ASK A BETTER QUESTION: UNLOCKING WHAT VISION-LANGUAGE MODELS ALREADY GET RIGHT

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models are queried through language, yet a frozen model's answer can vary substantially across questions intended to express the same visual request. This sensitivity creates both a practical failure mode and an opportunity: the same model may answer an alternative formulation correctly after failing on the original question. We ask whether the model's own first response can guide a better formulation at inference time. We introduce Inference-time response-CONditioned reframing (ICON), a lightweight text rewriter that observes the original question, answer options when present, and the frozen VLM's initial response, then produces a single revised question for a second pass through the same model. ICON is trained offline from supervised rewrite pairs, but deployment requires neither gold labels nor candidate search, and the target VLM remains frozen throughout. On Robo2VLM, relative to question-only rewriting, the full response-conditioned rewriter increases the mean accuracy gain from 1.65 to 8.46 percentage points with a 4B rewriter and from 2.12 to 8.54 points with an 8B rewriter. In the matched comparison, adding the initial response to question-and-options-only rewriting contributes an additional 4.84 and 4.30 points, respectively. Across eight benchmarks, ICON yields gains of 0.20 to 8.54 percentage points, and a Robo2VLM-trained rewriter transfers across several frozen VLMs on the same task. Together, these results identify the model's own response as a useful conditioning signal for targeted question reformulation, enabling the same frozen VLM to produce correct answers under revised formulations when the original wording leads to an error.

1 INTRODUCTION

Vision-language models (VLMs) are typically queried through natural-language interfaces (OpenAI, 2023; Gemini Team, Google, 2023; Liu et al., 2023b). Yet this interface is not neutral: questions intended to request the same visual judgment from the same image or video can elicit different answers from a frozen model. As a result, model performance can depend on how a request is phrased even when the visual input and intended judgment remain unchanged.

Consider the three scenes in Figure 1. In an indoor image from the EmbSpatial benchmark (Du et al., 2024), a frozen Qwen3-VL-2B model answers *picture* when asked which object has the *nearest placement*, but answers *sofa* correctly when asked which item is *closest to you*. In a robot frame from Robo2VLM (Chen et al., 2025), the same model answers *purple* when asked which colored arrow indicates the robot's next motion for achieving *to the right*, but answers *red* correctly when the question specifies the arrow pointing right. In a dashcam frame from LingoQA (Marcu et al., 2024), the model answers *yes* to *can you turn right?*, but correctly answers *no* when asked whether it is *safe to turn right* under the current traffic conditions. In each example, the visual input remains fixed and the questions are intended to request the same judgment, yet the model's answer changes.

These examples do not imply that every model error is caused by wording. Instead, they demonstrate a narrower and testable phenomenon: some errors are recoverable under alternative formulations intended to preserve the same request. Our detailed analyses on Robo2VLM and LingoQA in Section 3 show that the frozen model changes its answer on 73.2% and 43.0% of questions, respectively, across alternative formulations. Moreover, at least one sampled formulation corrects 47.0% and 42.8% of the errors made under the original wording, respectively (please check Section 3 for details).

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.

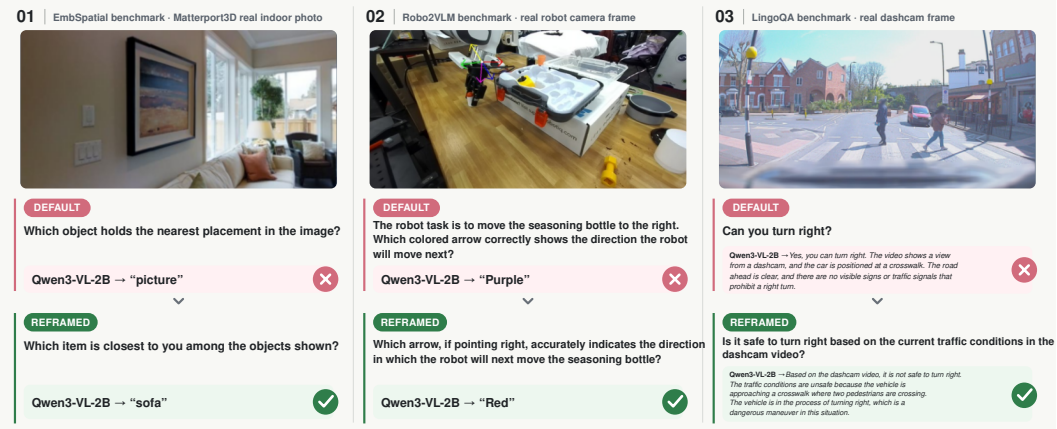

Figure 1: Examples of prompt sensitivity. **Left** (EmbSpatial, a Matterport3D indoor photo): when asked which object has the 'nearest placement," Qwen3-VL-2B answers *picture* (✗); when asked which item is 'closest to you," it answers *sofa* (✓). **Middle** (Robo2VLM, a real robot camera frame): when asked which 'colored arrow shows the direction the robot will move next," Qwen3-VL-2B answers *purple* (✗); when asked which arrow, 'pointing right, indicates the direction it will next move the seasoning bottle," it answers *red* (✓). **Right** (LingoQA, a real dashcam frame): when asked 'can you turn right?" Qwen3-VL-2B answers *yes, you can turn right* (✗); when asked whether it is 'safe to turn right" given the current traffic conditions, it answers that it is *not safe* because pedestrians are crossing (✓). Across all three examples, the visual input and underlying intent remain unchanged; only the phrasing changes. These failures therefore expose a gap between what the model can *recognize* and what a particular question formulation can elicit.

This observation raises a practical question: *can some of this lost accuracy be recovered at inference time without gold labels or changes to the frozen model?* We introduce Inference-time response-CONditioned reframing (ICON), which uses the frozen model's first response as context for asking the question again. ICON first queries the frozen VLM on the original question and then conditions a single reformulation on the original question, answer options when present, and the initial response a_{ori}. The same frozen VLM then answers the revised question. Importantly, ICON neither knows nor attempts to determine whether a_{ori} is correct. Instead, the initial response provides model-specific and request-specific context for targeted reformulation.

