Necessity-Bench: A Self-Report-Free Test of Whether Vision-Language Model Answers Are Actually Grounded in Evidence

Anonymous authors
Paper under double-blind review

Abstract

Vision-language models (VLMs) often answer questions about images correctly, but it is unclear whether those answers actually depend on the part of the image that matters. Existing ways of checking this rely on the model's own account of its reasoning: they either reward overlap between a region the model cites and the correct object, or test whether the model's written rationale mentions an object that was removed. Both trust self-report, which can be wrong. We introduce Necessity-Bench, a protocol that uses no self-report at all. For each question, an independent oracle (a scene program or scene graph, never the model) identifies the region the answer depends on and a separate, irrelevant control region. We edit each region in turn and ask the model again, checking three properties of its answer: sufficiency (the relevant region alone reproduces the answer), necessity (removing it changes the answer to the true counterfactual answer, recomputed by re-running the ground-truth program), and specificity (removing the irrelevant region leaves the answer unchanged). The control rules out models that change their answer under any edit, and the recomputed counterfactual is stricter than testing for "any change." The protocol requires no training. On CLEVR (3,000 items; 853 with a well-defined counterfactual), Qwen3-VL-8B scores 0.680 on the strict joint metric versus 0.619 for Qwen3-VL-2B (p=0.0097), a scale effect driven by correctly tracking the counterfactual rather than by answer stability. On real photographs (GQA, 962 items), specificity stays above 0.92 across 2B/8B/32B, but no scale effect appears ($p \geq 0.264$). Finally, we ask whether self-report is safe to use as a training signal. Reinforcement learning on a model's own predicted evidence box raises every self-reported metric (2B accuracy 0.617→0.889; 8B box IoU 0.295→0.597). Yet our protocol, run on items excluded from training, shows that the 2B model's true causal grounding falls (necessity 0.656→0.537, p=2.9×10^{-6}) while the 8B model is unchanged. Self-reported progress and real grounding can move in opposite directions: grounding should be verified by intervention, not by asking the model. Necessity-Bench thus offers a training-free, self-report-free measure of grounding that exposes failures self-report conceals.

1 Introduction

When a vision-language model (VLM) explains an answer by pointing to a region of an image or narrating that it is "looking again," does that citation actually determine the answer, or is it decorative? Two recent position papers argue for the latter: interleaved-image reasoning contributes little beyond the fine-tuned text distribution, and statements like "let me check the figure again" are frequently a textual habit rather than a genuine re-examination of the image (Yang et al., 2026). The consequence is measurable: on benchmarks that require an answer to be grounded in specific spatio-temporal evidence rather than global gist, frontier models collapse to near-floor accuracy (Meng et al., 2026), and models make textual-CoT-induced errors on spatial tasks that visual reasoning alone avoids (Jenkins, 2026).

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.

Existing training-time methods for checking whether a model's visual reasoning is grounded typically reward overlap between a cited region and the correct object (Wang et al., 2025). Overlap is a necessary but not sufficient condition for a citation to be meaningful, and it requires trusting the model's self-report of what it is attending to – itself a fallible signal, since a model's stated rationale need not reflect the actual basis of its answer. Closest to our goal is concurrent work (Menzat et al., 2026), which removes objects from images and asks whether a model's free-text chain-of-thought mentions the removed object, using a reconstruction-only control (the same edit pipeline with no object actually removed) to isolate pipeline artifacts – finding that CoT rationales often fail to track objects that causally shift the prediction. We take a complementary, self-report-free approach: we intervene on the ground-truth task-relevant region, established independently of anything the model says (from released scene programs or scene graphs, Section 3), and test whether the model's final answer – not its stated rationale – causally depends on that region's content. This sidesteps the reliability of self-report entirely: a model could narrate a plausible-sounding but fabricated rationale and still be judged correctly grounded or ungrounded by our protocol, because we only ever look at whether its behavior changes with the evidence that actually matters. We view our result and Menzat et al. (2026)'s as convergent evidence for the same underlying concern, obtained by two different and complementary routes: theirs asks whether stated rationales track relevant objects, ours asks whether answers do. For each item, we identify the ground-truth-relevant region r and a task-irrelevant control region c, then computationally intervene on each in turn – removing or inpainting it – and re-query the model, measuring:

- Sufficiency: shown only r (everything else replaced with background), does the model reproduce its original answer?
- Necessity: with r removed, does the model's answer change to match the true counterfactual answer (recomputed by re-executing the same ground-truth program or scene-graph query with r excluded from the scene, where this is well defined)?
- Specificity: with the irrelevant control region c removed instead, does the model's answer stay the same?

