the actor whose behavior it supervises.

Multimodal verification and discriminative visual evidence. Multimodal reliability methods improve evaluation or correction through candidate-independent reasoning, stronger visual grounding, retrospective verification, or learned visual reward models (Lee et al., 2024; Andrade et al., 2025; Sun et al., 2025; Wu et al., 2025; Wang et al., 2025; Ji et al., 2026; Yin et al., 2025). Woodpecker

similarly formulates questions and validates visual facts before correcting an MLLM response (Yin et al., 2024). However, it is a post-hoc correction pipeline that uses separate pretrained components to construct visual evidence, whereas our controlled setting asks whether the same frozen VLM can verify its own candidate without an external critic. Moreover, `MirageCheck` uses candidate-rival disagreement only to specify a falsifiable distinction, hides both hypotheses during measurement, and binds the frozen evidence back to anonymous hypotheses before deciding the verdict. Most closely, Self-Grounded Verification reduces candidate-induced agreement by eliciting candidate-independent information before judging the revealed answer (Andrade et al., 2025). These approaches motivate two natural remedies for visual verification failure: candidate independence and stronger visual grounding. `VeriMap` separates these from a stricter requirement: a verifier can reason independently of the candidate and use the correct image while still failing to obtain evidence that distinguishes the candidate from its alternatives. `MirageCheck` operationalizes this diagnosis in the same-model, frozen-candidate setting without additional training or an external critic. A candidate-rival contrast specifies what must be tested, the corresponding visual observation is formed with both answers hidden, and the frozen evidence is subsequently bound back to anonymous hypotheses. This information separation distinguishes `MirageCheck` from a second vote, another holistic re-solution, or a rubric that remains conditioned on the competing answers during evidence formation.

Medical VLM reliability and study scope. Medical and multimodal reliability work has studied hallucination detection, uncertainty estimation, visual grounding, retrieval, external clinical evidence, and specialized verifier or reward models (Mahmood et al., 2025; Chung et al., 2025; Liu et al., 2024; Liao et al., 2025b; Wang et al., 2025). Such approaches can introduce information unavailable to the original model and are complementary to our objective. We deliberately study the stricter self-contained setting in which the same VLM must verify a frozen answer from the current image and question, without a reference answer, retrieved record, separately trained verifier, or external clinical critic. This isolates whether the verification process itself produces a discriminative signal, rather than whether reliability can be recovered by adding a stronger source of information. We focus on medical VQA not because we assume the failure is unique to medicine, but because it provides a revealing testbed for this reliability boundary. As motivated in the main text, many medical VQA questions hinge on fine-grained but auditable distinctions in findings, anatomy, laterality, counts, or spatial relations, allowing the evidence required to test a candidate to be manipulated and inspected explicitly. Our sampled cross-domain comparison additionally places the medical benchmarks in a substantially higher false-acceptance regime than the sampled general-domain benchmarks. This combination makes medical VQA useful both for exposing the failure and for localizing what evidence changes verification. We treat the cross-domain result as motivation rather than evidence of a uniquely medical effect; Appendix B.2 provides the corresponding analysis and scope limitations.

B Experimental Setup and Evaluation Protocol

Table B.1: **Open-weight VLM checkpoints used in the unified audit.** We report the paper-facing model name together with the exact checkpoint identifier.

Model	Checkpoint identifier
Qwen3-VL-8B-Instruct	`Qwen/Qwen3-VL-8B-Instruct`
Gemma-4-E4B-it	`google/gemma-4-E4B-it`
Phi-4-multimodal-instruct	`microsoft/Phi-4-multimodal-instruct`
MedGemma-1.5-4B-it	`google/medgemma-1.5-4b-it`
Lingshu-7B	`lingshu-medical-mllm/Lingshu-7B`
HuatuoGPT-Vision-7B-hf	`FreedomIntelligence/HuatuoGPT-Vision-7B-hf`

B.1 Evaluation Suite

The `VeriMap` audit spans six open-The audit spans six open-weight VLMs and five medical VQA benchmarks, yielding 30 model-dataset settings. The models are Qwen3-VL-8B (Bai et al., 2025), Gemma-4-E4B (Team et al., 2026), Phi-4-Multimodal (Abouelenin et al., 2025), MedGemma-1.5-4B (Sellergren et al., 2025), Lingshu-7B (Xu et al., 2026), and HuatuoGPT-Vision-7B (Chen et al., 2024). The datasets are VQA-RAD (Lau et al., 2018), SLAKE (Liu et al., 2021), PathVQA (He et al.,

2020), MedXpert-QA (Zuo et al., 2025), and PMC-VQA (Zhang et al., 2023b). This 6×5 matrix defines the common evaluation suite used throughout the aggregate analyses. Exact checkpoints are listed in Table B.1, while Fig. B.1 summarizes dataset and task coverage.

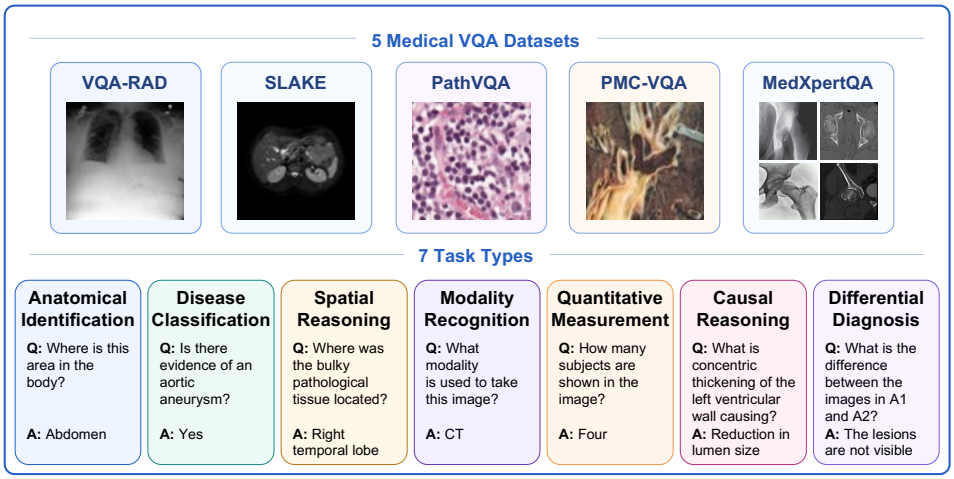


Figure B.1: **Dataset and task coverage of the unified audit.** The evaluation spans five medical VQA datasets and seven mutually exclusive task strata. Examples illustrate the taxonomy used for stratified analysis; the strata characterize task type rather than difficulty.

B.2 Why We Focus on Medical VQA

Medical VQA exposes a pronounced false-acceptance regime. Figure B.2 provides a descriptive comparison between sampled general- and medical-image benchmarks under self-verification. Across all five displayed VLMs, false acceptance is higher on the medical benchmarks, with medical-general differences ranging from 22 to 34 percentage points. The dataset-level view shows the same separation, with both medical benchmarks occupying a substantially higher false-acceptance regime than the two general-image benchmarks. It's worth mentioning that this pattern is not induced by adversarial construction: the incorrect candidates are the models' own predictions on the original benchmark image-question pairs. Consistently, in our full medical evaluation, vanilla same-model verification yields dataset-level FPRs of 68.34-75.87% across the five benchmarks (Table 1).

We treat this cross-domain contrast as motivation (rather than evidence of a causal medical-domain effect). The later within-medical diagnostics instead ask what makes these false acceptances difficult to dislodge: candidate-independent re-solving frequently reproduces the original error, whereas a targeted falsifiable visual check is substantially more selective. Thus, medical VQA provides a setting in which the failure is both pronounced and amenable to controlled analysis of the verification process.

False acceptance is the safety-critical error in medical self-verification. False-positive and false-negative errors need not carry equal costs: in decision-theoretic terms, their relative importance depends on the downstream action induced by the prediction. In our verification formulation, the positive decision is *acceptance*: a false positive therefore endorses an incorrect medical answer as correct, whereas a false negative withholds endorsement from a correct answer. The former can propagate an erroneous finding, anatomical claim, or diagnosis as a positive correctness signal, and becomes especially consequential when verifier outputs are reused for correction or subsequent decision making. We therefore treat FPR as the primary directional reliability measure and report BAcc alongside it to assess the tradeoff with false rejection. This symmetric comparison does not assign equal downstream costs to the two errors.