Our central empirical finding is that this response conditioning improves learned question reformulation. In the matched comparison on Robo2VLM, adding a_{ori} to question-and-options-only rewriting contributes an additional 4.84 and 4.30 percentage points for the 4B and 8B rewriters, respectively. This comparison holds the filtered training data, rewriter backbone, evaluation examples, and frozen VLM (answerer) fixed, isolating whether the initial response is available to the rewriter.

The resulting procedure is deliberately simple. ICON is implemented as a LoRA adapter on a frozen text backbone, while the target VLM remains frozen throughout. At inference time, it requires one rewriter call and one additional target-model pass, with no gold answer or candidate search. Relative to question-only rewriting, the full response-conditioned rewriter increases the mean accuracy gain on Robo2VLM from 1.65 to 8.46 percentage points with a 4B rewriter and from 2.12 to 8.54 points with an 8B rewriter. ICON produces positive mean gains across eight benchmarks, ranging from 0.20 to 8.54 points, and a Robo2VLM-trained rewriter transfers across several frozen answerers on the same task. In contrast, all tested zero-label baselines are non-positive on Robo2VLM. Together, these results identify the model's initial response as a useful conditioning signal for targeted question reformulation.

2 Related Work

Prompt sensitivity and question reformulation. Language models are sensitive to how a request is phrased (Zhao et al., 2021; Zhu et al., 2023; Perez et al., 2022). Prior work addresses this

sensitivity through several forms of prompt reformulation and optimization. Rephrase and Respond asks a model to reformulate a question before answering it (Deng et al., 2023). PRewrite learns a prompt rewriter through reinforcement learning (Kong et al., 2024), while RLPrompt optimizes discrete prompts (Deng et al., 2022). OPRO and APE instead search for task level instructions using development data (Yang et al., 2024; Zhou et al., 2023). RePARe is particularly related in multimodal question answering: it generates visually grounded reformulations and selects among candidates using model confidence (Prasad et al., 2024). ICON focuses on a complementary signal for per instance reformulation: the frozen answerer's own initial response. It first queries the target VLM and then conditions a single natural language rewrite on the question, answer options when present, and the resulting self answer. Our controlled comparison holds the filtered training data and rewriter backbone fixed and varies whether this initial response is available to the rewriter, directly testing the contribution of response conditioning. At deployment, ICON requires neither gold labels nor label based candidate selection.

Response refinement and test time adaptation. Chain of thought (Wei et al., 2022), self refinement (Madaan et al., 2023), and self consistency (Wang et al., 2023) elicit, revise, or aggregate responses. ICON instead uses the initial response as context for changing the *question* before asking the same frozen model again. Test time prompt tuning adapts a soft prompt for a frozen CLIP style model without labels (Shu et al., 2022). ICON likewise leaves the target model frozen, but produces a natural language reformulation rather than adapting a soft prompt. More broadly, our motivation relates to capability elicitation: a model may be able to produce a correct answer under some formulations even when it fails under another. We use the gap between pass@K and pass@1 (Brown et al., 2024) only as a diagnostic of recoverable performance, rather than as a proposed metric, and separately examine whether the pass@K–pass@1 gap predicts where ICON succeeds in Section 3.

Multimodal robustness and language reliance. Visual question answering research has studied robustness to question paraphrases, while image free baselines have long been used to expose language priors and dataset biases (Goyal et al., 2017; Li et al., 2023; Chen et al., 2024). More broadly, multimodal benchmarks increasingly span temporal, spatial, and multidisciplinary reasoning (Fu et al., 2023; Yue et al., 2024; Li et al., 2024b; Mangalam et al., 2023; Fu et al., 2025; Gan et al., 2025). Our results highlight a complementary evaluation issue: performance can vary substantially across alternative formulations of the same underlying request, motivating evaluation beyond a single question phrasing.

3 Problem Formulation and Prompt Sensitivity

Measuring prompt sensitivity. Let X_n denote the multimodal input for item n, M a frozen vision-language model, and q_{n0} the original benchmark question. For each item, we construct a set of K formulations, $\mathcal{P}_n(z_n) = \{q_{n0}, q_{n1}, \ldots, q_{n(K-1)}\}$, intended to express the same underlying request z_n. The set includes the original benchmark question q_{n0}, while the remaining $K-1$ alternative formulations are generated by GPT-5.5. Let $y_n(q)$ denote the normalized answer produced by M for input X_n when queried with formulation q. Ideally, formulations expressing the same underlying request should produce the same answer:

$$q \equiv_{z_n} q' \quad \Longrightarrow \quad y_n(q) = y_n(q'), \qquad q, q' \in \mathcal{P}_n(z_n). \tag{1}$$

This measures the percentage of examples for which at least two formulations yield different answers. Multiple-choice and yes/no responses are normalized to categorical labels before comparison.

Best-of-K accuracy and error correction. Answer changes alone do not reveal whether alternative wording can improve accuracy. Let $c_n(q) \in \{0, 1\}$ indicate whether the generated answer under formulation q is correct according to the benchmark's evaluation metric. We define best-of-K accuracy as

$$\text{Accuracy}_{\text{Best-of-}K} = \frac{100}{N} \sum_{n=1}^{N} \max_{q \in \mathcal{P}_n(z_n)} c_n(q). \tag{2}$$

Because the original question is included in $\mathcal{P}_n(z_n)$, best-of-K accuracy is the accuracy obtainable if the best sampled formulation could be selected for each example. It is a label-aware upper bound rather than a deployable result, since selecting the best formulation requires evaluating multiple alternatives and knowing the gold answer. For interpretability, we also report the error correction rate:

among examples answered incorrectly under the original question, this is the percentage for which at least one alternative formulation produces a correct answer. The error correction rate describes the same corrected examples as the best-of-K gain, but normalizes their number by the original error count rather than by the full evaluation set. Both quantities are used only to characterize prompt sensitivity and are unavailable to the deployed method.