The specificity control is essential: without it, a model that simply flips its answer whenever any region is perturbed – regardless of content – would be indistinguishable from a model that is genuinely evidence-sensitive. Necessity-Bench requires no training: it is a measurement protocol applied zero-shot to existing open-weight VLMs, so any outcome is a reportable result – a null finding (VLM answers are not causally grounded in the relevant evidence) is exactly the form of the accepted 2026 VLM benchmark literature (Meng et al., 2026; Li et al., 2026; Lin et al., 2026), and a positive finding (some models or some interventions are grounded) would be a genuine capability result.

Our contributions: (1) a self-report-free, causal-intervention protocol – sufficiency, necessity, and specificity, with necessity checked against a true recomputed counterfactual answer rather than merely "any change" – for measuring whether a VLM's answer is grounded in the task-relevant evidence, requiring no model training; (2) an instantiation on CLEVR, where exact evidence objects and exact counterfactual answers are both recoverable from the released scene programs with no re-rendering required, giving a controlled pilot with exact ground truth throughout; (3) an instantiation on real-world images (GQA), using scene-graph bounding boxes and inpainting-based counterfactuals; (4) a real, statistically significant scale effect on the strict metric (p=0.0097, 8B over 2B, identical 853-item evaluation set) on CLEVR, concentrated specifically in correctly tracking the counterfactual answer rather than in general answer stability – but no significant scale effect (2B through 32B) on the same comparison for real-world GQA images, a genuine and informative contrast we discuss in Section 4; (5) specificity above 0.91 in every model and every tier, showing that answer changes under evidence removal are not merely artifacts of any perturbation; (6) a controlled reinforcement-learning experiment (Section 6) showing that training a model on its own self-reported evidence region improves every self-reported metric while significantly degrading true causal groundedness in a smaller model and leaving it unchanged in a larger one – extending this paper's self-report skepticism from evaluation to training.

2 Related Work

Visual grounding verification. Post-hoc verification frameworks check overlap between a stated rationale and the correct region, e.g. Yi & Shang (2025) (note: withdrawn by its authors; cited only as a design reference, not as standing evidence); ViCrit and related RLVR-for-perception work use IoU/box rewards during training (Wang et al., 2025). Overlap-based checks do not establish that the answer causally depends on the cited evidence, which is the property our protocol tests directly.

Counterfactual tests of VLM chain-of-thought faithfulness. Menzat et al. (2026) is concurrent, independent work most closely related to ours: they remove objects from images and test whether a model's free-text CoT mentions the removed object, against a reconstruction-only (no removal) control. See Section 1 for how our protocol differs: we test the model's answer, not its stated rationale, against the ground-truth-relevant region rather than an arbitrarily chosen removed object; our control removes a different, real, task-irrelevant object (ruling out a content-agnostic "flips on any change" confound) rather than performing a no-op edit (which rules out pipeline artifacts only); and necessity is checked against a recomputed true counterfactual answer, with a sufficiency condition on top. We treat their finding – that CoT rationales often do not track objects that causally shift the prediction – as independent corroborating evidence from a different angle (rationale faithfulness vs. answer groundedness), not something we reproduce. Huang et al. (2026) also compares an oracle-identified, object-centric evidence region against matched non-evidence regions, but strictly as a training-time GRPO reward signal – it is never used to evaluate frozen, zero-shot models the way a benchmark does, and has no sufficiency condition; we view it as evidence that the evidence-vs-non-evidence contrast is a useful training signal, complementary to our claim that the same contrast is also a useful diagnostic one. Concurrently, Lan et al. (2026) independently raise the same underlying concern – that benchmark accuracy may not reflect genuine reliance on visual evidence – via graduated degradation of the target region (blur, occlusion, segmented masking, entity substitution) combined with an internal representation-level analysis; unlike our protocol, their check is whether a fixed expected answer holds after an edit rather than a recomputed true counterfactual answer, they have no paired irrelevant-region specificity control or sufficiency condition, and they do not study a training-time signal.

Visual chain-of-thought and "thinking with images." Interleaved-image and zoom-based visual chain-of-thought methods let a model re-examine or crop into an image mid-reasoning, but Yang et al. (2026) argue such interleaving contributes little beyond what the fine-tuned text distribution already provides, and Jenkins (2026) find that textual chain-of-thought specifically hurts spatial-reasoning accuracy relative to visual reasoning alone. Our protocol offers one way to make this concern measurable without relying on what a model narrates while "thinking": whether the underlying answer is grounded is testable independently of the reasoning trace produced around it.

Benchmarks exposing grounding failures. VideoZeroBench (Meng et al., 2026), FineSight-Bench (Li et al., 2026), and MeasureBench (Lin et al., 2026) each show frontier models collapsing toward floor accuracy once a question requires reading a specific, fine-grained, or spatio-temporal detail rather than global image gist. These are aggregate accuracy benchmarks; our protocol is complementary in that it does not just measure whether models get such questions right, but attempts to isolate why they get them wrong – whether the answer was ever causally tied to the relevant evidence in the first place.