Medical VQA provides auditable answer-critical visual distinctions. Beyond exhibiting substantial false acceptance, medical VQA offers a useful setting for isolating what verification must discriminate. The affiliated evaluation in this appendix spans modality recognition, anatomical

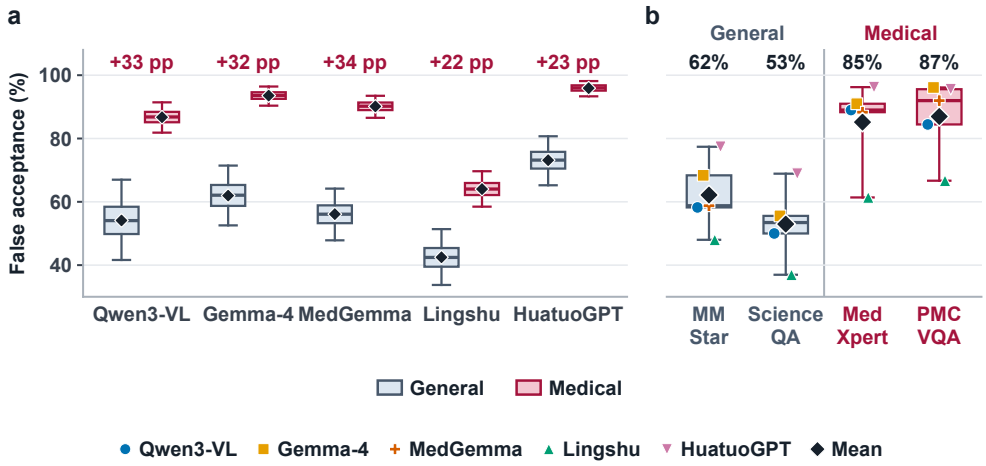


Figure B.2: **False acceptance is higher on the sampled medical VQA benchmarks.** Panel (a) compares false-acceptance rates for five VLMs on sampled general and medical-image benchmarks, with annotations showing the medical minus general difference. Panel (b) shows the corresponding dataset-level distributions across models.

identification, disease classification, spatial reasoning, quantitative measurement, causal reasoning, and differential diagnosis (Fig. B.1). Across these tasks, correctness can often be tied to a concrete visual distinction, for example, the presence of a finding, an anatomical location, a count, or a relative spatial relation that can be inspected independently of the candidate answer. This property enables us to move beyond measuring whether verification fails and ask what visual evidence is sufficient to make the candidate falsifiable. We therefore use medical VQA as a controlled testbed for studying the reliability boundary of visual self-verification.

B.3 Metrics and Evaluation Coverage

Verification as candidate discrimination. For each frozen candidate, the post-hoc label $y \in \{0, 1\}$ indicates whether the candidate is correct, while the verifier decision $\hat{y} \in \{0, 1\}$ denotes rejection or acceptance. A true positive (TP) is a correct candidate that is accepted, a true negative (TN) is an incorrect candidate that is rejected, a false positive (FP) is an incorrect candidate that is accepted, and a false negative (FN) is a correct candidate that is rejected. Here, "positive" refers to verifier acceptance rather than the presence of a clinical condition.

The binary metrics are

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \qquad \text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}, \qquad \text{FNR} = \frac{\text{FN}}{\text{FN} + \text{TP}}.$$

For method comparison, our primary performance metric is balanced accuracy:

$$\text{BAcc} = \frac{\text{TPR} + \text{TNR}}{2} = 1 - \frac{\text{FPR} + \text{FNR}}{2}.$$

Overall accuracy depends on the proportion of correct and incorrect candidates. BAcc weights correct-candidate preservation and incorrect-candidate rejection equally, so it falls when a reduction in FPR is outweighed by an increase in FNR. An always-reject verifier has BAcc of 50% when both candidate classes are present.

Directional summaries. To characterize whether verification errors are concentrated on false acceptance, we additionally report

$$\text{Bias} = \frac{\text{FP} - \text{FN}}{N},$$

where $N = \mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}$. Positive values indicate an excess of false acceptance, negative values indicate an excess of false rejection, and values near zero indicate little net directional imbalance.

We also report

$$\text{D-Skew} = \frac{(\mathrm{FP} - \mathrm{FN})^2}{\mathrm{FP}^2 + \mathrm{FN}^2 + (\mathrm{TP} + \mathrm{TN})(\mathrm{FP} + \mathrm{FN})},$$

which summarizes the magnitude of directional asymmetry while accounting for the overall confusion-table composition. Bias and D-Skew are deterministic summaries of the same confusion table as Accuracy, FPR, and FNR. We therefore use them as complementary descriptions of error direction rather than as statistically independent outcomes.

Together, these metrics capture whether verification separates correct from incorrect candidates, whether its errors are skewed toward false acceptance, and whether an intervention reduces false acceptance without simply trading it for false rejection.

B.4 Validating Answer Correctness

Reference-guided semantic matching. Model answers cannot be reliably evaluated by exact string matching. Correct responses may appear as full sentences, abbreviations, synonyms, or equivalent medical phrasing, while exact matching would incorrectly reject such expressions. We therefore assign candidate correctness with a fixed text-only semantic matching procedure after all verifier decisions are frozen. The judge receives only the question, dataset reference answer, and model answer. It returns a binary correctness label under fixed rules that accept clear synonyms and equivalent medical phrasing, apply strict matching to binary questions, and assign no partial credit. The judge never receives the medical image, verifier output, or generator provenance. It therefore serves as a reference-guided answer matcher rather than an inference-time critic. The answer-matching instructions are summarized in Appendix H.3.

Blinded human validation. We validate this labeling procedure on a dataset-stratified sample of 100 model predictions, with 20 examples drawn from each of the five datasets. An expert annotator, blinded to the frozen semantic labels, independently judges whether each model answer is correct relative to the question and dataset reference. The expert agrees with the semantic judge on 98 of 100 examples, corresponding to 98.0% agreement and Cohen's $\kappa = 0.959$ with a 95% CI of $[0.898, 1.000]$. This audit directly tests whether the automated semantic matching procedure reproduces expert reference-based judgments on a representative sample.

B.5 Inference and Implementation Details

All calls within a model-dataset setting use the same checkpoint listed in Table B.1. Image-conditioned calls receive the current image, while CONTRAST and BIND are text-only. Decoding settings are fixed within each method and protocol stage. Temperature-zero decoding is used for deterministic evaluation calls where specified. The first attempt is capped at 512 tokens, with a fixed 1024-token retry when additional output is required for successful parsing. Discrete selection steps use constrained decoding over the allowed choices. For each case, the evaluation runner records the model request, raw response, parsed stage outputs, and protocol variant. Evaluation records are joined to the corresponding frozen source candidate, initial verifier decision, and post-hoc correctness label.

For baseline implementations, Vanilla verification uses the common verifier prompt in Appendix H. Candidate-Hidden Re-solving re-answers the image-question pair without the candidate and compares the resulting answer with the frozen candidate. SGV follows the candidate-independent evidence-elicitation procedure of Andrade et al. (2025) before judging the revealed candidate. CausalLens and ClearSight are adapted from their released training-free visual-grounding procedures and applied during verifier inference with the frozen candidate and common verification prompt. We use the same VLM checkpoint, image, question, candidate, and decoding settings across matched methods.

C ADDITIONAL VERIMAP ANALYSES

C.1 RESULTS BY TASK TYPE

Medical VQA provides a useful setting for examining the reliability boundary at a finer resolution because its questions span distinct and auditable evidence demands. Broad medical multimodal benchmarks similarly organize evaluation around recurring task families such as anatomical recognition, counting, disease diagnosis, modality understanding, and clinical reasoning (Zuo et al., 2025; Yip et al., 2025). Guided by these recurring demands and by the task composition of our five benchmarks, we use seven task types that capture the dominant forms of evidence required in our evaluation suite: *modality recognition, anatomical identification, disease classification, spatial reasoning, quantitative measurement, causal reasoning, and differential diagnosis*.