Prompt sensitivity is substantial. We first measure this phenomenon over the full benchmark, independently of rewriter training. On Robo2VLM (Chen et al., 2025), using the full official test set, a frozen Qwen3-VL-2B changes its answer on 73.2% of questions across formulations. Its accuracy increases from 51.8% under the original questions to 74.5% under best-of-K selection, and at least one alternative formulation corrects 47.0% of the original errors. We observe a similar pattern in a different domain on LingoQA (Marcu et al., 2024) (N=500): the model changes its answer on 43.0% of questions, best-of-K accuracy increases from 45.8% to 69.0%, and at least one alternative formulation corrects 42.8% of the original errors. These results show that model behavior can depend substantially on question formulation and that alternative formulations can elicit correct responses when the original wording does not. These measurements characterize the available opportunity rather than deployable performance. Best-of-K selection evaluates multiple formulations and uses the gold answer to identify the best one. At inference time, neither the best formulation nor the correctness of the model's initial response is known.

From measurement to intervention. The analysis above measures prompt sensitivity over the full benchmark and is independent of rewriter training. Evaluating ICON requires training the rewriter, so the benchmark examples used in the rewriter study are divided into disjoint training and evaluation splits. The training split is used for offline pair construction and rewriter training, while all reported ICON results in Sections 5 and 6 are computed exclusively on the evaluation split. For Robo2VLM, the evaluation split contains 1335 items ($N = 1335$), on which the original Qwen3-VL-2B accuracy is 48.8%. LingoQA provides separate official training and validation splits, so we use the entire official 500-item validation set for evaluation. The two analyses therefore serve distinct purposes: the full benchmark analysis quantifies sensitivity to question formulation, whereas the experiments below measure the accuracy improvement achieved by a single learned reformulation without access to the gold answer at inference time.

4 Response Conditioned Reframing

Overview. Our key idea is to use the initial response as request specific context for a single reformulation. The target model first produces

$$a_{\mathrm{ori}} = M(X, q_0), \tag{3}$$

and the text rewriter then computes

$$q^\star = r_\theta(q_0, o, a_{\mathrm{ori}}), \qquad \widehat{y} = M(X, q^\star), \tag{4}$$

where o denotes answer options when present. The same target VLM is used in both passes and remains frozen throughout. The initial response is not treated as a correctness signal: at deployment, ICON neither knows nor attempts to determine whether a_{ori} is correct. Instead, it records how the frozen model responded to this particular request and supplies additional context for the reformulation. We illustrate this procedure in the "ICON INFERENCE" of Figure 2.

Offline pair construction. For each labeled training example (X, q_0, g), a proposer observes the original question and gold answer g and generates $K - 1$ candidate reformulations. The frozen target model answers each candidate, and correctness is determined by the benchmark's generated answer scorer. If the initial response is correct, the original question becomes an identity target, $q_0 \to q_0$. If the initial response is incorrect, a candidate that produces a correct generated answer can become the target after filtering. If no candidate succeeds, that training example is omitted. We illustrate this procedure in the "OFFLINE TRAINING OF ICON" of Figure 2. A useful rewriter should not change every question because a reformulation can break an answer that was initially correct. Identity targets therefore teach preservation, while successful candidates provide repair targets. This construction encourages selective rewriting without revealing whether the initial response is correct at deployment. Importantly, filtering is applied only when constructing the training data, all examples in the evaluation set remain in the evaluation and contribute to the reported metrics.

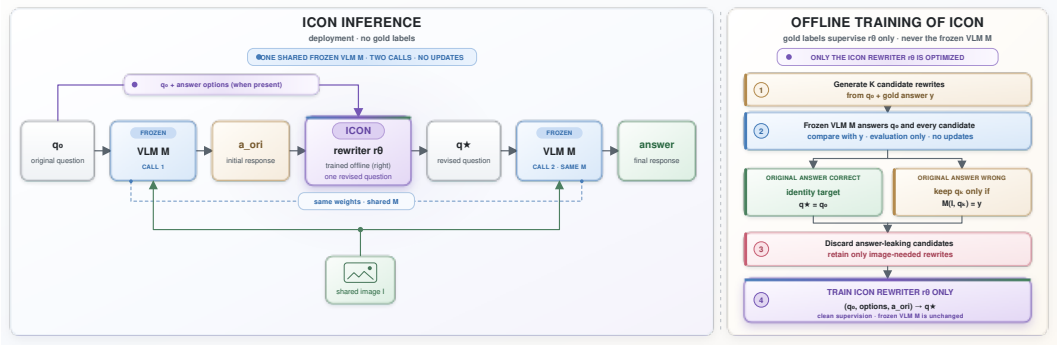

Figure 2: Standard ICON inference and offline pair construction. **Left:** the frozen VLM first answers q_0, producing a_{ori}. The text rewriter observes q_0, answer options when present, and a_{ori}, and emits one revised question $q^\star$. The same frozen VLM then answers $q^\star$. No gold label is available during deployment. **Right:** labels are used only offline to construct rewrite targets. Correct original questions provide identity targets, while incorrect original questions contribute a filtered candidate only when that candidate produces a correct generated answer.

Filtering answer-leaking candidates. A revised question may improve measured accuracy by revealing the answer in its wording rather than by better eliciting information from the visual input. To reduce this risk, we apply two checks before retaining a candidate as a training target in the "OFFLINE TRAINING OF ICON" of Figure 2. First, a rule-based answer-disclosure check identifies predefined patterns in which the revised question directly states or strongly implies the gold answer or the correct answer option. Candidates that match these patterns are discarded. Second, we replace the visual input with black pixels and evaluate both the original and revised questions. We discard a candidate if the revised question can be answered correctly after this replacement while the original question cannot. Gold labels are used only for these offline filtering decisions. These checks detect the specified forms of answer disclosure but do not establish semantic equivalence or complete reliance on the visual input.