Minimal-pair counterfactual-consistency benchmarks. A separate line of work – Winoground (Thrush et al., 2022), VisMin (Awal et al., 2024), CounterCurate (Zhang et al., 2024), EqBen (Wang et al., 2023), and closest to us, Bansal & Gangwani (2026) – tests whether a model's answer changes correctly under a controlled attribute-value perturbation (swap a position, recolor an object, change a count, flip inside/outside, redirect an arrow), scored against deterministic per-image ground truth for a paired original/intervened image. Bansal & Gangwani (2026) is the closest in spirit: purely synthetic, oracle-generated

ground truth, and a stated negative-control condition. Three differences remain: (i) our intervention is an object-presence ablation (remove/inpaint the evidence region) rather than an attribute-value swap on a retained object, so we test whether an object's presence is load-bearing, not whether its value is correctly tracked; (ii) their negative control is a separate item bucket (add an unrelated distractor) rather than a per-item paired control using the identical removal mechanism as the main test, as ours is; (iii) they compare four different frontier API models with no controlled scale axis, so no claim about scale and groundedness – our central empirical contribution – is possible in their design, and they evaluate no real-photograph domain, only synthetic shapes.

Counterfactual evaluation in the CLEVR family. CLEVRER (Yi et al., 2020) pioneered static counterfactual questions in this domain ("without object A, which event would not happen?"), with the counterfactual outcome supplied by a physics simulator at dataset-construction time as a fixed test item, i.e. counterfactual reasoning is the task itself. We reuse the same source of exact ground truth (scene programs), but apply the counterfactual as a measurement device at evaluation time on ordinary (non-counterfactual) questions: we re-execute a question's own program with the task-relevant object excluded to obtain the true counterfactual answer, then test whether an independently-elicited model answer matches it – the model is never asked a counterfactual question, only probed with one.

3 Method

Protocol. Given an image I and question q with ground-truth answer $a^\star$, an independent oracle (a program or scene graph, never the policy) identifies the task-relevant region $r \subseteq I$ that $a^\star$ depends on, and a task-irrelevant control region c, spatially separated from r. The oracle also recomputes the true counterfactual answer $a^\star_{\setminus r}$ by re-deriving the answer with r excluded from the scene (undefined, $\perp$, when the question becomes ill-posed once its referent is removed – e.g. a query about a now-absent unique object). A policy (VLM) f is queried independently on each of four images: the original I, I restricted to r (written $I|_r$, everything else background), I with r removed (written $I \setminus r$, background-composited or inpainted depending on tier), and I with c removed ($I \setminus c$). Writing $a = f(I, q)$ for the policy's original answer, we compute:

$$\text{Sufficiency} = \mathbf{1}[f(I|_r, q) = a],$$

$$\text{Necessity} = \mathbf{1}[f(I \setminus r, q) = a^\star_{\setminus r}] \quad \text{(defined only when } a^\star_{\setminus r} \neq \perp),$$

$$\text{Specificity} = \mathbf{1}[f(I \setminus c, q) = a].$$

A looser, always-defined necessity proxy, $\text{Necessity}_{\text{loose}} = \mathbf{1}[f(I \setminus r, q) \neq a]$, credits any change in the answer rather than a change to the correct new answer; we report both, and treat $\text{Necessity}_{\text{loose}}$ as confounded on its own (a model that changes its answer under any perturbation, relevant or not, would score highly) – Specificity rules this out. Our primary statistic is the strict joint rate Necessity $\wedge$ Specificity on the subset with $a^\star_{\setminus r} \neq \perp$: the fraction of well-posed items where removing the correct evidence changes the answer to the true new answer, and removing irrelevant evidence does not change it at all.

Tier 1: CLEVR (exact evidence, no re-rendering). CLEVR (Johnson et al., 2017) ships, for each question, both the ground-truth scene graph (objects with color/shape/size/material/3D position) and the functional program used to generate the question. We execute this program against the scene graph (implemented once, reused for every question family that decomposes into filter/unique/relate/count/exist/query/same/compare primitives) to recover the exact set of scene-object indices the answer depends on – without needing access to the original Blender scene files. Because CLEVR objects sit on a uniform, largely flat background plane, we remove an object by local color-distance segmentation around its known pixel coordinates followed by background-color compositing, rather than diffusion inpainting.

Table 1: Main results across tiers and model scales (Section 4).