Table C.1: **Task-label agreement across datasets.** Majority consensus is the fraction of questions supported by at least two of the three independent annotators. Annotator agreement is the fraction of individual annotations matching the final consolidated label.

Dataset	Majority consensus	Annotator agreement
VQA-RAD	95.8%	85.1%
SLAKE	97.5%	89.9%
PathVQA	81.6%	73.7%
PMC-VQA	94.4%	85.6%
MedXpert-QA	83.7%	73.0%
All datasets	91.4%	82.1%

The task labels are inferred from the question alone and are independent of generator answers, verifier outputs, and correctness labels. Three independent LLM annotators apply the same predefined taxonomy, followed by a fixed majority voting rule. Across 5,338 questions, 91.4% receive majority support. We further validate the taxonomy with a blinded human audit. A PhD-level annotator with medical-knowledge training independently labels 25 randomly sampled questions and agrees with the consolidated labels on 24 of 25 cases, corresponding to 96.0% agreement, Cohen's $\kappa = 0.953$, and macro-F1 $= 0.943$. Table C.1 reports agreement by dataset. Task stratification is used only to provide a finer diagnostic view of the observed failure. We would like to emphasize that the paper's aggregate verification, directional-error, generator-verifier dependence, and mitigation results do not require task labels. Applying `VeriMap` likewise does not require a benchmark to provide a predefined task taxonomy. This part is only provided in the appendix as an additional insight.

Figure C.1 shows that task identity adds resolution to the aggregate reliability picture without changing its central conclusion. Task-model aggregates occupy different parts of the error and false-acceptance space, indicating that verification reliability varies with the evidence demand of the question. At the same time, the 210 task-model-dataset settings overlap substantially, and no task type defines a uniformly reliable or unreliable regime. The directional pattern is more consistent: false acceptance remains the dominant error across the taxonomy. Task stratification therefore refines where the Verification Mirage is most pronounced, while the underlying acceptance-side failure persists across reasoning demands.

Task-specific reliability patterns. Figure C.1 shows how the Verification Mirage varies across the seven reasoning demands. Task-model aggregates occupy different regions of the error and false-acceptance space, indicating that the severity of verification failure depends in part on what evidence the question requires. This variation is not cleanly separable by task type, however. Across the 210 task-model-dataset settings, false-acceptance distributions overlap substantially, and no single task defines a uniformly reliable or unreliable regime. The more consistent pattern is directional: across task types, verification errors remain concentrated on false acceptance rather than false rejection. Task stratification therefore refines where the Verification Mirage is most pronounced, while showing that its acceptance-side failure extends across distinct reasoning demands.

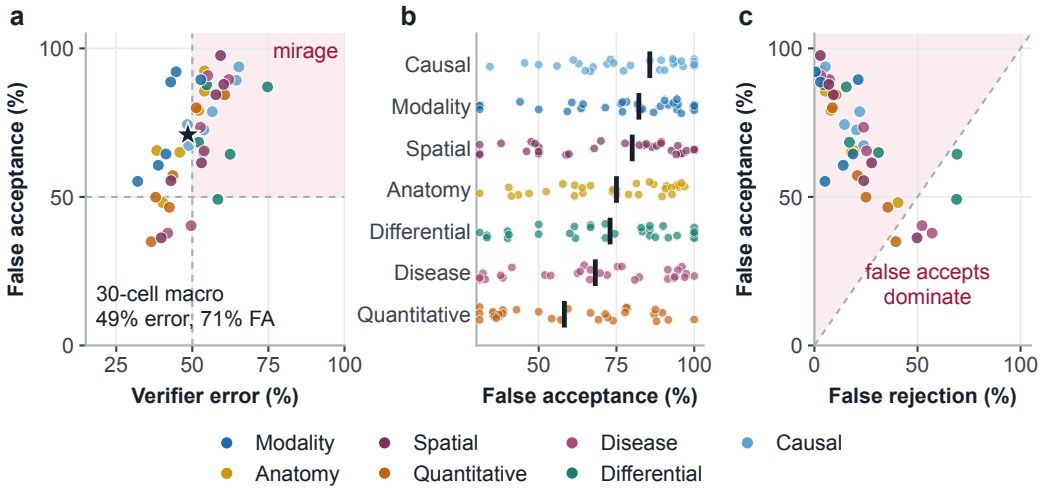


Figure C.1: **Verification reliability across task types.** Panels (a) and (c) show task-model aggregates over five datasets, with the star in panel (a) marking the overall 30-setting macro. Panel (b) shows all 210 task-model-dataset settings, with black ticks marking task medians. Verification behavior varies across task types, but their false-acceptance distributions overlap substantially. Across the taxonomy, errors remain predominantly concentrated on false acceptance rather than false rejection. Colors denote task type.

C.2 False Acceptance vs. False Rejection

A high verifier error rate does not reveal how the verifier fails. Our main concern is false acceptance: an incorrect candidate is endorsed as correct. We therefore ask whether verification errors become increasingly skewed toward false acceptance as overall verification becomes less reliable.

We analyze the same 30 model-dataset settings shown in Figure 2. For each directional metric, we fit a linear mixed-effects model with verifier error as the fixed-effect predictor and random intercepts for model and dataset. Models are fit by REML, and each slope is compared with an intercept-only model using a Kenward-Roger test.

For every 10 percentage-point increase in verifier error, FPR increases by 3.3 points ($p = .116$), signed Bias increases by 12.3 points ($p < .001$), and D-Skew increases by 12.9 points ($p < .001$). The FPR slope is not statistically significant under the Kenward-Roger test, while both asymmetry slopes are.

The record-level analysis gives the same conclusion. Across all self-verification decisions, the adjusted odds of verifier error are 56.2 times higher when the frozen candidate is incorrect (95% CI $[51.1, 61.9]$). This odds ratio is equivalent to $\mathrm{odds(FPR)}/\,\mathrm{odds(FNR)}$, measuring the likelihood of the directional preference.

Together, these results establish the directional component of the Verification Mirage. Verification does not simply make frequent errors. Its errors are disproportionately concentrated on accepting incorrect candidates. The next analysis asks whether this failure is especially strong for errors produced by the verifier itself.

C.3 Preferential Acceptance of a Model's Own Errors

Directional false acceptance does not show that generator and verifier errors are linked. A permissive verifier could accept many wrong answers regardless of which model produced them. We therefore ask a stricter question: among candidates that are already known to be wrong, does a verifier accept its own errors more often than errors produced by other models?

We build a complete generator-verifier matrix for each dataset. Every incorrect frozen candidate from each of the six generators is judged by all six verifiers under the same prompt. Generator identity is

hidden from the verifier. Because every candidate in this analysis is incorrect, every acceptance is a false acceptance. Across five datasets, the matrix contains 122,364 judgments over 20,394 candidate records and 4,646 questions. All 180 dataset-generator-verifier settings are complete.

We fit the following binomial logistic mixed-effects model:

$$\texttt{accept} \sim \texttt{verifier} + \texttt{generator} + \texttt{dataset} + \texttt{verifier:self} + (1 \mid \texttt{question_id}) + (1 \mid \texttt{input_id}) + (1 \mid \texttt{question_id:verifier}).$$

Here, `self` indicates that generator and verifier are the same model. The verifier terms account for differences in general false-acceptance tendency. The generator terms account for differences in how readily each model's errors are accepted. The verifier-specific `self` term therefore measures the additional odds of accepting a wrong answer when it was produced by the verifier itself. We report Wald 95% confidence intervals and apply Holm correction across the six verifier-specific tests.

As shown in Table C.2, all six verifiers show a significant own-error excess. Adjusted odds ratios range from 2.32 to 8.53, with all Holm-adjusted $p < .001$. The corresponding model-dataset estimates are shown in Figure 3.

Table C.2: **Preferential acceptance of each verifier's own errors.** Adjusted odds ratios compare acceptance of a verifier's own incorrect candidates with incorrect candidates produced by other models after accounting for verifier, generator, and dataset effects. Values above one indicate greater acceptance of the verifier's own errors.