Response conditioning, training, and inference. We consider three rewriter input configurations: Q, QO, and QOA, corresponding to the question only, question plus options, and question plus options plus initial response, respectively. QO and QOA use the same filtered targets, training split, rewriter backbone, evaluation items, and frozen target model, so their difference isolates the contribution of a_{ori}. ICON is trained as a LoRA adapter on a frozen text backbone, using rank $r = 64$ and $\alpha = 128$ across all benchmarks. The target VLM remains frozen. At deployment, ICON generates a single revised question with no candidate search or access to the gold answer.

Offline versus online cost. Pair construction evaluates multiple candidate formulations and uses the gold answer to select a target, but this process occurs only offline. Standard deployment neither retains the candidate set nor repeats the selection procedure. For each evaluation item, ICON makes one initial target call, one rewriter call, and one second target call. This distinction separates the label-aware sensitivity upper bound from the single-rewrite method and standard ICON from the costlier pooled consensus extension reported only as an auxiliary study.

5 Experiments

5.1 Rewriter Backbone Selection

Before the main comparisons, we evaluate seven text backbones for ICON while keeping the frozen Qwen3-VL-2B answerer, filtered training data, and training procedure fixed. We report Δ, the difference between revised and original accuracy, in percentage points. For a fair comparison, all results are averaged over training seeds 42, 43, and 44. Table 1 shows that performance on Robo2VLM is relatively insensitive to the rewriter backbone, with gains ranging from 7.54 to 8.54 points. CV-Bench (Tong et al., 2024) is more sensitive to this choice, with gains ranging from 1.74 to 5.81 points and the smallest Qwen2.5 backbone performing notably worse. Qwen3-VL-8B and

Table 1: Rewriter backbone comparison with a frozen Qwen3-VL-2B answerer and fixed training data. Values are mean accuracy changes in percentage points over three seeds. Bold marks the two backbones retained for subsequent experiments.

Rewriter base	Params	Robo2VLM Δ	CV-Bench Δ
Base Accuracy	–	48.8	72.8
Qwen2.5-1.5B-Instruct	1.5B	+8.04	+1.74
Qwen2.5-3B-Instruct	3B	+7.82	+3.23
Qwen2.5-7B-Instruct	7B	+8.39	+4.61
Qwen2.5-14B-Instruct	14B	+7.54	+4.89
Qwen3-VL-2B (text backbone)	2B	+7.64	+4.08
Qwen3-VL-4B (text backbone)	**4B**	**+8.46**	**+4.91**
Qwen3-VL-8B (text backbone)	**8B**	**+8.54**	**+5.81**

Qwen3-VL-4B rank first and second, respectively, on both benchmarks. We therefore retain these two backbones for all subsequent experiments. The 8B model performs best in this comparison, while the 4B model achieves similar performance on Robo2VLM with half the parameters. Reporting both avoids basing our conclusions on a single rewriter size.

5.2 Does the Initial Response Help?

We report the comparison with other methods in Table 2. The Q, QO, and QOA variants differ only in whether the rewriter receives answer options and the initial response. PRewrite, ReQueR, and both Advisor variants use the same filtered data but differ in their objectives or architectures, while the prompting baselines use no ICON training pairs. With the 4B rewriter, QO improves accuracy by 3.62 points, while QOA improves it by 8.46 points. With the 8B rewriter, the corresponding gains are 4.24 and 8.54 points. Providing the initial response in addition to the question and options therefore yields matched improvements of 4.84 and 4.30 points. By comparison, adding answer options to the question-only input yields improvements of 1.97 and 2.12 points. The full input comparison shows that the gain cannot be explained solely by providing answer options to the rewriter.

The broader baselines provide additional context. None of the evaluated zero-label prompting methods produces a positive net change in this setting. Trained refiners perform better than these prompting methods but remain below QOA. PRewrite achieves $+1.22/+1.87$ points and ReQueR achieves $+2.60/+2.22$ points, compared with $+8.46/+8.54$ points for ICON QOA. Because these methods also differ in training or inference, they do not provide controlled estimates of the contribution of a_{ori}. The matched QO versus QOA comparison provides that test.

We find this distinction is important. The question and options specify what is being requested, while a_{ori} captures the response produced by this particular frozen model for this particular request. Conditioning on a_{ori} therefore provides model-specific, request-specific context without revealing whether the response is correct. The experiment does not establish that a_{ori} exposes the model's internal state or identifies an error. Rather, it shows that providing this observable response improves learned question reformulation under a matched training and evaluation setup.

5.3 ICON across Diverse Multimodal Tasks

We next examine whether response-conditioned reformulation is effective across diverse multimodal tasks. These experiments use custom disjoint training and evaluation partitions of each benchmark dataset. We evaluate ICON on eight benchmarks: Robo2VLM, CV-Bench, LingoQA, ScienceQA, MMStar, MMMU, BLINK, and VSR, using a frozen Qwen3-VL-2B answerer. Each benchmark uses its official generated-answer scorer, and a separate rewriter is trained for each benchmark. This experiment therefore evaluates the effectiveness of the ICON approach across tasks. Table 3 summarizes the results. ICON yields positive gains across all eight benchmarks with both rewriter sizes. The largest improvements occur on Robo2VLM, CV-Bench, and LingoQA. ScienceQA and MMStar, MMMU, BLINK, and VSR show consistent positive gains. These results demonstrate that response-conditioned reformulation is effective across diverse tasks, although the magnitude of the improvement varies across benchmarks. In the Appendix, we show he training and evaluation split sizes for each benchmark are provided in Table 5 and further explore the cross-benchmark rewrite taxonomy in Table 6.

Table 2: Robo2VLM accuracy changes with a frozen Qwen3-VL-2B answerer.