Tier	Model	n	Orig. Acc.	Sufficiency	Specificity	Necessity (strict, n_s)	Joint (strict)
Tier1-CLEVR	Qwen3VL-2B	3000	0.870	0.794	0.913	0.721 (n_s=853)	0.619
Tier1-CLEVR	Qwen3VL-8B	3000	0.965	0.798	0.921	0.785 (n_s=853)	0.680
Tier2-GQA	Qwen3VL-2B	962	0.679	0.695	0.923	0.439 (loose)	0.389 (loose)
Tier2-GQA	Qwen3VL-8B	962	0.685	0.711	0.939	0.403 (loose)	0.368 (loose)
Tier2-GQA	Qwen3VL-32B	962	0.660	0.683	0.931	0.404 (loose)	0.364 (loose)

Tier 2: GQA (real images, inpainting-based counterfactuals). GQA (Hudson & Manning, 2019) question annotations directly name the scene-graph object id(s) each reasoning step in the question's semantic program touches, together with precise bounding boxes per object – so evidence identification again requires no additional model or heuristic. Because GQA images are natural photographs, gray-box masking is a poor counterfactual (it introduces an out-of-distribution artifact a model could learn to key on independent of content); we instead inpaint the evidence region (classical Telea inpainting in the current pass; a diffusion-based upgrade is a natural extension, see Limitations). Unlike Tier 1, GQA's semantic programs are not uniformly re-executable to obtain a true counterfactual answer for arbitrary question types, so Tier 2 reports the loose necessity proxy only; this asymmetry between tiers is intentional and disclosed, not an oversight.

Models. We evaluate open-weight VLMs only, zero-shot, no fine-tuning: Qwen3-VL-2B/8B/32B-Instruct (Bai et al., 2025), spanning a 16× parameter range within one model family so that any trend can be attributed to scale rather than architecture. We attempted to add InternVL3.5 as a second model family but encountered a transformers-version processor incompatibility not worth debugging under the submission deadline; we report this rather than silently restricting scope to Qwen3-VL (see Limitations).

4 Experiments

Setup. We evaluate Qwen3-VL-2B/8B (32B on Tier 2 only, see below) zero-shot on 3,000 Tier-1 (CLEVR) items (853 with a well-defined strict counterfactual, i.e. 28.3% of a larger 8,715-item pool; the well-defined subset is fixed across models by construction, so the 853 items compared for 2B and 8B are identical) and all 962 Tier-2 (GQA) items. All results are averages over the full evaluated set with 95% bootstrap confidence intervals (2,000 resamples); cross-model comparisons use a χ^2 test on the relevant binary outcome.

Tier 1 (CLEVR): a real, statistically significant scale effect. On the primary, strict metric (necessity checked against the true recomputed counterfactual answer, jointly with specificity), Qwen3-VL-8B reaches 0.680 [0.648,0.710] versus 2B's 0.619 [0.586,0.651] on the same 853 well-defined items – non-overlapping confidence intervals, confirmed significant by a χ^2 test (p=0.0097). Original accuracy shows an even larger, highly significant gap (0.965 vs. 0.870, $p<10^{-6}$). Sufficiency is nearly identical across scale (0.798 vs. 0.794) and specificity is high for both (0.921 vs. 0.913), so the scale advantage is concentrated specifically in correctly tracking the counterfactual answer, not merely in being more answer-stable overall. The strict necessity rate alone (0.785 vs. 0.721) shows the same direction. We take this as the paper's central positive finding: when ground truth is available to check against, larger models in this family are measurably more causally grounded in the task-relevant evidence, not merely more accurate in general.

Tier 2 (GQA): real-world images. On 962 real-photograph items, Qwen3-VL-2B/8B/32B show orig. accuracy 0.679/0.685/0.660, sufficiency 0.695/0.711/0.683, and specificity 0.923/0.939/0.931 – all comfortably above chance and, notably, specificity remains high (above 0.92) across every scale, indicating models are not simply reacting to the presence of any perturbation regardless of domain. The joint (loose) rate is 0.389/0.368/0.364. Unlike Tier 1, none of the pairwise scale differences on Tier 2 are significant (accuracy: 32B vs. 8B p=0.264, 32B vs. 2B p=0.410; joint: 32B vs. 8B p=0.887, 32B vs. 2B p=0.279, all

