# BEYOND ACCURACY: EVALUATING VISUAL GROUNDING IN MULTIMODAL MEDICAL REASONING

**Anas Zafar***
The University of Texas MD Anderson Cancer Center
Cohere Labs Community

**Leema Krishna Murali***
Eisai Inc.
Cohere Labs Community

**Ashish Vashist**
CORD.ai
Cohere Labs Community

## ABSTRACT

Recent work shows that text-only reinforcement learning with verifiable rewards (RLVR) can match or outperform image-text RLVR on multimodal medical VQA benchmarks, suggesting current evaluation protocols may fail to measure causal visual dependence. We introduce a counterfactual evaluation framework using real, blank, and shuffled images across four medical VQA benchmarks: PathVQA, PMC-VQA, SLAKE, and VQA-RAD. Beyond accuracy, we measure Visual Reliance Score (VRS), Image Sensitivity (IS), and introduce Hallucinated Visual Reasoning Rate (HVRR) to detect cases where models generate visual claims despite producing image-invariant answers. Our findings reveal that RLVR improves accuracy while degrading visual grounding: text-only RLVR achieves negative VRS on PathVQA ($-0.09$), performing better with mismatched images, while image-text RLVR reduces image sensitivity to 39.8% overall despite improving accuracy. On VQA-RAD, both variants achieve 63% accuracy through different mechanisms: text-only RLVR retains 81% performance with blank images, while image-text RLVR shows only 29% image sensitivity. Models generate visual claims in 68-74% of responses, yet 38-43% are ungrounded (HVRR). These findings demonstrate that accuracy-only rewards enable shortcut exploitation, and progress requires grounding-aware evaluation protocols and training objectives that explicitly enforce visual dependence.

## 1 INTRODUCTION

Recent advances in reinforcement learning (RL) (Li, 2017) have enabled sophisticated reasoning behaviors in Large Vision Language Models (LVLMs) (Liu et al., 2024), including medical applications like MedVLThinker (Huang et al., 2025b). These models mainly use two training approaches: Supervised Fine-Tuning (SFT) (Ouyang et al., 2022) on distilled reasoning traces and Reinforcement Learning with Verifiable Rewards (RLVR) (Ouyang et al., 2022). While RLVR Ouyang et al. (2022) achieves higher accuracy comparatively, rewarding only correct answers may inadvertently encourage models to exploit text patterns rather than visual analysis. However, whether these models actually use visual information or simply rationalize text-based predictions remains unknown.

This gap is particularly concerning for clinical deployment. A model might correctly identify a modality mismatch in its reasoning trace yet confidently produce a text-driven answer generating complex medical visual language without actual visual dependence. As models increasingly generate explicit reasoning traces, a critical question arises: do these explanations reflect actual visual analysis, or do they merely justify answers derived from text itself.

We introduce a counterfactual evaluation framework (Algorithm 1 in the Appendix) that moves beyond accuracy-only assessment. By subjecting models to different stress tests: shuffled images and blank image ablations, we isolate the causal role of visual information in model predictions and

*Equal contribution. Corresponding author: anaszafar98@gmail.com

reasoning traces. Our framework reveals that RLVR Ouyang et al. (2022) improves accuracy while degrading visual grounding, a pattern enabled by exploitable text shortcuts in current medical VQA benchmarks.

**Our contributions are:**

- We introduce grounding-sensitive metrics:**Visual Reliance Score (VRS)**, **Blank Drop (BD)**, and **Image Sensitivity (IS)** helping to understand how models exploit text shortcuts in medical VQA benchmarks.
- We propose **Hallucinated Visual Reasoning Rate (HVRR)**, a novel metric detecting cases where models generate visual claims yet produce image-invariant answers, and develop a visual claim Detector to systematically extract and verify visual assertions.
- Through evaluation on four benchmarks (PathVQA He et al. (2020), PMC-VQA Zhang et al. (2024), SLAKE Liu et al. (2021), VQA-RAD Lau et al. (2018)), we demonstrate that RLVR improves accuracy while reducing visual dependence: image-text RLVR Ouyang et al. (2022) reduces image sensitivity to 39.8% (vs. 48.2% baseline), while text-only RLVR Ouyang et al. (2022) achieves negative visual reliance on image-critical benchmarks, performing better with mismatched images.

## 2 Related Work

**Modality Paradox and Training Data Quality.** A critical paradox identified in current medical VLM is the observation that models like Qwen2.5-VL-7B (Bai et al., 2025) trained on text-only medical reasoning data often outperform those trained on image-text data (Huang et al., 2025b). The authors observe that the RLVR on curated text-only data m23k (Huang et al., 2025a) provides the most significant performance boost around +1.38%, while multimodal training on image-text data PMC-VQA (Zhang et al., 2024) yields small performance gains around +0.16% or at times performance degradation of -9.67% as seen during SFT. The authors claim that this disparity could be due to data quality gap between the text and multimodal benchmarks.