	Rewriter base (Δ, 3-seed)	
Method	Qwen3-VL-4B	Qwen3-VL-8B
Zero-label baselines (rewriter-independent; frozen 2B answerer)		
Naive self-rephrase (no a_{ori})	-0.22	
OPRO (Yang et al., 2024)	-1.35	
RePARe, label-free (Prasad et al., 2024)	-0.37	
Self-consistency @K=8 (Wang et al., 2023)	-0.60	
Self-refine (Madaan et al., 2023)	-7.04	
Chain-of-thought (Wei et al., 2022)	-8.16	
Rephrase-and-Respond (Deng et al., 2023)	-10.41	
*Trained rewriter baselines (*ICON*'s identical clean data)*		
PRewrite (Kong et al., 2024)	$+1.22$	$+1.87$
ReQueR (Zhou et al., 2026)	$+2.60$	$+2.22$
Advisor (reactive) (Asawa et al., 2026)	-2.70	$+0.45$
Advisor (blind) (Asawa et al., 2026)	-2.42	-4.02
ICON Q-cond (control, question only)	$+1.65$	$+2.12$
ICON QO-cond (control, no a_{ori})	$+3.62$	$+4.24$
ICON QOA-cond (ours)	$\mathbf{+8.46}$	$\mathbf{+8.54}$

Table 3: Benchmark-specific ICON rewriters with a frozen Qwen3-VL-2B answerer. Base denotes the original accuracy. Results are mean changes in percentage points over three seeds.

Benchmark	Base	4B change	8B change
Robo2VLM (Chen et al., 2025)	48.8	$+8.46$	$+8.54$
CV-Bench (Tong et al., 2024)	72.8	$+4.91$	$+5.81$
LingoQA (Marcu et al., 2024)	45.8	$+2.67$	$+4.00$
ScienceQA (Lu et al., 2022)	83.7	$+1.77$	$+1.69$
MMStar (Chen et al., 2024)	44.3	$+2.17$	$+1.11$
MMMU (Yue et al., 2024)	41.1	$+1.02$	$+1.10$
BLINK (Fu et al., 2024)	43.9	$+1.40$	$+0.82$
VSR (Liu et al., 2023a)	73.5	$+0.47$	$+0.20$

5.4 Transfer across Frozen Answerers

We next keep a Robo2VLM rewriter fixed while varying the frozen answerer. The rewriter is trained using data collected from Qwen3-VL-2B, but at transfer time it receives each new answerer's own initial response. This setting evaluates transfer across answerers on the same task. Table 4 reports results for all evaluated answerers with both retained rewriter backbones. Outside the original setting, gains range from $+0.10$ to $+5.54$ points. The rewriter transfers across Qwen3-VL model scales, to other model families including Phi-3.5-Vision, Idefics3, and Molmo, and to three commercial APIs. Phi-3.5-Vision shows the largest transferred gains. This transfer setting differs from training a new rewriter for each target answerer. For example, supervised adaptation to Qwen3-VL-8B yields a gain of $+2.95$ points, whereas applying the original rewriter unchanged yields gains of $+0.97$ and $+1.57$ points for the 4B and 8B rewriter backbones, respectively.

6 Understanding When and Why Reframing Helps

6.1 Selective Rewriting without Correctness Labels

At deployment, the rewriter does not know whether the initial response is correct. Yet when we retrospectively group the 1335 Robo2VLM evaluation items by initial correctness, the 8B rewriter changes 71.2% of initially incorrect questions but only 6.9% of initially correct questions, a $10.3\times$ difference. This behavior is consistent with the role of identity targets, which encourage the rewriter to preserve questions that already produce correct responses during training. Importantly, this retrospective pattern does not imply that the rewriter can identify incorrect responses at deployment. Fixed transformations further illustrate why selective rewriting matters. As is shown in Table 2, we apply lexical rewriting, premise clarification, and perceptual grounding separately to every item and

Table 4: Robo2VLM accuracy changes when fixed rewriters are applied to other frozen answerers. Base is absolute accuracy in percent. Rewriter columns are percentage point changes. The 2B row averages three seeds.

Axis	Answerer	Base	4B change	8B change
Scale	Qwen3-VL-2B (Bai et al., 2025)	48.8	+8.46	+8.54
	Qwen3-VL-4B	53.9	+2.40	+1.95
	Qwen3-VL-8B	54.7	+0.97	+1.57
	Qwen3-VL-32B	60.4	+1.50	+1.20
Model family	Phi-3.5-Vision (Abdin et al., 2024)	25.0	+5.54	+4.72
	Idefics3-8B (Laurençon et al., 2024)	33.7	+1.95	+2.32
	Molmo-7B-O (Deitke et al., 2024)	23.7	+2.10	+2.25
Commercial API	Claude Opus 4.8 (Anthropic, 2026)	43.0	+2.24	+2.19
	Gemini 3.5 Flash (Google DeepMind, 2026)	45.5	+1.65	+0.67
	GPT-5.5 (OpenAI, 2026)	42.4	+0.10	+0.47

find that they change accuracy by -1.65, -4.42, and -3.97 percentage points, respectively. Although these transformations correct some initial errors, they can also break responses that were initially correct, creating a fix-and-break trade-off. For example, perceptual grounding corrects 130 initially incorrect items but breaks 183 initially correct ones. If originally correct answers are protected using a label-aware upper bound, the corresponding gains become $+4.42$, $+2.85$, and $+9.74$ points. These results suggest that effective reformulation requires both useful rewrite content and an appropriate decision about when to preserve or revise the original question.

6.2 What Changes in Successful Rewrites?

We next examine the content of successful reformulations. Our analysis of Robo2VLM assigns 47.9% of successful rewrites to perceptual grounding and 43.1% to premise explicitation. Perceptual grounding restates the requested attribute in a form that is more directly connected to the visual input, while premise explicitation makes an implicit task condition explicit. Together, these categories account for approximately 91% of the labeled successful rewrites.

These categories characterize recurring patterns in the rewritten questions rather than establish a causal mechanism, and their exact proportions depend on the labeling procedure. ICON uses a text-only rewriter. We additionally evaluate whether providing the image directly to the rewriter improves performance. We strictly follow the same setting in Table 3. With frozen Idefics3-8B answerers, image-aware rewriters achieve gains of $+2.10/+2.92$ points, compared with $+1.95/+2.32$ points for their text-only counterparts. With Molmo-7B-O, the corresponding gains are $+2.47/+3.45$ and $+2.10/+2.25$ points. Direct image access therefore provides an additional $+0.15$ to $+1.20$ points in these settings. These results suggest that the target model's initial textual response already provides useful context for reformulation, while direct visual access offers a modest additional benefit.