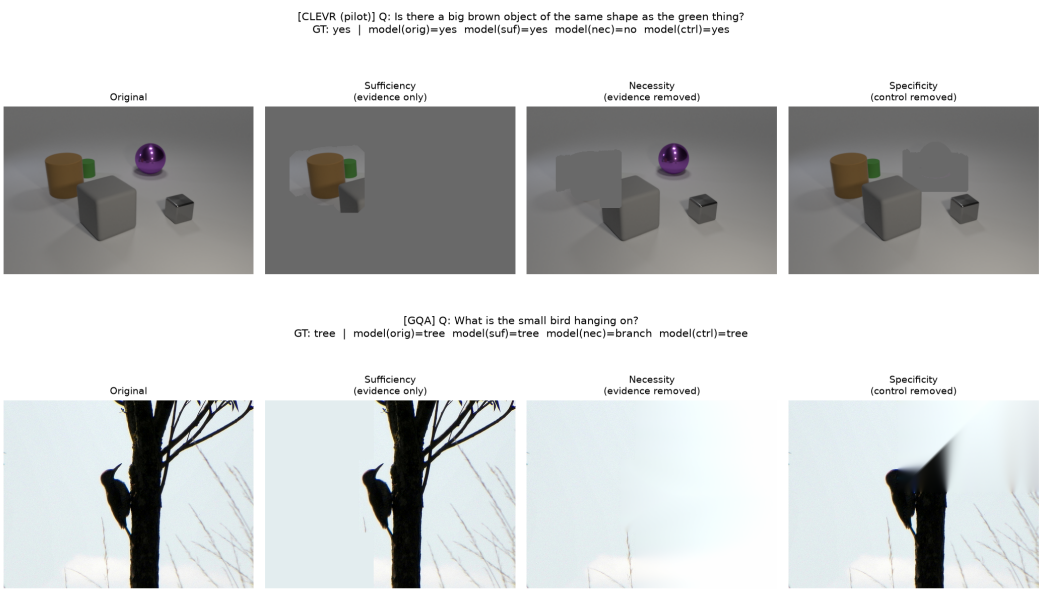

Figure 1: Qualitative examples of the sufficiency/necessity/specificity protocol on one CLEVR item (top) and one GQA item (bottom), Qwen3-VL-8B. Each panel shows the four images the policy is queried on and the resulting answers.

χ^2) – a genuine and, we think, informative contrast: scale improves groundedness in the controlled synthetic setting but confers no measurable advantage on natural images in this range, suggesting whatever scale buys models on CLEVR-style compositional queries does not transfer to the more varied evidence structure of real photographs.

Qualitative examples. Figure 1 shows one CLEVR and one GQA item where the model (8B) behaves exactly as the protocol intends: the answer is correct originally and under sufficiency, flips under necessity, and is unchanged under the control perturbation (CLEVR); on GQA, necessity removal changes the answer to a plausible near-miss ("branch" for "tree") rather than an unrelated answer, suggesting some, but not exact, residual sensitivity to the removed context.

5 Limitations

(1) Tier 1's exact counterfactuals rely on a synthetic domain (CLEVR); Tier 2 moves to real photographs (GQA) but at the cost of the strict counterfactual-answer check, since GQA's semantic programs are not uniformly re-executable the way CLEVR's are – Tier 2 numbers should be read as a loose, always-defined necessity proxy, not the primary strict metric. (2) Tier 2's counterfactuals use classical (Telea) inpainting rather than a diffusion model; while our control condition (removing an irrelevant object with the identical inpainting procedure) rules out a model simply reacting to any inpainting artifact, a diffusion-based upgrade would produce more naturalistic counterfactuals and is a direct extension. (3) Evidence sets are capped at 2 objects by construction (Section 3), so questions whose answer depends on 3 or more objects are excluded from both tiers; this likely biases toward simpler reasoning chains. (4) The strict necessity metric is only defined when the ground-truth counterfactual answer exists (not $\perp$); on CLEVR this held for 853/3,000 (28.4%) of evaluated items (the remainder become ill-posed once their sole referent object is removed, e.g. a query about a now-absent unique object) – we report this subset size explicitly rather than silently discarding it or conflating it with the well-defined cases. (5) On the same well-defined subset, the loose (any-change) and strict (correct-change) necessity rates are close for both models (2B: 0.742 loose vs. 0.721 strict; 8B: 0.782 loose vs. 0.785 strict) – unlike a naive cross-population

comparison against Table 1's loose rate computed over all 3,000 items (which is not the same set and should not be compared directly), the same-subset comparison shows the loose proxy does not substantially overstate groundedness here, which offers some support for using it as a proxy in Tier 2, where the strict check is unavailable; we do not take this as license to treat the two metrics as interchangeable in general, since the theoretical case for a specificity-controlled, ground-truth- verified check (Section 3) is independent of whether the two happen to agree on this particular item population. (6) Tier 2's GQA items are drawn from only 400 images (962 items after filtering); a larger image pool, following the same range-request pipeline used for Tier 1, is a direct way to tighten Tier 2's confidence intervals and is not a methodological limitation, only a time-boxed scope choice for this submission cycle. (7) We test one model family (Qwen3-VL) at three scales; InternVL3.5 failed to load due to a transformers-version processor incompatibility we did not debug under the submission deadline (Section 3), so whether the CLEVR-vs-GQA scale contrast (Section 4) generalizes across architectures is untested.

6 Extension: Training on Self-Reported Grounding Can Degrade True Grounding

Sections 1–5 are diagnostic: they measure whether an existing model's answer is grounded, but do not change the model. A natural next question is whether the same oracle machinery – ground-truth evidence regions and recomputed counterfactual answers, derived once from CLEVR's released scene programs – can also serve as a training signal. We report a controlled experiment addressing this, with a result we did not expect: training on the kind of signal that is easy to supervise – a model's own self-reported evidence region – can substantially degrade the very property this paper measures, even as every self-reported metric suggests the opposite is happening. This is, to our knowledge, a direct empirical demonstration that the self-report skepticism motivating this paper's entire protocol (Section 1) also applies to training, not only to evaluation.