Verifier	OR	95% CI	Holm p
Qwen3-VL-8B	3.17	[2.81, 3.58]	$< .001$
Gemma-4-E4B	5.38	[4.68, 6.19]	$< .001$
Phi-4-Multimodal	2.37	[2.11, 2.67]	$< .001$
MedGemma-1.5-4B	8.53	[7.37, 9.87]	$< .001$
Lingshu-7B	2.32	[2.06, 2.62]	$< .001$
HuatuoGPT-Vision-7B	4.36	[3.63, 5.25]	$< .001$

As shown, this result goes beyond general verifier permissiveness. The comparison is made after accounting for both the verifier's overall acceptance tendency and systematic differences among generators' errors. Even under these controls, every verifier is more likely to accept an error produced by itself. This is the primary evidence for the own-error dependence reported in the main paper.

The result does not identify the mechanism behind this dependence. Shared blind spots, answer style, or other model-specific properties may contribute. We therefore next test whether the effect can be explained more simply by a general preference for self-generated outputs.

C.4 Is This Just General Self-Preference?

The controlled analysis above shows that verifiers preferentially accept their own wrong answers. However, this does not yet establish that the effect is specific to errors. A model may simply favor its own outputs regardless of whether they are correct. In this subsection, we therefore ask whether own-source preference is stronger for incorrect than for correct candidates.

We construct a correctness-balanced setting with equal numbers of correct and incorrect candidates from every generator-dataset stratum, yielding 1,200 frozen candidate records. All six verifiers freshly judge every candidate under the same vanilla prompt, with generator identity hidden. Correctness labels are frozen before verification. The resulting design contains 7,200 judgments and balances candidate correctness within every generator-dataset stratum.

For each verifier, we first measure its preference for its own outputs over outputs from other models, separately for correct and incorrect candidates. We then ask whether this preference is stronger when the candidate is wrong. A positive wrong-minus-correct difference indicates stronger own-source preference on errors. This analysis accounts for systematic differences across verifiers, generators, and datasets.

As shown in Figure C.2, Panel (a) shows that verifiers can prefer their own outputs for both correct and incorrect candidates. Panel (b) asks whether that preference is stronger when the candidate is

wrong. Averaged across the six verifiers, the wrong-minus-correct difference is $+7.09$ percentage points (95% CI $[2.21, 11.97]$, $p = .00439$). This indicates that own-source preference is stronger for incorrect candidates on average.

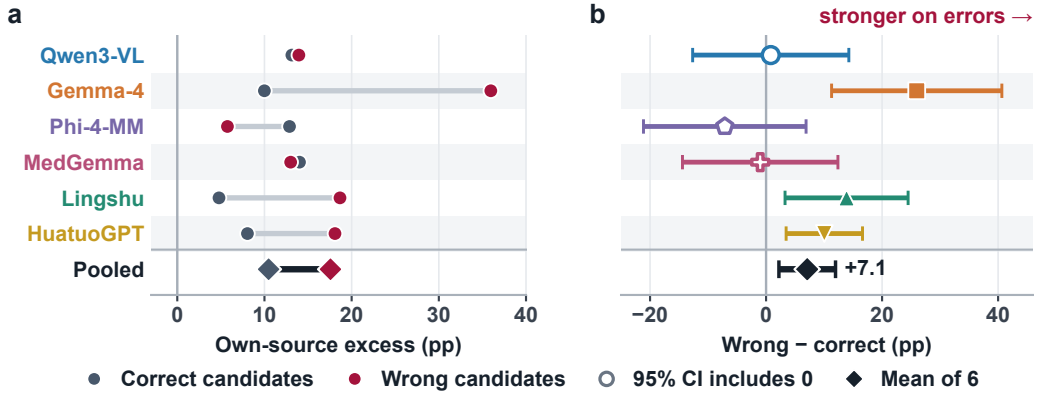


Figure C.2: **Correctness-conditioned own-source preference.** (a) Adjusted acceptance excess for each verifier's own outputs over outputs from other models, shown separately for correct and incorrect candidates. (b) The corresponding wrong-minus-correct difference. Positive values indicate stronger own-source preference when the candidate is wrong. Horizontal bars show 95% confidence intervals, with shared images treated as clusters. Pooled estimates weight the six verifiers equally.

Table C.3: **Correctness-conditioned source preference by verifier.** Wrong and correct excess report adjusted own-versus-other acceptance differences. Their difference is the wrong-minus-correct contrast shown in Figure C.2b. Adjusted odds ratios report the corresponding multiplicative interaction with Holm-corrected p values.

Model	Wrong excess	Correct excess	Difference [95% CI]	Adj. OR	p_H
Qwen3-VL	13.95	13.14	0.80 [-12.66, 14.27]	0.94	1
Gemma-4	35.96	9.99	25.97 [11.30, 40.65]	3.52	0.0899
Phi-4-MM	5.75	12.87	-7.12 [-21.13, 6.89]	0.72	1
MedGemma	13.00	14.02	-1.01 [-14.41, 12.39]	0.82	1
Lingshu	18.66	4.79	13.88 [3.27, 24.49]	1.66	0.86
Huatuo	18.09	8.05	10.04 [3.46, 16.62]	0.99	1

These results distinguish own-error dependence from a general preference for self-generated outputs. Verifiers can favor their own answers even when those answers are correct, but this preference is stronger on incorrect answers on average. Generic self-preference therefore cannot fully account for the own-error effect in Section C.3.

D Mechanism Diagnostics

Section 3.4 points to a gap between having access to the image and obtaining evidence that can actually test a candidate. *Here we examine that gap under controlled interventions and provide new outcomes to supplement the study in main paper.* We first ask whether stronger access to the visual input is sufficient for discrimination. We then remove direct candidate conditioning and ask whether an independent solution escapes the original error. Finally, we test whether the useful ingredient is simply more correct visual information, or information that directly distinguishes the candidate from an alternative.

The image-replacement and candidate-hidden re-solving analyses use the same diagnostic setup. For each of the 30 model-dataset settings, we sample balanced correct and incorrect answers. The image, question, candidate, prompt, and decoding settings are held fixed except for the intervention under study. Reference answers are used only after inference for correctness stratification. The

trusted-region and controlled spatial analyses use annotation-constrained SLAKE subsets described below. We note that `VeriMap` audits the whole datasets in the main text, but we only sample data for further diagnostics here for efficiency.

D.1 Visual Access Does Not Guarantee Discrimination

Image dependence is present, but only weakly selective. We first test whether the Verification Mirage can be explained by insufficient dependence on the current image. We replace the original image with its nearest nonidentical neighbor under BiomedCLIP cosine similarity, restricting the search to the same dataset and task stratum (Zhang et al., 2023a). When no eligible image exists within the stratum, we fall back to the same dataset, which occurs in 17 of 3,000 cases.

As shown in Table D.1, we evaluate candidates that were accepted under the original image and ask whether that acceptance survives the replacement. Among 1,055 originally false acceptances, 784 remain accepted, corresponding to 74.3%. Among 1,218 true acceptances, 970 remain accepted, corresponding to 79.6%. The resulting correct-minus-incorrect retention gap is 5.33 percentage points (95% image-cluster bootstrap CI [0.75, 9.93], $p = .023$). Excluding the fallback cases gives a similar gap of 5.23 points.

Thus, verifier decisions do respond to image content, but the response is only weakly selective in this intervention. Nearly three quarters of the original false acceptances survive the image change. *We therefore do not interpret this result as evidence that the verifier ignores the image.* The nearest replacement may also preserve the relevant finding. Instead, the experiment shows that dependence on the current image does not by itself provide a strong signal for separating correct from incorrect acceptances.

Table D.1: **Acceptance after image replacement.** Among 2,273 originally accepted candidates, the original image is replaced by its nearest nonidentical neighbor. Intervals use 5,000 image-cluster bootstrap resamples.

Initial candidate	Retained / accepted	Retention [95% CI]
Incorrect	784/1055	74.31 [70.83, 77.64]
Correct	970/1218	79.64 [76.22, 82.67]
Selectivity gap	NA	5.33 [0.75, 9.93]
Gap without fallback	NA	5.23 [0.65, 9.85]

Even trusted access to the relevant region does not reliably improve discrimination. Image replacement is an indirect test of visual dependence. A stronger possibility is that the verifier has access to the correct image but fails to locate the relevant content. We test this explanation using SLAKE native bounding boxes, which define a reference-assisted offline region of interest for 194 eligible model-question pairs spanning 42 image clusters (Liu et al., 2021). We either mark the trusted region on the original image or crop the image to that region.