7 Discussion

From Best-of-K Potential to Deployable Performance. The prompt sensitivity analysis and the deployed method address different questions. Best-of-K accuracy measures whether at least one sampled formulation produces a correct answer when multiple formulations can be evaluated using the gold answer. It is therefore a label-aware upper bound over the sampled formulations rather than an achievable inference-time score. Standard ICON, by contrast, generates a single reformulation without access to the gold answer. Its net gain depends both on producing a useful reformulation and on preserving questions that should not be changed.

The fixed-transformation experiments illustrate this distinction. A transformation may correct a substantial number of original errors while reducing overall accuracy when applied indiscriminately. Reporting best-of-K accuracy together with single-rewrite performance and fix-versus-break behavior prevents potential improvement from being mistaken for deployable performance. More broadly, ICON yield positive mean gains across eight benchmarks, and one fixed Robo2VLM rewriter transfers across multiple frozen answerers, including different model families and commercial APIs.

Scope and Deployment. The answerer-transfer study keeps the Robo2VLM task and rewriter fixed while varying the frozen answerer. It therefore evaluates portability across model scales, model families, and commercial APIs on the same task. The rewriter also receives each new answerer's own initial response instead of reusing the response produced by Qwen3-VL-2B. The positive results show that one task-specific rewriter can remain useful across several frozen answerers. ICON requires one initial target-model call, one text-rewriter call, and one second target-model call. It uses no gold answer, target-model update, or candidate search at deployment. Gold answers and benchmark-specific data are nevertheless required during offline pair construction.

Transfer across Diverse Multimodal Tasks. Beyond the per-benchmark in-domain setting, we train a *single* pooled rewriter on a fixed pool of six multimodal sources (Robo2VLM, SEED-Bench (Li et al., 2024a), A-OKVQA (Schwenk et al., 2022), ScienceQA, MathVista (Lu et al., 2024), MathVision (Wang et al., 2024), about 4.9K response-conditioned examples collected from a frozen answerer's own answers, with no gold label at test time), then apply it *unchanged* to a strong frozen answerer (Qwen3-VL-8B) on benchmarks it *never trained on*, scoring each on its *full* official test set. Because the rewriter is disjoint from these benchmarks there is no train/test leakage to avoid, so we evaluate the complete official set rather than a subset. Unconditional reframing alone does not transfer, since per-item help and hurt nearly cancel, but a base-anchored cross-rewrite consensus, a zero-label inference-time rule, converts it into a net gain on multiple-choice reasoning benchmarks: MMMU $+1.67/+2.26$, MMStar $+0.93/+1.24$, and BLINK $+0.88/+0.33$ (Qwen3-VL-4B/8B rewriter, three seeds, 0% answer-leak rate). This is a genuine cross-benchmark, cross-model transfer on complete official sets. We verify that a ICON rewriter that has never seen a benchmark can help at inference time, without access to test labels, when the target reasoning has room for improvement.

Limitations: Where Does Reframing Help Less? We examine settings in which ICON provides limited improvement. In additional evaluations, MMBench (Liu et al., 2024) is essentially unchanged, while GQA (Hudson & Manning, 2019) and A-OKVQA show negligible changes. MMBench and GQA place substantial demands on visual understanding and compositional reasoning, while A-OKVQA additionally requires external knowledge. These results suggest that reformulation may provide limited benefit when successful answering depends more heavily on visual understanding, reasoning, or external knowledge than on how the question is expressed. We further examine whether best-of-K accuracy, domain similarity, or original accuracy can help characterize where ICON is effective. Best-of-K accuracy measures whether at least one sampled formulation can produce a correct answer under label-aware selection, but it does not determine whether a learned rewriter can produce a useful formulation with a single rewrite at inference time. The existence of a better formulation and the ability of ICON to produce one are therefore related but distinct properties.

8 Conclusion

In this paper, we study a practical source of unreliability in vision-language models: the same frozen model can produce different answers to questions intended to express the same visual request. We introduce Inference-time response-CONditioned reframing (ICON), a lightweight rewriter that uses the model's initial response to generate a single revised question for a second pass through the same frozen VLM. At deployment, ICON keeps the target VLM frozen, requires no gold answer or candidate search, and adds only one rewriter call and one additional target-model pass. The results show that ICON produces positive mean gains across eight benchmarks, and a Robo2VLM-trained rewriter transfers across multiple frozen answerers on the same task.

These results show that a VLM's own response can serve as a powerful conditioning signal for targeted question reformulation, revealing that the language interface is itself a meaningful axis for improving the performance of a frozen model. Rather than treating the initial response as the endpoint of inference, ICON uses it to determine how the same request should be posed again based on the model and instance. This turns interaction into a lightweight form of adaptation: the model remains completely unchanged, while the interface adapts to the model's observed behavior. More broadly, our findings point to a complementary direction for improving multimodal systems beyond parameter updates and increasingly elaborate inference-time search. Some failures are not failures of capability but failures of elicitation: the desired answer may already be within the model's accessible behavior, yet remain unrecovered under the initial formulation. When this is the case, the right intervention is not to change the model but to ask it a better question.

Reproducibility Statement

The main text and appendix provide detailed definitions, supervision and data construction procedures, implementation settings, baseline and transfer results, rewrite analyses, and additional evaluations covering text answerability, boundary cases, cross-benchmark generalization, and stronger answerers. The target model remains frozen throughout all standard ICON evaluations.

AI Use Statement

AI tools were used to generate supervision labels for offline rewriter training and to assist with manuscript editing for readability and organization. They were not used to generate or alter the evaluation outputs or reported numerical results. The authors remain responsible for the experimental design, generated labels, analyses, claims, citations, and final manuscript.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

Anthropic. Introducing Claude Opus 4.8. `https://www.anthropic.com/news/claude-opus-4-8`, May 2026. Published May 28, 2026. Accessed September 2026.