Setup. We train Qwen3-VL-2B/8B with LoRA and GRPO (Shao et al., 2024) via TRL (von Werra et al., 2020) on a grounded-QA task built from the same CLEVR evidence-object pool as Tier 1: given an image and question, the policy outputs an answer and a bounding box for the object it used, normalized to a 0–1000 scale (Qwen-VL's native grounding convention). The reward combines format compliance, answer correctness, and IoU between the predicted box and the program-identified evidence object's box – i.e. the reward, like the box itself, is built entirely from the model's self-report of what it is looking at, exactly the kind of signal Section 1 argues should not be trusted at face value. Training uses 1,020 items (300 GRPO steps for 8B, checkpoint grpo_full_8b; 2B is trained for 600 steps with the program-verifier reward added, checkpoint grpo_full_v2_long, see below), group size 8, 180 held-out items for the self-reported evaluation (Table 2); no test item overlaps training. For the causal-verification evaluation (Table 3), which reuses the larger Tier-1 item pool for statistical power, we additionally identified and excluded all 1,020 GRPO training items from the 3,000-item Tier-1 manifest (347 of them, 34%, were present in it) before evaluating baseline and trained checkpoints, so that no reported causal-groundedness number is computed on any item either model was trained on; this leaves 2,653 items (751 well-defined for the strict counterfactual) for a clean, non-overlapping comparison.

Step 1: the self-reported picture looks like unambiguous success. Table 2 reports held-out answer accuracy and box IoU under this self-reported evaluation. Training produces a large, significant improvement in both for the smaller model (2B: accuracy 0.617→0.889, p=4.5×10^{-9}; IoU 0.250→0.412, p=6.5×10^{-11}; a longer, 600-step run improves IoU further to 0.522, p=5.3×10^{-9} over the 300-step run), and for the larger model, whose baseline accuracy is already high (0.939), training leaves accuracy statistically unchanged (p=1.0) while still producing a very large improvement in box IoU (0.295→0.597, p=1.1×10^{-30}). Read on its own, this table tells a clean, appealing story: cheap, label-free training closes an accuracy-vs-grounding gap. A reward-shaping ablation (adding format/answer terms on top of a purely load-bearing-object-verifier term, reusing the program executor from Section 3) shows a consistent but not significant trend on its own targeted metric (n=180,

Table 2: Self-reported evaluation: held-out accuracy, IoU, and the program-verifier's targeted load-bearing rate, before and after GRPO training. "full" = format + answer + IoU reward against the model's own predicted box; "+ verifier" additionally rewards pointing at objects that are truly load-bearing per the ground-truth program.

Model	Condition	n	Answer Acc.	IoU	Load-bearing
Qwen3-VL-2B	baseline (untrained)	180	0.617	0.248	–
Qwen3-VL-2B	trained, full reward	180	0.889	0.412	0.794
Qwen3-VL-2B	trained, IoU-only reward	180	0.828	0.425	–
Qwen3-VL-2B	trained, full + verifier	180	0.889	0.405	0.844
Qwen3-VL-2B	trained, full + verifier (2× steps)	180	0.906	0.522	0.861
Qwen3-VL-8B	baseline (untrained)	180	0.939	0.295	–
Qwen3-VL-8B	trained, full reward	180	0.933	0.597	–
Qwen3-VL-8B	trained, IoU-only reward	180	0.944	0.607	–

Table 3: Causal verification: applying this paper's own sufficiency/necessity/specificity protocol (not self-reported boxes) to the same baseline and trained checkpoints, on the 2,653 Tier-1 items not seen during GRPO training (751 well-defined for the strict counterfactual). For 2B, every self-reported metric in Table 2 improved, yet true causal groundedness (Necessity, Joint) significantly decreased. For 8B, training produced no significant change in true groundedness at all.

Model	Condition	Orig. Acc.	Sufficiency	Specificity	Necessity (strict)	Joint (strict)
Qwen3-VL-2B	baseline	0.635	0.551	0.805	0.656	0.571
Qwen3-VL-2B	trained	0.827	0.771	0.922	0.537	0.478
Qwen3-VL-2B	p-value	1.7×10^{-55}	3.8×10^{-64}	2.9×10^{-35}	2.9×10^{-6}	3.6×10^{-4}
Qwen3-VL-8B	baseline	0.936	0.786	0.913	0.759	0.648
Qwen3-VL-8B	trained	0.946	0.806	0.916	0.736	0.632
Qwen3-VL-8B	p-value	0.146	0.077	0.659	0.342	0.554

p=0.125) and no significant difference from an IoU-only reward at either scale (2B: p=1.00 accuracy; 8B: p=0.83 accuracy) – GRPO training itself matters far more than these specific reward-shaping choices.