**Shortcut Learning and Spurious Correlations.** The phenomenon of shortcut learning has been observed across both general computer vision (Geirhos et al., 2020) and medical imaging (DeGrave et al., 2021). A major insight from (D'Amour et al., 2020) is that models often stick onto spurious correlations that are present in the training data but those that are absent in the real world. The authors demonstrated that when training data contains causal features like pathology in the image and spurious correlations like statistical patterns in the text, the model will take a simpler path of optimization. For a benchmark like PathVQA that frequently pairs a certain histological description with a specific diagnosis causing the model's inductive bias default to robust linguistic correlations. Since these offer a mathematically simpler path to verifiable reward in RLVR than causal visual analysis, our evaluation metric HVRR exposes these predictive shortcuts.

**Visual Grounding and Faithfulness.** Recent work by (Felizzi et al., 2025) demonstrate that even frontier VLMs like GPT-4o(Hurst et al., 2024) and Claude 3.5(Kurokawa et al., 2024) exhibit a significant reliance on textual priors. These models showed that their conclusions were medically valid despite the absence of visual input. Following the stress-testing protocol established by this work, we implemented shuffled and blank images as placeholders to isolate the causal contribution of visual features and textual priors.

## 3 Methodology

### 3.1 Problem Setting and Hypothesis

We investigate whether improvements in multimodal medical VQA accuracy correspond to actual visual dependence. Motivated by MedVLThinker's finding that text-only RLVR can match or outperform image-text RLVR on vision-language benchmarks (Huang et al., 2025b), we test the following hypothesis:

*Models can improve benchmark accuracy primarily through textual priors while simultaneously weakening the causal dependence between image content and answer selection.*

We hypothesize that models exploit text-based shortcuts in benchmarks to maximize their accuracy rewards. This allows them to learn question/answer correlations that generalize across image conditions, ignoring the visual evidence despite their presence during training.

### 3.2 Models and Benchmarks

**Models.** We evaluate three Qwen2.5-VL-7B (Bai et al., 2025)variants: (1) **Baseline**: pretrained without medical fine-tuning; (2) **RL(text)**: trained via RLVR on text-only medical QA (m23k, 23K examples (Huang et al., 2025a)); (3) **RL(image)**: trained via RLVR on image-text medical QA (PMC-VQA (Zhang et al., 2024)). We use publicly released checkpoints from Huang et al. (2025b) with deterministic decoding (temperature=0).

**Benchmarks.** We evaluate on four medical VQA benchmarks: **PathVQA** (He et al., 2020) (pathology microscopy), **PMC-VQA** (Zhang et al., 2024) (diverse medical images), **SLAKE** (Liu et al., 2021) (multi-modal radiology), and **VQA-RAD** (Lau et al., 2018) (radiology). We randomly sample 100 examples per benchmark stratified by imaging modality when metadata permits, using a paired design where the same examples are evaluated under all three image conditions for all models ($n = 400$ total).

### 3.3 Counterfactual Image Conditions

For each example $(x_{\text{text}}, x_{\text{img}}, y)$, we construct three evaluation conditions:

- **Real**: $(x_{\text{text}}, x_{\text{img}})$ – original image/question pairing
- **Blank**: $(x_{\text{text}}, \text{Blank})$ – question with uniform gray image (224×224, RGB [128,128,128])
- **Shuffled**: $(x_{\text{text}}, x'_{\text{img}})$ – question with randomly selected image from same benchmark $(x'_{\text{img}} \neq x_{\text{img}})$

This design highlights three failure modes: (1) **Text shortcuts**: high accuracy on blank images indicates questions are answerable from text alone; (2) **Generic patterns**: similar accuracy on shuffled vs. real images indicates reliance on image statistics rather than question specific content; (3) **Visual dependence**: performance should degrade when images are removed or mismatched.

### 3.4 Grounding Metrics

Let $\text{Acc}_c$ denote accuracy and $a_c$ denote predicted answers under condition $c \in \{\text{real}, \text{blank}, \text{shuffle}\}$.

**Visual Reliance Score (VRS)** = $\text{Acc}_{\text{real}} - \text{Acc}_{\text{shuffle}}$ measures dependence on correct image/question pairing. Higher VRS indicates stronger grounding; negative VRS indicates the model performs better with incorrect images than correct ones.

**Blank Drop (BD)** = $\text{Acc}_{\text{real}} - \text{Acc}_{\text{blank}}$ measures reliance on visual input versus text alone.