Parth Asawa, Alan Zhu, Abigail O'Neill, Matei Zaharia, Alexandros G. Dimakis, and Joseph E. Gonzalez. How to train your advisor: Steering black-box LLMs with Advisor Models. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V. Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*, 2024.

Kaiyuan Eric Chen, Shuangyu Xie, Zehan Ma, Pannag Sanketi, and Ken Goldberg. Robo2VLM: Improving visual question answering using large-scale robot manipulation data. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2025.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, et al. Molmo and PixMo: Open weights and open data for state-of-the-art vision-language models. *arXiv preprint arXiv:2409.17146*, 2024.

Mingkai Deng, Jianyu Wang, Cheng-Ping Hsieh, Yihan Wang, Han Guo, Tianmin Shu, Meng Song, Eric P. Xing, and Zhiting Hu. RLPrompt: Optimizing discrete text prompts with reinforcement learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3369–3391, 2022.

Yihe Deng, Weitong Zhang, Zixiang Chen, and Quanquan Gu. Rephrase and respond: Let large language models ask better questions for themselves. *arXiv preprint arXiv:2311.04205*, 2023.

Mengfei Du, Binhao Wu, Zejun Li, Xuanjing Huang, and Zhongyu Wei. EmbSpatial-Bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL, Volume 2: Short Papers)*, pp. 346–355, 2024.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Chaoyou Fu et al. MME: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

Xingyu Fu et al. BLINK: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision (ECCV)*, 2024.

Yulu Gan, Ligeng Zhu, Dandan Shan, Baifeng Shi, Hongxu Yin, Boris Ivanovic, Song Han, Trevor Darrell, Jitendra Malik, Marco Pavone, and Boyi Li. FoundationMotion: Auto-labeling and reasoning about spatial movement in videos. *arXiv preprint arXiv:2512.10927*, 2025.

Gemini Team, Google. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Google DeepMind. Gemini 3.5 Flash: Model card. `https://deepmind.google/models/model-cards/gemini-3-5-flash/`, May 2026. Published May 19, 2026. Accessed September 2026.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Weize Kong, Spurthi Amba Hombaiah, Mingyang Zhang, Qiaozhu Mei, and Michael Bendersky. PRewrite: Prompt rewriting with reinforcement learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL, Volume 2: Short Papers)*, pp. 594–601, 2024.

Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: Insights and future directions. *arXiv preprint arXiv:2408.12637*, 2024.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. SEED-Bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024a.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. MVBench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024b.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics (TACL)*, 11, 2023a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023b.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. MMBench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision (ECCV)*, 2024.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems*, volume 35, pp. 2507–2521, 2022. doi: 10.52202/068431-0182.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*, 2024.

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. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. EgoSchema: A diagnostic benchmark for very long-form video language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Ana-Maria Marcu et al. LingoQA: Visual question answering for autonomous driving. In *European Conference on Computer Vision (ECCV)*, 2024.

OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

OpenAI. GPT-5.5 system card. `https://openai.com/index/gpt-5-5-system-card/`, April 2026. Published April 23, 2026. Accessed September 2026.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3419–3448, 2022.

Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. Rephrase, augment, reason: Visual grounding of questions for vision-language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-OKVQA: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision (ECCV)*, 2022.

Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with MATH-Vision dataset. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun Chen. Large language models as optimizers. In *International Conference on Learning Representations (ICLR)*, 2024.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9556–9567, 2024.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.

Yixiao Zhou, Dongzhou Cheng, Zhiliang Wu, Yi Yang, Yu Cheng, and Hehe Fan. One refiner to unlock them all: Inference-time reasoning elicitation via reinforcement query refinement. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 38957–38978, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.1807. URL https://aclanthology.org/2026.acl-long.1807/.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. Large language models are human-level prompt engineers. In *International Conference on Learning Representations (ICLR)*, 2023.

Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Yue Zhang, Neil Zhenqiang Gong, and Xing Xie. PromptRobust: Towards evaluating the robustness of large language models on adversarial prompts. *arXiv preprint arXiv:2306.04528*, 2023.

A REPRODUCIBILITY: DATA SPLITS AND EVALUATION PROTOCOL

All final results use each benchmark's official evaluation metric. The reported gains are averaged over three training seeds (42, 43, and 44) using the answer-leak-filtered rewriter, with results reported for both the Qwen3-VL-4B and Qwen3-VL-8B rewriter backbones.

A.1 TRAIN/TEST SPLITS.

Table 5 lists the training, evaluation, and total question counts for each benchmark in the main results (Table 3 in the main paper). For the seven multiple-choice/regression benchmarks, we construct a deterministic train/evaluation split using fixed seed 42. The evaluation fraction is determined by dataset size, ranging from approximately 20% for the larger Robo2VLM dataset to approximately 50% for the smaller MMMU dataset, with most benchmarks using 40–45%. This provides sufficient data for both rewriter training and evaluation. Rewriter training data are collected only from the training split, and all reported gains are computed exclusively on the disjoint evaluation split. We verify zero ID overlap between the two splits and that train+evaluation=total exactly. For each benchmark, the same training data are used for the 4B and 8B rewriter backbones. The data are collected using the frozen Qwen3-VL-2B answerer, and only the rewriter backbone differs. Thus, each benchmark uses one training split rather than a separate split for each rewriter size. LingoQA is the only exception. Its 3504 training questions come from the separate official LingoQA training set, which is disjoint from the 500-item official validation set used for evaluation. Therefore, its training and evaluation examples are not obtained by partitioning the same set.

Table 5: Per-benchmark training, evaluation, and total question counts for the main results in Table 3 in the main paper. For the seven multiple-choice/regression benchmarks, we use a deterministic train/evaluation split with fixed seed 42, zero ID overlap between splits, and train+evaluation=total. The rewriter is trained only on the training split, and the N reported in Table 3 in the main paper corresponds to the evaluation split. †LingoQA is the only exception and does not use a same-set partition. This is because it provides separate official training and validation sets, so its 3504 rewriter-training questions come from the official training set, with zero ID overlap with the 500-item official validation set used for evaluation. Its "Total" therefore denotes the full 500-item evaluation set rather than train+evaluation, since the 3504 training questions come from a separate training set.