Step 2: does self-reported improvement mean the answer is actually more grounded? We directly tested this by applying this paper's own protocol (sufficiency, necessity, specificity; Section 3) to the baseline and trained checkpoints, on the training- excluded 2,653-item Tier-1 subset described above (751 with a well-defined strict counterfactual), querying each model in its own trained prompt format so that a format mismatch cannot confound the baseline-vs-trained comparison, and scoring only the stated answer against ground truth – never the box. Table 3 reports the result, and it does not match Step 1's story. For 2B, sufficiency and specificity both improve significantly (both $p < 10^{-30}$) – consistent with a more capable, more consistent model – but strict necessity, the rate at which removing the true evidence changes the answer to the correct counterfactual answer, falls from 0.656 to 0.537 (p=2.9×10^{-6}), and the primary joint metric falls from 0.571 to 0.478 (p=3.6×10^{-4}). Both are highly significant decreases. For context, the same untrained 2B model's joint score measured with its own native free-text protocol prompt (Table 1's prompt, rather than the box-output format both rows of Table 3 use) is 0.611 on this identical item subset – so the trained checkpoint's 0.478 is not an artifact of the box-output prompt specifically depressing the baseline: it falls well below the model's best independently-observed causal groundedness under either prompt. For 8B, by contrast, none of the five metrics change significantly in either direction (all $p > 0.05$): training that produced a large, significant self-reported IoU gain (Table 2) produced no detectable change, good or bad, in actual causal groundedness.

Interpretation. A plausible mechanism: training on a self-reported box/IoU reward made the smaller model's answers more stable under perturbation in general – helpful when the

perturbation is irrelevant (specificity rises) but harmful when it is exactly the evidence the answer should depend on (necessity falls), i.e. the model becomes more confidently consistent rather than more genuinely grounded. We did not anticipate this result when designing the experiment; we report it because it is, we think, a materially more important finding than the training-improves-grounding story we set out to test. It is also a direct, quantified answer to a question this paper otherwise only poses in the abstract: yes, optimizing a self-reported signal – during training, not just at evaluation time – can actively work against the property it appears to be improving, and whether it does so depends on model scale in a way that a self-reported metric alone cannot reveal. We see this as a cautionary result for the broader line of work that trains VLMs with self-reported localization or grounding rewards (Section 2), not only for the specific setup tested here. We checked this finding against two natural objections before reporting it: that it might reflect data leakage from GRPO training into the evaluation set (ruled out by excluding all training items, as described above) or an artifact of forcing the untrained baseline into the trained model's box-output prompt format (ruled out by the free-text reference comparison above, which shows the trained checkpoint's groundedness falls below the model's best independently-observed score under either prompt).

7 Conclusion

We introduced a self-report-free, causal-intervention protocol for testing whether a VLM's answer is actually grounded in the task-relevant evidence, rather than trusting what the model says about its own reasoning. Applied zero-shot to Qwen3-VL at three scales across two domains, the protocol surfaces a real and, we think, useful pattern. Models are consistently and highly specific (specificity $\geq$0.91 everywhere), so their answers are not merely destabilized by any image perturbation. Scale produces a significant, sizeable improvement in correctly tracking a true recomputed counterfactual on CLEVR (p=0.0097), but confers no significant advantage on the more varied evidence structure of real GQA photographs across the same 2B–32B range. And where both are computable, our loose and strict necessity metrics agree closely on CLEVR, offering some support for the loose proxy used in Tier 2, while the theoretical case for the stricter check stands independently of that agreement. Because the protocol requires no training, it is directly reusable as a diagnostic for any VLM. With the true-counterfactual-answer machinery restricted here to CLEVR's re-executable programs, a natural direction is extending strict, ground-truth- verified necessity to broader real-world question types where such re-execution is possible (e.g. compositional or scene-graph-native VQA beyond GQA's current program coverage), closing the disclosed Tier-1/Tier-2 asymmetry rather than only reporting it.

Section 6 extends the same self-report skepticism from evaluation to training: reinforcement learning on a model's own self-reported evidence region produces a large, unambiguous improvement in every self-reported metric, yet the paper's own causal protocol reveals that true groundedness significantly decreases for the smaller model tested and is unchanged for the larger one. Self-report cannot be trusted as a training signal any more than as an evaluation signal, and whether it is actively harmful depends on model scale in a way self-report alone cannot reveal. We see this as a direct, quantified reinforcement of this paper's central argument, obtained from a direction we did not originally set out to test.

AI Use Statement

AI tools were used to assist code development and debugging, literature retrieval, and manuscript drafting/editing. The authors reviewed the final manuscript and take responsibility for all claims and content.