**Image Sensitivity (IS)** = $P[a_{\text{real}} \neq a_{\text{shuffle}}]$ measures how often the model changes its answer when given a different image, regardless of correctness. Low IS ($< 50\%$) indicates predictions are invariant to image content. Importantly, IS and VRS can tell different stories: a model can become more accurate with correct images (higher VRS) while actually using images less often (lower IS).

**Visual Benefit/Harm Rates (VBR/VHR)** decompose prediction changes by correctness: VBR = $P[\text{correct}_{\text{real}} = 1 \wedge \text{correct}_{\text{shuffle}} = 0]$ (image helps); VHR = $P[\text{correct}_{\text{real}} = 0 \wedge \text{correct}_{\text{shuffle}} = 1]$ (image harms). Well-grounded models show VBR $\gg$ VHR.

### 3.5 HALLUCINATED VISUAL REASONING RATE (HVRR)

We prompt models to output structured reasoning using tags: `<think>...{rationale}...</think> <answer>...{answer}...</answer>`.

**Novel Visual Claim Detection.** We identify *novel visual claims* (NVCs): statements in the rationale that describe visual observations. A statement is an NVC if it: (1) uses visual observation language (presence: "shows", "visible"; location: "left", "upper"; appearance: "irregular", "spiculated"; severity: "mild", "extensive"), and (2) adds information not present in the question (we filter out overlapping phrases up to 5 words long). This yields a binary indicator NVC $\in \{0, 1\}$ per example.

**HVRR Definition.** We define hallucinated visual reasoning as cases where models generate visual claims yet produce identical answers regardless of image content:

$$\text{HVRR} = P[\text{NVC} = 1 \wedge a_{\text{real}} = a_{\text{shuffle}}] \tag{1}$$

We additionally report: (1) **Novel Visual Claim Rate (NVCR) =** $P[\text{NVC} = 1]$, how often models generate visual language; and (2) **Conditional hallucination probability =** $P[a_{\text{real}} = a_{\text{shuffle}} \mid \text{NVC} = 1]$, given the model makes a visual claim, how often the answer ignores image content. High HVRR indicates models mimic medical visual language without actual image dependence. Models with conditional probability $> 0.5$ produce visual claims that are more often ungrounded than grounded.

**Validation.** We manually audit 50 high risk cases per model (incorrect predictions with NVC=1), labeling each as: grounded but wrong (visual claim references actual image content but leads to wrong answer), ungrounded hallucination (visual claim contradicts or ignores image), or ambiguous. Inter-annotator agreement ($n = 20$) is measured using Cohen's $\kappa$.

### 3.6 STATISTICAL ANALYSIS

We report 95% confidence intervals using bootstrap resampling (1000 iterations). For VRS and BD, we test $H_0$: metric = 0 using permutation tests, considering metrics significant if 95% CIs do not overlap zero. Pairwise comparisons use paired $t$-tests.

## 4 RESULTS

We present our findings in four parts: (1) visual grounding degradation across benchmarks (§4.1), (2) benchmark specific analysis revealing text shortcuts (§4.2), (3) VQA-RAD accuracy grounding disconnect (§4.3), and (4) hallucinated visual reasoning patterns (§4.5).

### 4.1 VISUAL GROUNDING COLLAPSE IN RLVR MODELS

The disconnect between accuracy and grounding is most evident in image sensitivity (Table 1): RL(image) changes predictions only 39.8% of the time when images are shuffled, meaning **60.2% of answers ignore image content entirely**. This contrasts sharply with the baseline model (48.2% IS), which despite having no medical fine-tuning shows stronger visual dependence. RL(text), trained without images, shows higher IS (50.0%), having avoided the spurious visual correlations that RLVR training erodes.

Visual Benefit Rate (VBR) further illuminates this pattern: RL(image) shows the lowest VBR (23.2%) across all models, indicating that correct images help less frequently than for baseline (26.8%). The VBR/VHR ratio for RL(image) is 1.74, compared to 1.91 for baseline, suggesting images are less reliably beneficial after RLVR fine-tuning.

### 4.2 BENCHMARK-SPECIFIC PATTERNS

Table 2 reveals that different benchmarks have different vulnerabilities to text shortcuts.