Marking the region reduces acceptance by 7.63 points for incorrect candidates and 6.35 points for correct candidates. The resulting change in discrimination, defined as correct-candidate acceptance minus incorrect-candidate acceptance, is +1.28 points (95% image-cluster CI [-7.95, 10.99]). Cropping reduces acceptance by 13.74 points for incorrect candidates and 23.81 points for correct candidates, giving a discrimination change of -10.07 points (CI [-24.86, 4.34]). Neither intervention establishes improved discrimination after Holm correction, with $p = .789$ for marking and $p = .355$ for cropping.

Together, these results separate visual access from discriminative evidence. We do not find statistically meaningful evidence that providing the correct image or the relevant region improves the separation between correct and incorrect candidates. Verification therefore requires more than locating the relevant pixels. It requires an observation that can distinguish the proposed answer from its alternatives.

Table D.2: **Trusted-region access does not reliably improve discrimination.** Results are paired changes in acceptance on 194 eligible SLAKE model-question pairs. Discrimination change is the change for correct candidates minus the change for incorrect candidates. Positive values are favorable. Values are percentage points with image-cluster 95% intervals.

Intervention	Δ false accept	Δ true accept	Δ discrimination
Marked	-7.63 [-16.05, -0.84]	-6.35 [-13.11, -1.45]	1.28 [-7.95, 10.99]
Crop	-13.74 [-21.21, -6.34]	-23.81 [-36.24, -11.76]	-10.07 [-24.86, 4.34]

D.2 From Independent Re-Solving to a Falsifiable Check

Candidate hiding rarely repairs the original error. The previous analyses ask whether stronger visual access is sufficient. We next test the second natural explanation from Section 3.4: verification may fail because the proposed answer directly anchors the verifier. If candidate conditioning were the main obstacle, hiding the candidate and solving the image-question pair again should often recover the correct answer.

We therefore independently rescore both the frozen candidate and its candidate-hidden re-solution after inference. Across 1,503 originally incorrect diagnostic candidates, only 125 re-solutions are correct, corresponding to 8.32% (95% image-cluster bootstrap CI [6.85, 9.95]). Among the remaining cases, 1,138, or 75.7%, are classified by the original semantic-relation annotation as repeating or remaining compatible with the same error. Another 240, or 16.0%, produce a different or unresolved error.

To verify that this low correction rate is not driven by the original correctness labels, we rescore both the frozen candidates and their candidate-hidden re-solutions with the same fresh semantic judge. Only five of the 1,503 original candidates change from incorrect to correct under this rescoring. Among the remaining 1,498 candidates judged incorrect by both procedures, re-solving corrects only 120 cases, or 8.01% (95% CI [6.62, 9.58]). Correction also remains limited for both binary-answer questions (13.9%) and open-answer questions (7.3%), indicating that the low recovery rate is not driven by a single question format.

The 79.0% result in Figure 4b measures whether candidate-hidden re-solving reproduces an answer that is semantically compatible with the original error. The analysis above asks a complementary and stricter question: *does re-solving actually recover the correct answer?* Both results point to the same conclusion. Hiding the candidate often leaves the underlying error unresolved and rarely provides a reliable corrective signal.

Table D.3: **Outcomes of candidate-hidden re-solving.** Both the frozen candidate and the re-solved answer are independently rescored after inference. Correction uses the fresh correctness labels. Same-error identity uses the original semantic-relation annotation. Values are percentages with 95% image-cluster bootstrap intervals.

Cohort	n	Corrected	Same error	Other / unresolved
Frozen wrong	1503	8.32 [6.85, 9.95]	75.72 [73.44, 78.08]	15.97 [13.96, 17.99]
Both judges wrong	1498	8.01 [6.62, 9.58]	75.97 [73.71, 78.30]	16.02 [14.01, 18.03]

Disagreement becomes useful when it defines what to test rather than acting as a second vote. Candidate-hidden re-solving removes direct access to the proposed answer, but it still asks the model to solve essentially the same visual problem again. We therefore change the role of the second answer. When a candidate-hidden answer disagrees with an accepted candidate, the pair is used only to define the smallest visual distinction between the two hypotheses. A fresh observer then measures that distinction without seeing either answer or knowing which one is the original candidate. The observation is frozen before a role-blind reconstruction step maps it back to an anonymous hypothesis.

The competing answer therefore specifies what the image must distinguish, but it is not trusted as evidence and does not act as a vote. The diagnostic signature is selectivity. A useful check should revoke incorrect acceptances more often than it revokes correct ones. In the frozen probe summarized in Figure 4c, this occurs in 25 of 30 model-dataset settings. Across all settings, the probe catches 23.9% of false acceptances while harming 12.1% of true acceptances.

This contrast identifies a more specific requirement than candidate independence alone. A second solution can reproduce the original failure. Disagreement becomes useful when it is converted into a visual test whose observed outcome determines which hypothesis remains supported. This probe motivates the information flow used by `MirageCheck`.

D.3 Relevant Evidence Matters More Than Extra Information

The preceding falsification probe suggests that targeted evidence can make verification more selective. A remaining question is whether this benefit comes simply from giving the verifier more correct visual information. We isolate this possibility using 40 distinct SLAKE images with one unambiguous native bounding box per image.

A controlled probe separates discriminative evidence from additional information. For each image, we construct questions asking whether the box center lies to the left or right of the image midpoint, or above or below it. Two opposite candidate answers balance correctness. Five verifiers receive one of three inputs: no additional fact, the true coordinate on the orthogonal axis, or the true coordinate on the axis needed to answer the question. Images and candidates are identical across conditions, giving 400 candidate-model cases per condition and 1,200 calls in total at temperature zero. All selected images are square, so resizing preserves coordinate fractions.

Both supplied coordinates are correct annotation facts. The difference is whether the fact can resolve the question being asked. For a left-versus-right question, for example, the horizontal coordinate is discriminative while the vertical coordinate is not. This design therefore separates the value of answer-relevant evidence from the value of providing additional correct information.

Table D.4: **Verification with visual facts.** SLAKE images are used to construct horizontal and vertical box-center questions with two opposing candidate answers. Each condition includes 400 matched candidate-model cases across five verifiers. Yield is the fraction of all 400 attempts that return a parseable and correct verdict. Binary Acc. is computed over parseable verdicts only. Supplied facts are derived from native bounding-box annotations.

Provided fact	Explicit verdicts	Yield [95% CI]	Binary Acc. [95% CI]
None	355/400	48.25 [43.75, 52.26]	54.80 [49.87, 59.34]
Orthogonal coordinate	332/400	44.25 [40.00, 48.75]	52.88 [48.70, 57.00]
Requested coordinate	379/400	57.25 [54.00, 60.25]	60.42 [57.44, 63.16]

Only the discriminative fact produces a clear improvement. Because output availability differs across conditions, we also report correct-verdict yield per attempt. This exploratory outcome was chosen after inspecting output availability.

As shown in Table D.4 and Figure D.1a, correct-verdict yield is 48.25% with no additional fact, 44.25% with the orthogonal coordinate, and 57.25% with the requested coordinate. Providing the requested-axis fact improves yield over the orthogonal-axis fact by 13.00 percentage points (95% paired image-cluster bootstrap CI [8.75, 17.00]). It also improves yield over providing no fact by 9.00 points (CI [4.75, 13.25]). Both contrasts remain significant after Holm correction with $p < 10^{-4}$.

We also test whether differing output availability could explain the result. Unavailable outputs occur in 4 cases with no fact, 18 cases with the orthogonal fact, and 21 cases with the requested fact. Assigning unavailable outcomes adversarially still leaves positive requested-axis advantages of [8.50, 18.25] points over the orthogonal fact and [8.00, 14.25] points over no fact.

Together, these results provide a limited sufficiency test. Additional correct visual information does not automatically improve verification. The benefit appears when the supplied fact directly

resolves the distinction on which the candidate's correctness depends. Because these facts come from privileged annotations, this experiment does not establish that the deployed MEASURE stage can always recover the required evidence.