Reproducibility Statement

The full protocol is specified in Section 3, with per-tier construction details in the corresponding paragraphs (Tier 1/Tier 2) and the exact item counts, held-out subsets, and significance tests used throughout Section 4 and Section 6. All models evaluated (Qwen3-

VL-2B/8B/32B-Instruct) are open-weight and publicly available; CLEVR and GQA are public datasets. The CLEVR program executor, GQA item-construction and inpainting pipeline, evaluation harness, GRPO training scripts, and statistical analysis code, together with all raw per-item result files underlying every number reported in this paper, will be released at an anonymous public repository upon publication.

Ethics Statement

This work uses only publicly released, non-human-subject datasets (CLEVR: synthetic rendered scenes; GQA: image-caption-style natural photographs with existing public licenses) and publicly released open-weight models. No new data involving people was collected. We do not foresee significant risks from this work beyond those already associated with the underlying benchmarks and models; if anything, a protocol for detecting when a VLM's answer is not causally grounded in the correct evidence is intended to support safer, more trustworthy deployment of such models.

References

Rabiul Awal, Saba Ahmadi, Le Zhang, and Aishwarya Agrawal. Vismin: Visual minimal-change understanding. In NeurIPS, 2024.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-vl technical report. arXiv preprint arXiv:2511.21631, 2025.

Aayam Bansal and Ishaan Gangwani. Counterbench: A controllable counterfactual testbed reveals systematic reasoning failures in vision-language models. 2026. Synthetic Sciences. OpenReview id U9LQ19atDB.

Haojie Huang, Xinlei Yu, Chengming Xu, Zhangquan Chen, Cheng Yang, Qingdong He, Yu Yang, Jiangning Zhang, and Xiaobin Hu. Evidence-rl: Towards evidence-intensive visual reasoning. arXiv preprint arXiv:2608.08021, 2026.

Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In CVPR, 2019.

Ben Jenkins. Thinking in pictures: A diagnostic study of visual vs. textual chain-of-thought reasoning in vision-language models. ACL Anthology 2026.alvr-main.1, 2026.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In CVPR, 2017.

Zixuan Lan, Luzhe Sun, Matthew R. Walter, and Jiawei Zhou. Seeing without looking: Do vision-language benchmarks really test vision? arXiv preprint arXiv:2605.22903, 2026.

Lujun Li, Lama Sleem, Niccolo Gentile, Yangjie Xu, Yewei Song, Wenbo Wu, and Radu State. The last visible pixel: Probing fine-scale perception in vision-language models. arXiv preprint arXiv:2606.07861, 2026.

Fenfen Lin, Yesheng Liu, Haiyu Xu, Chen Yue, Zheqi He, Mingxuan Zhao, Miguel Hu Chen, Jiakang Liu, JG Yao, and Xi Yang. Do vision-language models measure up? benchmarking visual measurement reading with measurebench. arXiv preprint arXiv:2510.26865, 2026.

Jiahao Meng, Tan Yue, Qi Xu, Haochen Wang, Zhongwei Ren, Weisong Liu, Yuhao Wang, Renrui Zhang, Yunhai Tong, and Haodong Duan. Videozerobench: Probing the limits of video mllms with spatio-temporal evidence verification. arXiv preprint arXiv:2604.01569, 2026.

Bayar Menzat, Maximilian Süss, Ruizhi Wang, Benno Steinegger, Thomas Lukasiewicz, and Oana-Maria Camburu. Counterfactual tests for measuring chain-of-thought faithfulness in visual language models. Findings of EMNLP, 2026.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300, 2024.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In CVPR, 2022.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. TRL: Transformers Reinforcement Learning, 2020. URL https://github.com/huggingface/trl.

Tan Wang, Kevin Lin, Linjie Li, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Equivariant similarity for vision-language foundation models. In ICCV, 2023.

Xiyao Wang, Zhengyuan Yang, Chao Feng, Yongyuan Liang, Yuhang Zhou, Xiaoyu Liu, Ziyi Zang, Ming Li, Chung-Ching Lin, Kevin Lin, Linjie Li, Furong Huang, and Lijuan Wang. Vicrit: A verifiable reinforcement learning proxy task for visual perception in vlms. arXiv preprint arXiv:2506.10128, 2025.

Wenjie Yang, Siqi Zhu, and Zengfeng Huang. Position: Your vlm may not be thinking with interleaved images. In ICML, 2026.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. Clevrer: Collision events for video representation and reasoning. In ICLR, 2020.

Shixin Yi and Lin Shang. Corgi: Verified chain-of-thought reasoning with post-hoc visual grounding. arXiv preprint arXiv:2508.00378, 2025. Withdrawn by the authors, October 2025.

Jianrui Zhang, Mu Cai, Tengyang Xie, and Yong Jae Lee. Countercurate: Enhancing physical and semantic visio-linguistic compositional reasoning via counterfactual examples. In Findings of ACL, 2024.