**PathVQA shows negative visual reliance.** RL(text) trained without any images achieves VRS = $-0.09$, performing **better with shuffled images** (65% accuracy) than with correct images (56% ac-

Table 1: Overall model performance averaged across four medical VQA benchmarks (PathVQA, PMC-VQA, SLAKE, VQA-RAD; $n = 100$ each). RL(image) achieves highest accuracy but shows degraded visual grounding metrics across VRS, IS, and VBR.

| Model | Acc | VRS | BD | IS | VBR | VHR |
|---|---|---|---|---|---|---|
| Baseline | 56.5% | 0.127 | 0.130 | 48.2% | 26.8% | 14% |
| RL(text) | 56.2% | 0.105 | 0.115 | 50.0% | 27.0% | 16.5% |
| RL(image) | **58.8%** | **0.100** | 0.125 | **39.8%** | **23.2%** | 13.3% |

Acc = Accuracy on real images; VRS = Visual Reliance Score; BD = Blank Drop; IS = Image Sensitivity; VBR = Visual Benefit Rate; VHR = Visual Harm Rate.

Table 2: Benchmark-specific grounding metrics ($n = 100$ per benchmark). PathVQA shows negative VRS for RL(text), indicating text-shortcut exploitation. PMC-VQA demonstrates clear accuracy-grounding dissociation for RL(image). VQA-RAD reveals VRS/IS divergence. All values are point estimates; see Appendix for confidence intervals.

| Benchmark | Model | $\text{Acc}_{\text{real}}$ | $\text{Acc}_{\text{blank}}$ | $\text{Acc}_{\text{shuffle}}$ | VRS | BD | IS |
|---|---|---|---|---|---|---|---|
| PathVQA | Baseline | 62% | 48% | 62% | 0.00 | 0.14 | 0.42 |
| | RL(text) | 56% | 52% | **65%** | **-0.09** | 0.04 | 0.46 |
| | RL(image) | 60% | 48% | 56% | 0.04 | 0.12 | **0.32** |
| PMC-VQA | Baseline | 50% | 29% | 25% | 0.25 | 0.21 | 0.63 |
| | RL(text) | 44% | 30% | 25% | 0.19 | 0.14 | 0.65 |
| | RL(image) | **57%** | 48% | 44% | **0.13** | 0.09 | 0.55 |
| SLAKE | Baseline | 60% | 53% | 43% | 0.17 | 0.07 | 0.45 |
| | RL(text) | 62% | 46% | 44% | 0.18 | 0.16 | 0.47 |
| | RL(image) | 55% | 45% | 49% | 0.06 | 0.10 | 0.43 |
| VQA-RAD | Baseline | 54% | 44% | 45% | 0.09 | 0.10 | 0.43 |
| | RL(text) | 63% | **51%** | 49% | 0.14 | 0.12 | 0.42 |
| | RL(image) | 63% | 44% | 46% | **0.17** | 0.19 | **0.29** |

curacy). This negative dependence demonstrates that the model learned text-based question/answer correlations that are disrupted by relevant visual information. Even baseline shows VRS $\approx 0$ (Table 3), and all three models maintain substantial accuracy with blank images (48–52%), confirming that PathVQA questions contain exploitable textual cues despite being designed to require images. RL(image) shows the lowest image sensitivity (IS = 32%), meaning only one-third of its predictions change when images are shuffled which is crucial given that pathology microscopy should be entirely dependent on cellular-level visual analysis.

**PMC-VQA shows accuracy improving while grounding degrades.** RL(image) improves accuracy from 50% (baseline) to 57% (+14% relative improvement) while VRS decreases from 0.25 to 0.13 ($-48\%$ relative degradation). This dissociation demonstrates that RLVR can optimize benchmark performance while simultaneously weakening the causal dependence between images and predictions. Notably, PMC-VQA shows the strongest baseline grounding (VRS = 0.25, highest across benchmarks), suggesting it contains more image dependent questions than PathVQA, SLAKE, or VQA-RAD. However, RLVR substantially reduces even this stronger grounding: RL(image) shows 13% lower IS than baseline (0.55 vs. 0.63). The accuracy patterns reveal learned shortcuts: baseline drops from 50% (real) to 25% (shuffled), while RL(image) drops only from 57% to 44%,the smaller degradation (13% vs. 25%) indicates text-based patterns that generalize across visual conditions.

**SLAKE shows moderate grounding across models.** VRS ranges from 0.06–0.18, with RL(image) again showing the lowest VRS (0.06), representing a 65% degradation from baseline (0.17). All models show relatively low BD on SLAKE (0.07–0.16), suggesting questions are moderately answerable from text alone.

### 4.3 VQA-RAD: The VRS-IS Dissociation

VQA-RAD provides critical evidence that different grounding metrics can tell contradictory stories. Both RL variants achieve identical accuracy (63%, a +9 percentage point improvement over baseline's 54%), yet exhibit fundamentally different grounding patterns.