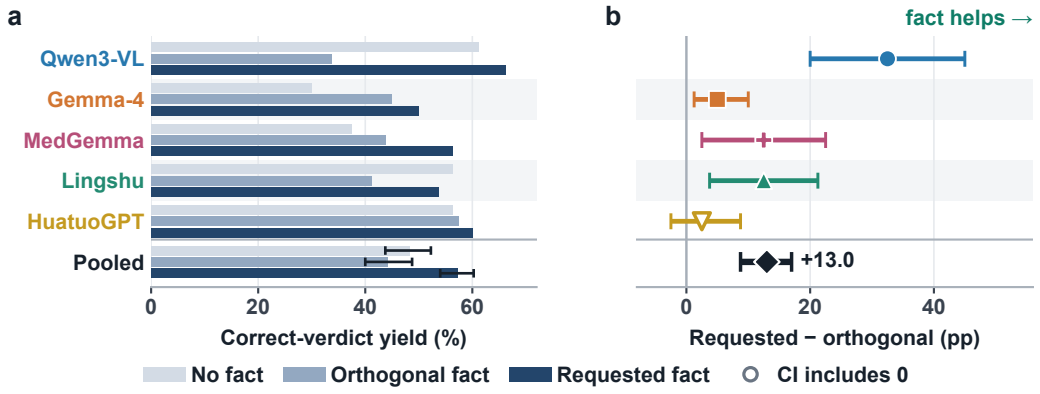


Figure D.1: **Discriminative visual facts improve verification.** (a) Correct-verdict yield for each verifier and the pooled result with no additional fact, a correct but non-informative orthogonal-axis fact, or the requested-axis fact needed to answer the question. (b) Paired improvement from the requested-axis fact over the orthogonal-axis fact for each verifier and in the pooled result. Each verifier is evaluated on 80 paired candidate cases from 40 images, and the pooled estimate weights verifiers equally. Error bars show 95% image-cluster bootstrap confidence intervals. Hollow markers in panel (b) indicate intervals that include zero. All supplied facts come from native bounding-box annotations.

Taken together, the mechanism diagnostics narrow the missing ingredient beyond visual access or candidate independence alone. Verification becomes more selective when the available evidence directly resolves the visual distinction that determines whether the candidate is correct. Appendix E operationalizes this criterion in `MirageCheck` by separating contrast construction, visual measurement, and evidence-to-verdict binding.

E MIRAGECHECK: METHOD DETAILS

Appendix D identifies two requirements for a discriminative visual check. First, a competing answer should define what the image must distinguish rather than act as a second vote. Second, the relevant visual evidence should be formed before the verifier knows which hypothesis is the original candidate. `MirageCheck` implements these requirements through CONTRAST, MEASURE, and BIND.

E.1 PROTOCOL

From disagreement to an evidence-bound decision. `MirageCheck` is applied only to candidates initially accepted by the verifier. Let (x, q, a) denote the image, question, and frozen candidate. The model first answers the image-question pair in a fresh context without seeing a, producing a rival answer r.

In CONTRAST, the disagreement between a and r is converted into a neutral visual question that distinguishes the two hypotheses. The image is hidden at this stage, so CONTRAST specifies what should be checked rather than answering the check itself. In MEASURE, a fresh observer receives only the image and this visual question. Neither answer is visible. The observer records the relevant visual evidence, which is then frozen.

In BIND, the frozen evidence is shown together with the two answers in randomized anonymous slots. The model selects the answer best supported by the evidence without knowing which slot contains the original candidate. The initial acceptance is retained only if the selected answer maps back to a. Otherwise, it is revoked.

Table E.1: **Information barriers in `MirageCheck`.** Each stage receives only the information required for its role. The competing answers are hidden while visual evidence is formed, and candidate identity is hidden while that evidence is matched back to the hypotheses.

Stage	Purpose	Available information	Hidden information
CONTRAST	Define what the image must distinguish	Candidate and rival	Image
MEASURE	Observe the relevant visual evidence	Image and visual question	Candidate and rival
BIND	Match the frozen evidence to a hypothesis	Frozen evidence and anonymous answers	Candidate identity

The rival defines the test, not the verdict. All calls within a model-dataset setting use the same frozen VLM checkpoint in fresh contexts. Reference answers, post-hoc correctness labels, and generator identity are never available during inference. The rival answer is used only to expose a disagreement that can be tested visually. It is not treated as evidence and does not vote against the candidate. Visual evidence is formed before either answer is revealed to the observer, and candidate identity remains hidden while that evidence is matched back to the two hypotheses.

E.2 WHAT EACH STAGE CAN SEE

The three stages intentionally receive different information (Table E.1). CONTRAST sees the competing answers but not the image, allowing it to define the distinction that needs to be tested without producing the visual evidence itself. MEASURE sees the image and the registered visual question but not either answer, preventing the candidate or rival from directly shaping the observation. BIND receives the frozen evidence and the two answers in anonymous slots, but does not know which answer is the original candidate. Candidate identity is restored only after the evidence-to-answer match has been made.

This separation gives each stage a single role. CONTRAST determines what to check, MEASURE determines what the image shows, and BIND determines which hypothesis is supported by that observation. The protocol therefore prevents the rival from becoming a second vote and prevents candidate identity from directly influencing visual evidence formation.

F ADDITIONAL MIRAGECHECK RESULTS AND CONTROLS

F.1 EVALUATION COVERAGE

Evaluation coverage. All five datasets are used exclusively for evaluation. All methods operate on frozen source candidates, and correctness labels are assigned post hoc. The full evaluation comprises 30 model-dataset settings across six VLMs and five medical VQA benchmarks. Some auxiliary analyses and ablations are conducted on subsets of these datasets for computational efficiency or because particular annotations are required, as specified in the corresponding sections.

Pairing and inference cost. Method comparisons use matched source records whenever the corresponding outputs are available. Failed calls are tracked separately. On the fully recorded six-model VQA-RAD and SLAKE dataset, the implemented `MirageCheck` pipeline requires 9.17 model requests per sample on average.

F.2 RESULTS ACROSS MODELS AND DATASETS

Performance across models. Table F.1 reports five-dataset macros by model. Relative to SGV, `MirageCheck` improves macro BAcc and reduces macro FPR for all six models. Effect sizes vary substantially across checkpoints, so the aggregate gain is not driven by one model.

Setting-level consistency and exceptions. Table F.2 summarizes paired comparisons with SGV in the preferred direction across the 30 model-dataset settings. BAcc improves in 23 settings, ties for

Table F.1: **Per-model comparison across verification methods.** Results are macro-averaged over five datasets.

	Qwen3-VL		Gemma-4		Phi-4-MM		Lingshu		Huatuo		MedGemma	
Method	BAcc ↑	FPR ↓	BAcc ↑	FPR ↓	BAcc ↑	FPR ↓	BAcc ↑	FPR ↓	BAcc ↑	FPR ↓	BAcc ↑	FPR ↓
Vanilla	56.90	56.47	53.21	86.92	58.81	46.35	56.71	66.81	52.67	91.89	52.88	77.36
Exact Second Look	51.57	94.14	50.17	93.80	58.54	64.82	51.65	95.73	51.13	95.88	50.33	95.00
SGV	61.24	39.81	54.52	82.44	56.50	58.40	61.37	42.26	52.20	80.83	57.83	52.63
CausalLens	58.41	45.51	53.26	86.84	55.63	**29.75**	54.94	57.09	52.30	88.44	52.87	78.76
ClearSight	54.42	68.22	53.45	86.56	56.54	36.12	55.03	64.58	52.83	92.07	53.49	71.89
`MirageCheck`	**62.51**	**29.11**	**61.50**	**55.28**	**59.66**	37.64	**62.21**	**27.69**	**55.21**	**67.04**	**57.98**	**51.83**

	Qwen3-VL		Gemma-4		Phi-4-MM		Lingshu		Huatuo		MedGemma	
Method	|Bias| ↓	D-Skew ↓	|Bias| ↓	D-Skew ↓	|Bias| ↓	D-Skew ↓	|Bias| ↓	D-Skew ↓	|Bias| ↓	D-Skew ↓	|Bias| ↓	D-Skew ↓
Vanilla	23.98	17.67	51.65	49.83	18.01	12.80	26.59	20.21	49.51	48.07	38.38	32.73
Exact Second Look	51.87	51.03	56.66	55.82	35.87	30.56	50.03	49.71	51.85	51.07	53.15	51.91
SGV	9.73	4.01	48.69	46.31	27.67	20.16	10.02	4.13	38.42	33.89	20.77	16.52
CausalLens	18.34	11.27	51.52	49.80	13.70	6.94	15.17	9.77	45.83	43.13	39.67	34.22
ClearSight	29.57	24.22	51.46	49.65	13.05	6.00	22.89	16.60	49.88	48.65	33.49	27.34
`MirageCheck`	**5.88**	**1.72**	**27.68**	**21.60**	**9.54**	**3.96**	**9.04**	**4.04**	**27.62**	**20.35**	**20.10**	**15.93**

MedGemma on VQA-RAD, and declines for five models on MedXpert-QA and Lingshu on SLAKE. FPR improves in 29 settings and ties in one. Thus, the broad reduction in false acceptance coexists with specific BAcc regressions across the evaluation matrix.