Benchmark	Train	Test (eval N)	Total
Robo2VLM	5341	1335	6676
CV-Bench	1451	1187	2638
LingoQA	3504	500	500†
ScienceQA	1209	808	2017
MMStar	900	600	1500
MMMU	477	423	900
BLINK	1045	856	1901
VSR	728	494	1222

A.2 CROSS-BENCHMARK REWRITE TAXONOMY.

Table 6 summarizes the combined share of perceptual grounding and premise explicitation among successful rewrites across the eight benchmarks in Table 3 in the main paper. For each benchmark, we analyze all successful rewrites and report the number of labeled rewrites in the corresponding denominator. Each benchmark contains at least 116 such examples. The relative prevalence of rewrite types varies across tasks rather than following a single universal pattern. Perceptual grounding is more common on visually demanding tasks such as Robo2VLM, CV-Bench, BLINK, and VSR, while premise explicitation is more prominent on LingoQA, MMStar, and MMMU. ScienceQA instead shows a larger contribution from decomposition. Thus, the 91% combined share observed on Robo2VLM should not be interpreted as a general property of all benchmarks. We also compare two automatic labeling procedures, one rule based and one LLM based. Their agreement on the exact fine-grained category is moderate. The broader pattern is more stable: successful rewrites often clarify perceptual targets or make task conditions more explicit, while purely lexical changes are less dominant. We therefore treat the exact category proportions as descriptive rather than definitive.

Table 6: Rewrite taxonomy across the eight benchmarks in Table 3 in the main paper. The fourth column reports the combined share of successful rewrites categorized as perceptual grounding or premise explicitation. N denotes the number of successful rewrites included in the taxonomy after excluding cases that introduce answer information.

Benchmark	Task type	N	Grounding+premise	Most common categories
Robo2VLM	robot MCQ	1002	91.0%	perceptual 48 / premise 43
CV-Bench	spatial/count MCQ	116	49.1%	perceptual 44 / option-anchor 28
LingoQA	driving open QA	1396	83.5%	premise 56 / perceptual 28
ScienceQA	science MCQ	133	36.1%	decomposition 33 / premise 23
MMStar	reasoning MCQ	187	49.7%	premise 33 / lexical 20
MMMU	knowledge MCQ	156	59.6%	premise 50 / option-anchor 15
BLINK	perception MCQ	138	48.6%	perceptual 36 / option-anchor 34
VSR	spatial T/F	195	63.1%	perceptual 62 / polarity 18

B Prompt Sensitivity Evaluation

Candidate formulation sets. For item n, let q_{n0} denote the original benchmark question and let $\mathcal{P}_n = \{q_{n0}, q_{n1}, \ldots, q_{n(K-1)}\}$ contain K formulations in total. The set includes the original question q_{n0} and $K-1$ alternative formulations generated by the proposer. The proposer is instructed to preserve the requested judgment without adding visual hints or answer statements.

Answer change rate. Let $y_n(q)$ denote the normalized answer produced by M for item n under formulation q. We define the answer change rate as

$$\text{AnswerChangeRate} = \frac{100}{N} \sum_{n=1}^{N} \mathbb{1}[\exists\, q, q' \in \mathcal{P}_n : y_n(q) \neq y_n(q')]\,. \tag{5}$$

This measures the percentage of examples for which at least two formulations yield different answers.

Best-of-K accuracy. Let $c_n(q) \in \{0, 1\}$ indicate whether the generated answer under formulation q is correct according to the benchmark's evaluation metric. We define

$$\text{Accuracy}_{\text{Best-of-}K} = \frac{100}{N} \sum_{n=1}^{N} \max_{q \in \mathcal{P}_n} c_n(q). \tag{6}$$

Because the original question is included in $\mathcal{P}_n$, best-of-K accuracy is the accuracy obtainable if the best sampled formulation could be selected for each example. This quantity uses the gold answer to select among multiple formulations and is therefore a label-aware upper bound rather than a deployable result.

Error correction rate. Let $c_n(q_{n0})$ denote correctness under the original question and

$$N_{\text{err}} = \sum_{n=1}^{N} (1 - c_n(q_{n0}))\,. \tag{7}$$

We define the error correction rate as

$$\text{ErrorCorrectionRate} = \frac{100}{N_{\text{err}}} \sum_{n=1}^{N} (1 - c_n(q_{n0})) \max_{q \in \mathcal{P}_n \setminus \{q_{n0}\}} c_n(q). \tag{8}$$

This measures the percentage of examples answered incorrectly under the original question for which at least one alternative formulation yields a correct answer. We evaluate the entire Robo2VLM and LingoQA evaluation sets and present the results below.

Benchmark	N	Original	Best-of-K	Answer change rate	Error correction rate
Robo2VLM (Chen et al., 2025)	6676	51.8	74.5	73.2	47.0
LingoQA (Marcu et al., 2024)	500	45.8	69.0	43.0	42.8

C Rewriter Adaptation to Stronger Answerers

We also examine whether adapting the rewriter to stronger frozen answerers provides additional benefits. For Qwen3-VL-8B and Qwen3-VL-32B, we collect new training pairs using each answerer's own initial responses, train the rewriter with supervised fine-tuning, and optionally continue training with reinforcement learning, while keeping the answerer frozen and using no gold answer during evaluation. The resulting gains remain modest. Continued reinforcement learning on the 8B answerer can increase task accuracy, but this increase is accompanied by greater answerability after the original visual input is replaced with a black image, indicating that some of the apparent improvement may arise from additional textual cues in the revised questions. We therefore interpret the later reinforcement learning results cautiously and retain the more conservative supervised fine-tuning results as the primary comparison. Overall, adapting the rewriter remains useful with stronger answerers, but additional optimization does not consistently produce larger improvements that still depend on the original visual input.