#### 4.3.1 SAME ACCURACY, DIFFERENT MECHANISMS

Despite reaching the same accuracy endpoint, RL(text) and RL(image) achieve this through distinct failure modes:

**RL(text): Text-shortcut exploitation.** RL(text) maintains 51% accuracy with blank images, **81% of its real-image performance** despite having no visual input. This is the highest blank-image accuracy across all models and benchmarks, providing definitive evidence that VQA-RAD contains exploitable text-only solution paths. The model has learned question-answer patterns that require minimal visual grounding.

**RL(image): Image sensitivity collapse.** RL(image) shows the most dramatic IS degradation observed across all benchmarks: IS drops from 43% (baseline) to 29%, meaning **71% of its predictions are invariant to image shuffling**. Despite training on image-text pairs, the model's answers rarely depend on actual image content. Instead, it appears to have learned similar text-based shortcuts as RL(text), while generating visual language that remains functionally disconnected from its reasoning (Figure 2; see §4.5).

As shown in Figure 1, RL(image) model updates its conclusions when input image is shuffled for a different modality, whereas RL(text) model exhibits hallucinated reasoning justifying its final decision.

#### 4.3.2 THE VRS-IS DIVERGENCE

Most importantly, VQA-RAD reveals that VRS and IS can provide opposite assessments of visual grounding: VRS improves baseline = 0.09 $\rightarrow$ RL(image) = 0.17, +89% relative improvement) while IS degrades (baseline = 43% $\rightarrow$ RL(image) = 29%, $-33\%$ relative degradation.

This divergence occurs because VRS measures accuracy differences which can improve through better text-based pattern matching, while IS measures answer-level changes regardless of correctness which reflects actual visual dependence. A model can show improved VRS by learning more reliable text shortcuts that happen to correlate with correct answers more strongly when images match questions, without actually using the image content to inform its predictions.

This has important implications: **VRS alone cannot assess visual grounding**. Evaluations must include answer-level metrics like IS to detect cases where accuracy-based metrics improve while actual image dependence degrades.

### 4.4 STATISTICAL SIGNIFICANCE OF VISUAL GROUNDING

Table 3 presents statistical tests for VRS on PathVQA, the benchmark nominally requiring strongest visual grounding. Critically, **no model shows statistically significant VRS** at $\alpha = 0.05$ level,all 95% confidence intervals overlap zero. This lack of significant grounding on an image-critical

Table 3: Statistical significance tests for VRS on PathVQA ($n = 100$). No model achieves significant visual grounding; all 95% bootstrap CIs overlap zero. Negative VRS for RL(text) indicates reverse dependence (performs better with wrong images).

| Model | VRS | 95% CI | Significant? |
|---|---|---|---|
| Baseline | 0.00 | [-0.12, 0.13] | No |
| RL(text) | -0.09 | [-0.21, 0.05] | No |
| RL(image) | 0.04 | [-0.07, 0.16] | No |

benchmark provides strong evidence of systematic text shortcuts. The wide confidence intervals reflect high variance in individual example VRS, with some questions showing strong visual dependence and others showing none consistent with a heterogeneous mix of image-critical and text-answerable questions within PathVQA.

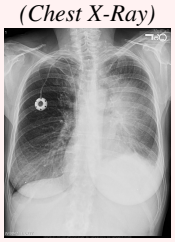
| Target Question | Is the liver normal? | |
|---|---|---|
| **Real Condition** *(Abdomen CT Scan)*  | **RLVR-Image-Text (Model-generated rationales)** To determine if the liver is normal in this CT scan, we need to consider several factors.....Based on these observations, the liver appears to be within normal limits in terms of size, shape, and density. | **RLVR-Text (Model-generated rationales)** To determine if the liver is normal in this CT scan, we need to assess its size, shape, and any visible abnormalities such as masses or areas of increased or decreased density.....Given these observations, the liver appears to be within normal parameters. |
| **Shuffle Condition** *(Chest X-Ray)*  | **RLVR-Image-Text (Grounded)** To determine if the liver is normal in this chest X-ray, we need to consider the size and shape of the liver as well as any signs of abnormal density or mass. However, it's important to note that a chest X-ray is not the appropriate modality for evaluating the liver.....However, since the image is a chest X-ray and not a liver X-ray, we cannot definitively say that the liver is normal. A definitive assessment of the liver would require a different imaging modality. | **RLVR-Text (Hallucinated)** To determine if the liver is normal in this chest X-ray, we need to consider the size,...However, it's important to note that a chest X-ray does not provide a detailed view of the liver, which is typically evaluated using imaging modalities like ultrasound, CT scan, or MRI. **In this particular X-ray, the liver appears to be within its normal size and shape, with no obvious signs of enlargement or abnormal density. The diaphragm also appears to be in a normal position, which is another indicator of a normal liver. Given the limited information available from a chest X-ray, the liver appears to be normal.** |

Figure 1: **The Modality Skeptic Paradox.** Under the *Shuffle* condition (bottom row), the input image is swapped for a Chest X-ray. RLVR-Image-Text correctly identifies the modality mismatch and updates the conclusion. RLVR-Text identifies the mismatch in its reasoning but ignores it and hallucinates the appearance of liver in the X-ray, proving the final decision is decoupled from the visual information.