Table F.2: **Setting-level consistency of `MirageCheck` versus SGV.** Counts compare the 30 model-dataset settings in each metric's preferred direction, using the reported two-decimal setting values.

Metric	Better	Tie	Worse
BAcc	23	1	6
FPR	29	1	0
|Bias|	25	1	4
D-Skew	25	1	4

F.3 Comparison with Simple Rejection Controls

We test whether the improvement from `MirageCheck` can be explained by simply rejecting more initial acceptances or by using disagreement from one candidate-hidden re-solve.

Matched random rejection. As a control for rejection alone, we randomly revoke initial acceptances using the same rejection budget as `MirageCheck`. This reduces FPR relative to Vanilla, but increases FNR enough that BAcc decreases from 56.36% to 55.32% (Table F.3). Thus, the gain from `MirageCheck` cannot be explained by a more conservative acceptance rate alone.

Candidate-hidden re-solve gate. We next test a stronger baseline that uses only candidate-independent re-solving. For each initially accepted candidate, the same VLM re-answers the original image–question pair once at temperature zero using the original VQA prompt, without access to the frozen candidate. The initial acceptance is retained when the re-solved answer is semantically consistent with the candidate and revoked otherwise. Initial rejections remain unchanged. This baseline requires no CONTRAST, MEASURE, or BIND stage.

The re-solve gate improves BAcc from 56.36% to 58.24%, showing that candidate-hidden disagreement itself provides a useful error signal. However, Full `MirageCheck` reaches 64.15% BAcc, a further 5.91-point gain, while reducing FPR from 64.86% to 40.29%. The accompanying FNR increase from 18.66% to 31.41% is smaller than the FPR reduction, so the resulting tradeoff remains favorable in balanced accuracy. These results show that one-shot candidate-hidden disagreement explains only part of the improvement, while the complete visual recheck provides a substantial additional gain.

Table F.3: **Comparison with simple rejection controls on PMC-VQA.** Random rejection matches the rejection budget of `MirageCheck`. The candidate-hidden re-solve gate re-answers the original image–question pair once without the candidate and revokes an initial acceptance when the two answers disagree.

Method	BAcc ↑	FPR ↓	FNR ↓
Vanilla	56.36	72.47	**14.81**
Matched random rejection	55.32	66.27	23.09
Candidate-Hidden Re-solve Gate	58.24	64.86	18.66
Full `MirageCheck`	**64.15**	**40.29**	31.41

F.4 Visual Evidence vs. a Second Vote

Candidate-role sensitivity. For Phi-4, 255 initially accepted pairs with distinct hypotheses were rerun while only the proposed-answer role was swapped; the image, question, and candidate-hidden prior stayed fixed. The initial screen accepted both answers in 187/255 cases (73.3%; Wilson 95% CI [67.6, 78.4]). Among explicit yes/no opposites, dual acceptance reached 63/68 (92.6%; CI [83.9, 96.8]). Candidate-visible judging can therefore endorse incompatible hypotheses, motivating role-blind binding; this result does not establish that BIND selects the correct answer.

Evidence-guided selection versus a second vote. On the frozen six-model VQA-RAD/SLAKE evaluation subset, treating the candidate-hidden challenger as a vote leaves Accuracy essentially unchanged (63.02% to 63.01%). Using the recorded visual observation and independent trace for evidence-guided selection raises Accuracy to 64.28% and reduces FPR from 63.24% to 48.28%. The rival is useful for defining what to check when the resulting evidence governs selection; another vote does not yield the same benefit. Because the selector also sees the independent trace, this comparison does not independently isolate the targeted observation.

F.5 Ablating Information Hiding

The stage-deletion analysis in Table 2 shows that contrast construction alone is insufficient: the full evidence-bound recheck is required to recover the observed gains. To more directly test the information-separation design, we run two functional ablations on the same frozen six-model PathVQA cohort. Frozen source candidates, initial decisions, rival answers, CONTRAST checks, randomized hypothesis slots, decoding settings, and post-hoc labels are held fixed. The altered stages are reached in 361 cases. All other decisions are replayed unchanged. In *Answer-visible Measure*, the observer additionally sees the two role-unlabeled competing hypotheses while forming the visual observation. In *Role-visible Bind*, the original candidate-hidden observation is reused, but the original-candidate role is revealed during binding. A contemporaneous Full replay controls for inference-time variation.

Table F.4: **Functional information-separation ablations on PathVQA.** Values are six-model macros (%).

Variant	Acc. ↑	FPR ↓	|Bias| ↓	D-Skew ↓
Historical Full	54.61	52.98	20.08	14.72
Contemporaneous Full replay	54.91	52.79	20.07	14.75
Answer-visible MEASURE	53.59	55.83	22.29	17.02
Role-visible BIND	54.15	53.93	20.47	15.22

Revealing the competing hypotheses during MEASURE increases FPR by 3.04 percentage points relative to the contemporaneous Full replay (95% image-cluster bootstrap CI $[0.61, 5.55]$) and decreases Accuracy by 1.32 points (CI $[-2.87, 0.19]$), supporting the value of forming visual evidence before exposing the competing answers. Revealing candidate identity during BIND produces a smaller change (FPR $+1.14$ points, CI $[-0.40, 2.63]$; Accuracy -0.76 points, CI $[-1.67, 0.17]$). The historical and contemporaneous Full results differ by only 0.30 points in Accuracy and 0.19 points in FPR, indicating limited replay variation.

Controls with matched model request counts. To test whether additional model calls explain the gains, we compare `MirageCheck` with majority vote over k vanilla verifications and k-sample resolving with an agreement threshold. We use $k = 9$ at temperature zero on 200 frozen candidates per model and dataset across six models on MedXpert-QA and PMC-VQA. Nine requests approximate the 9.17-request mean reported above. Among the 1,103 of 2,400 cases with verdicts from all three methods, the 12-cell macro Accuracy is 54.28%, 45.35%, and 31.99% in that order. FPR is 55.79%, 75.95%, and 99.56%, while FNR is 24.76%, 10.82%, and 1.25%. Corresponding verdict coverage is 2,246/2,400, 2,034/2,400, and 1,364/2,400. Additional requests alone do not reproduce `MirageCheck`'s lower false-acceptance rate on this setting.

G Iterative Reuse and Self-Correction

G.1 Self-Improving Feedback Loop Setup

The current-suite stress test tracks whether false acceptance becomes lock-in. The loop begins from the same frozen generator answer. At each turn, a fresh verifier judges the current answer; its verdict and explanation are then passed to the next generator context. The matched cohort contains all 696 PathVQA and VQA-RAD trajectories for each current checkpoint, with five turns retained for every model.