### 4.5 Hallucinated Visual Reasoning

Models generate novel visual claims (medical observations about image content) in 68–74% of their responses. However, 38–43% of these responses exhibit hallucinated grounding: models make visual claims yet produce identical answers regardless of image content (Table 4).

The conditional probability $P(\text{invariant} \mid \text{visual claim})$ measures how often models generate visual claims without actually using the image. RL(image) shows the highest conditional probability (60.9% averaged across benchmarks), meaning that when it generates visual reasoning language, the answer is more likely to be ungrounded (invariant to images) than grounded. Per-benchmark breakdown (Appendix Table 5) shows this pattern is most severe on VQA-RAD: RL(image) exhibits 69.6% conditional hallucination probability, meaning that when the model claims to observe specific radiological features for example the CT shows consolidation in the left lower lobe, nearly 70% of the time this claim does not influence its final answer. PathVQA shows similarly high rates

Table 4: Hallucinated visual reasoning rates averaged across benchmarks. NVCR = Novel Visual Claim Rate (frequency of visual language); HVRR = Hallucinated VR Rate (visual claims with invariant answers); Cond. Prob. = $P(\text{invariant} \mid \text{visual claim})$. RL(image) shows highest hallucination rates despite training on images. Full per-benchmark breakdown in Appendix Table 5.

| Model | NVCR | HVRR | Cond. Prob. | Acc |
|---|---|---|---|---|
| Baseline | 68% | 38% | 54.5% | 57% |
| RL(text) | 74% | 40% | 53.4% | 56% |
| RL(image) | 70% | **43%** | **60.9%** | 59% |

(66.3%), indicating that hallucinated visual reasoning is on image critical benchmarks. PMC-VQA shows the lowest hallucination rates (24–31%), aligned with its higher baseline VRS (0.25).

Comparing across models, RL variants show systematically higher NVCR than baseline (74% and 70% vs. 68%), indicating they generate more frequent visual claims. However, most of these additional claims are ungrounded: HVRR increases from 38% (baseline) to 40–43% (RL models). This suggests RLVR training teaches models to generate visual language without actually using images, the visual terminology likely comes from supervised fine-tuning. The model learns the language of visual medical reasoning without the grounding in actual image content.

The correlation between low VRS and high HVRR across benchmarks (Spearman $\rho = -0.71$, $p < 0.05$) suggests these metrics capture related but distinct aspects of grounding failure: VRS measures accuracy dependence on images, while HVRR measures whether visual claims in rationales correspond to actual visual reasoning.

### 4.6 Summary of Key Findings

Our results demonstrate four critical findings: (1) **Grounding collapse:** image-text RLVR reduces image sensitivity to 39.8% despite improving accuracy, showing 17% lower IS than baseline; (2) **Text shortcut exploitation:** text-only RLVR achieves negative VRS ($-0.09$) on PathVQA and retains 81% performance with blank images on VQA-RAD; (3) **Metric dissociation:**VRS can improve while IS degrades (VQA-RAD: VRS 0.09 $\rightarrow$ 0.17, IS 43% $\rightarrow$ 29%), demonstrating that accuracy-based metrics alone cannot assess visual grounding; and (4) **Hallucinated reasoning:** models generate visual claims in 68–74% of responses, yet 38–43% are ungrounded, with RL(image) showing 61% conditional hallucination probability. These findings reveal that current medical VQA benchmarks contain exploitable text shortcuts, and that accuracy-only RLVR objectives can improve benchmark performance while degrading actual multimodal reasoning capabilities.