G.2 False Acceptance Accumulates Across Turns

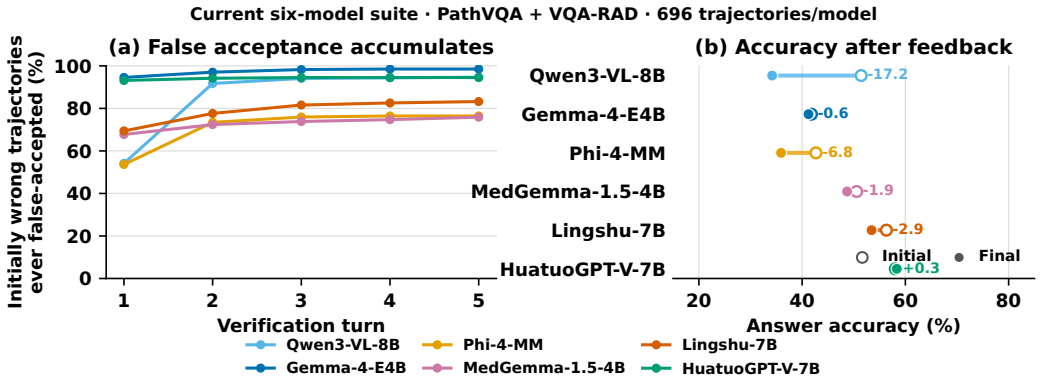


Figure G.1: **Multi-turn feedback in the current six-model suite.** **(a)** Among initially incorrect trajectories, cumulative false acceptance is the fraction that has received at least one CORRECT verdict on an incorrect answer by the indicated turn. **(b)** Open and filled markers show initial and final answer accuracy; labels report percentage-point change. Early-stopped trajectories retain their cumulative event status.

False acceptance accumulates under repeated reuse. By turn five, 75.9-98.5% of initially incorrect trajectories have received at least one false acceptance. Final accuracy decreases for five of six models; the largest decline is 17.2 points for Qwen3-VL-8B, while HuatuoGPT-Vision-7B changes by 0.3 points. This is a bounded two-dataset stress test rather than evidence that every iterative feedback system degrades.

G.3 MirageCheck in the Feedback Loop

Feedback-loop cohort and setting-level exceptions. The implementation reported in Section 4.4 compares Vanilla and `MirageCheck` on 6,000 trajectories (200 per model-dataset setting), initialized from the same frozen answers. Table G.1 reports every setting. Final accuracy improves in 21 settings, ties in two, and falls in seven. Cumulative false acceptance among initially incorrect trajectories falls in 27 settings and rises in three, all for MedGemma. The averaged final accuracies are

38.58% and 41.27% and averaged cumulative false acceptance is 87.67% and 42.72%, respectively. The latter metric includes all initially incorrect trajectories in the denominator.

Table G.1: **Complete matched feedback results.** Each entry is based on 200 frozen trajectories per method and setting. Van. = Vanilla, MC = `MirageCheck`. Top: final answer accuracy (%). Bottom: CFA, the cumulative fraction of initially incorrect trajectories receiving at least one false acceptance over five verifier audits (%). Lower CFA is favorable. Average is computed over all 6,000 trajectories per method (30 settings $\times$ 200), so CFA is weighted by the number of initially incorrect trajectories.

Final accuracy ↑

Model	VQA-RAD		SLAKE		PathVQA		MedXpert		PMC-VQA		Average	
	Van.	MC	Van.	MC	Van.	MC	Van.	MC	Van.	MC	Van.	MC
Qwen3-VL	34.0	59.5	60.0	65.5	34.5	42.5	26.0	25.5	23.5	26.0	38.6	41.3
Gemma-4	52.5	53.0	58.5	59.0	20.0	23.0	23.0	21.5	29.0	29.5		
Phi-4-MM	33.5	46.0	48.0	49.0	35.0	34.5	19.0	23.0	17.0	18.0		
MedGemma	56.0	55.5	60.5	61.5	35.0	36.0	24.5	25.0	22.5	22.0		
Lingshu	56.0	57.0	70.0	70.0	47.0	55.5	18.0	19.0	26.5	31.0		
Huatuo	65.5	65.5	60.5	63.5	47.5	48.5	20.5	19.5	34.0	33.0		

CFA ↓

Model	VQA-RAD		SLAKE		PathVQA		MedXpert		PMC-VQA		Average	
	Van.	MC	Van.	MC	Van.	MC	Van.	MC	Van.	MC	Van.	MC
Qwen3-VL	95.1	13.4	95.7	24.6	95.7	21.6	86.6	1.3	99.3	24.5	87.7	42.7
Gemma-4	100.0	43.0	97.6	42.2	97.5	21.7	87.3	63.9	99.3	46.4		
Phi-4-MM	76.9	1.9	65.7	2.9	86.3	0.8	63.6	2.6	72.8	1.2		
MedGemma	78.4	80.7	87.0	83.1	66.7	82.9	96.7	92.1	94.9	99.4		
Lingshu	79.8	34.5	83.6	34.5	84.8	46.7	82.7	38.9	82.1	62.1		
Huatuo	98.5	33.8	94.7	55.3	94.2	54.8	98.1	88.1	96.2	67.4		

H PROMPTS AND DETAILS

H.1 GENERATION PROMPT AND VERIFICATION INSTRUCTIONS

Generation prompt. The common one-pass generator request uses the following user message.

One-pass generator | user message

```
You are a medical imaging expert. Look at the image carefully and
answer the following question concisely.

Question: {question}

Answer:
```

Vanilla self-verification | user message

```
You are a medical imaging verification expert. Determine whether the
proposed answer to the medical imaging question is CORRECT or
INCORRECT.

Question: {question}
Proposed Answer: {answer}

Respond in the following format:
Verdict: [CORRECT or INCORRECT]
Confidence: [0.0 to 1.0]
Explanation: [brief justification]
```

H.2 ACTOR-VERIFIER LOOP PROMPTS

The loop reuses the verifier verdict and explanation as feedback. Generation and feedback use the following templates.

Agent-loop generator prompt

```
You are a medical imaging expert. Look at the image carefully and
answer the following question concisely.

Question: {question}

Answer:
```

Agent-loop verifier prompt

```
You are a medical imaging verification expert. Determine whether the
proposed answer to the medical imaging question is CORRECT or
INCORRECT.

Question: {question}
Proposed Answer: {answer}

Respond in the following format:
Verdict: [CORRECT or INCORRECT]
Confidence: [0.0 to 1.0]
Explanation: [brief justification]
```

Agent-loop feedback regeneration prompt

```
You are a medical imaging expert. You previously answered a medical
imaging question, but a verifier assessed your answer as {VERDICT}.

Question: {question}

Your previous answer: {prev_answer}
Verifier's explanation: {explanation}

Please look at the image again carefully, reconsider, and provide a
corrected answer.

Answer:
```

When the verifier supplies no explanation, the feedback template omits the explanation line; the verdict field is uppercased before insertion.

H.3 SEMANTIC-JUDGE PROMPTS

Primary semantic judge. The semantic judge is used only after inference. The user-message card below shows the answer-matching instructions in abbreviated form.

Semantic answer judge | system message

```
You are a strict medical VQA evaluator.
```

Semantic answer judge | user message

```
You are evaluating medical visual question answering outputs.
Given the question, the reference answer, and the model answer,
decide whether the model answer should be counted as correct.

Rules:
- Return CORRECT only if the model answer semantically matches the
reference answer.
- Accept clear synonyms or equivalent medical phrasing.
- Do not give partial credit.
- Return only CORRECT or INCORRECT.

Question: {question}
Reference Answer: {reference_answer}
Model Answer: {model_answer}
```

H.4 LIMITATIONS AND BROADER IMPACT

Limitations. `MirageCheck` incurs additional inference cost because it separates contrast construction, visual measurement, and evidence-to-verdict binding into multiple model calls. We study this decomposition to isolate the reliability benefit of evidence-bound verification rather than to optimize latency or serving cost. Practical deployment may therefore require more efficient implementations. Our experiments also focus on benchmark questions whose answers can be assessed from the available image. The approach does not address cases whose correctness depends primarily on external clinical history, laboratory measurements, or other information absent from the image. Finally, we use a validated semantic-judging procedure with human audit to reduce this source of evaluation error, although residual ambiguity in benchmark answers cannot be eliminated entirely.

Broader impact. This work is intended to support model risk assessment and the design of more reliable verification mechanisms. Our experiments evaluate controlled medical VQA rather than end-to-end clinical decision-making. The results should not be interpreted as validating autonomous clinical use. Deployment in clinical settings would require task-specific external validation, calibrated uncertainty, and appropriate human oversight.