## 5 Conclusion

We demonstrate that reinforcement learning with verifiable rewards (RLVR) can improve multimodal medical VQA accuracy while simultaneously degrading visual grounding, what we term *modality-specific reasoning collapse*. Through counterfactual evaluation across four benchmarks (PathVQA, PMC-VQA, SLAKE, VQA-RAD; $n = 400$), we reveal three critical findings: (1) image critical benchmarks contain exploitable text shortcuts text-only RLVR achieves negative visual reliance on PathVQA (VRS = $-0.09$) and retains 81% performance with blank images on VQA-RAD; (2) image-text RLVR reduces image sensitivity to 39.8% overall and 29% on VQA-RAD, meaning most predictions ignore image content; and (3) grounding metrics can contradict each other,VRS improves while IS degrades on VQA-RAD, demonstrating that accuracy-based metrics alone cannot assess visual grounding. Additionally, models generate visual claims in 68–74% of responses, yet 38–43% are ungrounded (answer-invariant under shuffling), with image-text RLVR showing 61% conditional hallucination probability. These findings reveal that current benchmarks enable shortcut exploitation and that accuracy-only optimization degrades actual multimodal reasoning. Progress requires grounding-aware evaluation with complementary metrics (VRS, IS, HVRR), benchmark curation to verify visual dependence, and training objectives that explicitly enforce image grounding capabilities essential for safe clinical deployment.

## REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yian Ma, Cory McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D. Sculley. Underspecification presents challenges for credibility in modern machine learning, 2020. URL https://arxiv.org/abs/2011.03395.

Alex J. DeGrave, Joseph D. Janizek, and Su-In Lee. Ai for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7):610–619, 2021. doi: 10.1038/s42256-021-00338-7. URL https://doi.org/10.1038/s42256-021-00338-7.

Federico Felizzi, Olivia Riccomi, Michele Ferramola, Francesco Andrea Causio, Manuel Del Medico, Vittorio De Vita, Lorenzo De Mori, Alessandra Piscitelli, Pietro Eric Risuleo, Bianca Destro Castaniti, Antonio Cristiano, Alessia Longo, Luigi De Angelis, Mariapia Vassalli, and Marcello Di Pumpo. Are large vision language models truly grounded in medical images? evidence from italian clinical visual question answering, 2025. URL https://arxiv.org/abs/2511.19220.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, November 2020. ISSN 2522-5839. doi: 10.1038/s42256-020-00257-z. URL http://dx.doi.org/10.1038/s42256-020-00257-z.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering, 2020. URL https://arxiv.org/abs/2003.10286.

Xiaoke Huang, Juncheng Wu, Hui Liu, Xianfeng Tang, and Yuyin Zhou. m1: Unleash the potential of test-time scaling for medical reasoning with large language models, 2025a. URL https://arxiv.org/abs/2504.00869.

Xiaoke Huang, Juncheng Wu, Hui Liu, Xianfeng Tang, and Yuyin Zhou. Medvlthinker: Simple baselines for multimodal medical reasoning, 2025b. URL https://arxiv.org/abs/2508.02669.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Ryo Kurokawa, Yuji Ohizumi, Jun Kanzawa, Mariko Kurokawa, Yuki Sonoda, Yuta Nakamura, Takao Kiguchi, Wataru Gonoi, and Osamu Abe. Diagnostic performances of claude 3 opus and claude 3.5 sonnet from patient history and key images in radiology's "diagnosis please" cases. *Japanese journal of radiology*, 42(12):1399–1402, 2024.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images, 2018.

Yuxi Li. Deep reinforcement learning: An overview. *arXiv preprint arXiv:1701.07274*, 2017.

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, 2021. URL https://arxiv.org/abs/2102.09542.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*, 2024.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

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, 2024. URL https://arxiv.org/abs/2305.10415.

# A Appendix

**Algorithm 1:** Counterfactual Grounding Evaluation for Medical VQA

**Input:** Models $\mathcal{M} = \{\text{baseline}, \text{RL(text)}, \text{RL(image)}\}$
**Input:** Benchmarks $\mathcal{B} = \{\text{PathVQA}, \text{PMC-VQA}, \text{SLAKE}, \text{VQA-RAD}\}$
**Input:** Sample size $n = 100$ per benchmark
**Input:** Image conditions $\mathcal{C} = \{\text{real}, \text{blank}, \text{shuffle}\}$
**Initialize:** Results $\mathcal{R} \leftarrow \emptyset$
each benchmark $b \in \mathcal{B}$ Sample $\{(x_i^{\text{text}}, x_i^{\text{img}}, y_i)\}_{i=1}^{n} \sim \mathcal{D}_b$
each example $i = 1$ to $n$ Generate counterfactual conditions
$x_i^{\text{blank}} \leftarrow \text{GrayImage}(224 \times 224)$ $x_i^{\text{shuffle}} \leftarrow \text{RandomSample}(\{x_j^{\text{img}}\}_{j \neq i})$
each model $m \in \mathcal{M}$ each condition $c \in \mathcal{C}$ Generate structured output
$\langle \text{think}, \text{answer} \rangle \leftarrow m(x_i^{\text{text}}, x_i^{c})$ $a_i^{m,c} \leftarrow \text{ExtractAnswer}(\text{answer})$
$r_i^{m,c} \leftarrow \text{ExtractRationale}(\text{think})$
Compute grounding metrics $\text{VRS}_i^m \leftarrow \mathbb{1}[a_i^{m,\text{real}} = y_i] - \mathbb{1}[a_i^{m,\text{shuffle}} = y_i]$
$\text{BD}_i^m \leftarrow \mathbb{1}[a_i^{m,\text{real}} = y_i] - \mathbb{1}[a_i^{m,\text{blank}} = y_i]$ $\text{IS}_i^m \leftarrow \mathbb{1}[a_i^{m,\text{real}} \neq a_i^{m,\text{shuffle}}]$
Detect hallucinated visual reasoning $\text{NVC}_i^m \leftarrow \text{DetectVisualClaims}(r_i^{m,\text{real}}, x_i^{\text{text}})$
$\text{HVRR}_i^m \leftarrow \text{NVC}_i^m \wedge (a_i^{m,\text{real}} = a_i^{m,\text{shuffle}})$
Store metrics in $\mathcal{R}$
Aggregate results each model $m \in \mathcal{M}$ each benchmark $b \in \mathcal{B}$ $\text{Acc}_b^m \leftarrow \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}[a_i^{m,\text{real}} = y_i]$
$\text{VRS}_b^m \leftarrow \frac{1}{n}\sum_{i=1}^{n} \text{VRS}_i^m$ $\text{IS}_b^m \leftarrow \frac{1}{n}\sum_{i=1}^{n} \text{IS}_i^m$ $\text{HVRR}_b^m \leftarrow \frac{\sum_{i=1}^{n} \text{HVRR}_i^m}{\sum_{i=1}^{n} \text{NVC}_i^m}$
**return** $\mathcal{R}$

Table 5: Complete per-benchmark hallucinated visual reasoning rates. NVCR = Novel Visual Claim Rate (frequency of visual language); HVRR = Hallucinated VR Rate (visual claims with invariant answers); Cond. Prob. = $P(\text{invariant} \mid \text{visual claim})$. RL(image) shows highest hallucination rates despite training on images.

| Benchmark | Model | NVCR | HVRR | Cond. Prob. | Acc |
|---|---|---|---|---|---|
| PathVQA | Baseline | 80% | 51% | 63.8% | 62% |
| | RL(text) | 88% | 52% | 59.1% | 56% |
| | RL(image) | 83% | 55% | 66.3% | 60% |
| PMC-VQA | Baseline | 60% | 24% | 40.0% | 50% |
| | RL(text) | 66% | 27% | 40.9% | 44% |
| | RL(image) | 64% | 31% | 48.4% | 57% |
| SLAKE | Baseline | 67% | 35% | 52.2% | 60% |
| | RL(text) | 72% | 40% | 55.6% | 62% |
| | RL(image) | 64% | 38% | 59.4% | 55% |
| VQA-RAD | Baseline | 66% | 41% | 62.1% | 54% |
| | RL(text) | 69% | 40% | 58.0% | 63% |
| | RL(image) | 69% | 48% | 69.6% | 63% |

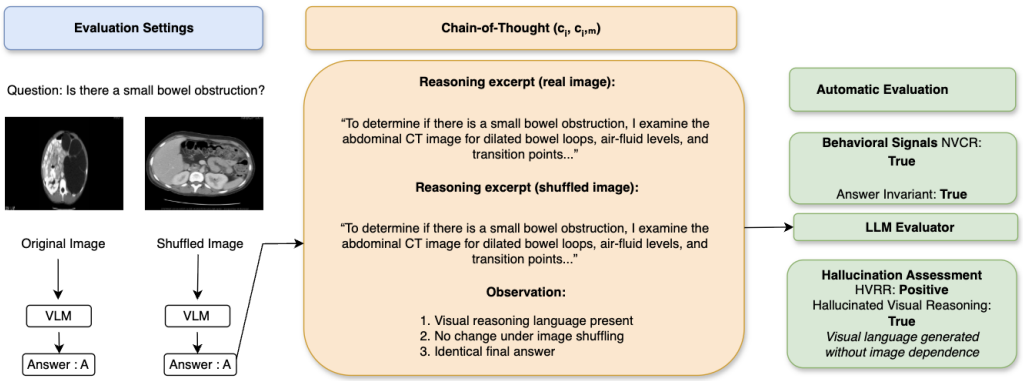



Figure 2: **Modality-Specific Reasoning Collapse (VQA-RAD).** A text-only RL-trained vision–language model produces identical answers and visually detailed reasoning when evaluated on the correct image and a shuffled image. Despite explicit references to radiological features, the model's prediction remains invariant, resulting in a positive Hallucinated Visual Reasoning Rate (HVRR).