

To Answer or Not to Answer? Reliability-Oriented Reinforcement Learning for Visual Language Models

Jade Chien^{1*†}, Julian Wu^{2†}, Leo Wei^{1†}, Dylan Chen²,
Garvin Guo², Qingpei Guo¹, Jianing Li^{1 ‡}

¹Ant Group

²School of Artificial Intelligence, University of Chinese Academy of Sciences

3195794910@qq.com

{qingpei.gqp, tiansuo.ljn}@antgroup.com

Abstract

Reinforcement learning with verifiable rewards (RLVR) has recently demonstrated substantial gains in reasoning capabilities of vision-language models (VLMs). However, existing RLVR paradigms rely on deterministic rewards that implicitly penalize uncertainty expression and abstention, causing VLMs to over-produce speculative answers and degrading reliability. Through quantitative analysis, we find that vanilla RLVR, while effective in improving reasoning accuracy, significantly impairs VLMs’ ability to appropriately abstain from unanswerable queries. To systematically address this, we introduce **Reliable-V-Bench**, a new benchmark designed to evaluate VLMs’ abstaining capacity on challenging and realistic unanswerable visual questions. We further propose **Reliable-V-RL**, a reliability-aware RL framework that incorporates unanswerable and ambiguous visual questions into training, guided by a fine-grained, rubric-based reward mechanism that encourages VLMs to jointly optimize reasoning accuracy and reliable judgment. Experimental results demonstrate that Reliable-V-RL substantially improves the reliability and abstention capability of VLMs without sacrificing performance on standard multimodal benchmarks.

1 Introduction

Recent years have witnessed the significant success of reinforcement learning with verifiable rewards (RLVR) (Wen et al., 2025) in enhancing the reasoning capabilities of visual language models (VLMs) (Chen et al., 2025; Huang et al., 2025). By leveraging algorithms such as group relative policy optimization (GRPO) (Shao et al., 2024), this paradigm incorporates chain-of-thought (CoT) (Wei et al., 2022) and automated verification mechanisms, yielding substantial gains in multimodal

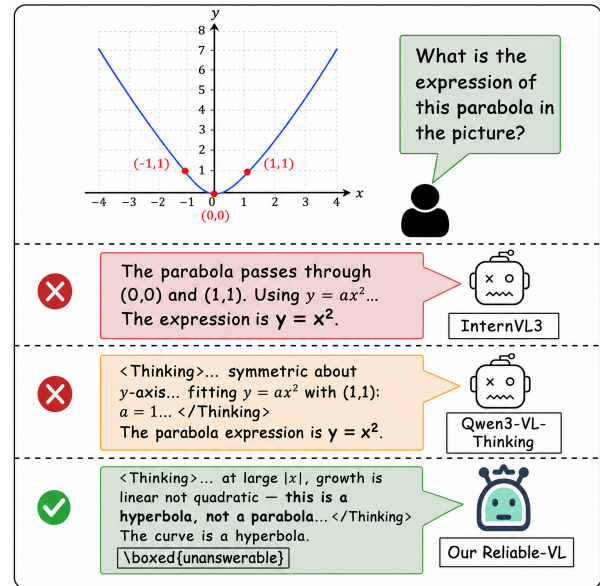


Figure 1: Existing VLMs often exhibit an answer-seeking bias, hallucinating responses to unanswerable questions even when faced with clear visual contradictions (e.g., misidentifying a "vehicle" as a "shoe"). To enhance reliability, our Reliable-VL is designed to detect such inconsistencies and appropriately abstain.

tasks (Zhang et al., 2024; Yue et al., 2024). Despite these advancements, the mainstream vanilla RLVR paradigm suffers from a critical limitation: its heavy reliance on deterministic, outcome-based reward. Specifically, this paradigm operates under the implicit assumption that all queries are answerable and inherently have a golden answer. Automated verifiers then assign binary rewards based solely on how VLMs’ final output matches these golden answers. However, this fails to account for realistic scenarios involving visual contradictions or information ambiguity (Jian et al., 2025), a typical example is illustrated in Figure 1. This reward mechanism of vanilla RLVR paradigm implicitly penalizes VLMs for expressing uncertainty or opting to abstain when encountering such unanswerable queries. Consequently, driven by the objective

*This work is done during internship at Ant Group.

†Equal contribution.

‡Corresponding Author.

of maximizing rewards, VLMs tend to speculatively generate plausible-sounding but misleading responses. This ultimately compromises the reliability of VLMs in realistic settings involving unanswerable questions—a concern that has increasingly drawn the community’s attention (Xue et al., 2025b; Yan et al., 2025).

Consistent with the previous discussion, we conducted a detailed analysis of vanilla RLVR and found that it severely impairs VLMs’ ability to abstain from unanswerable questions. Specifically, to assess its reliability, we follow Miyai et al. (2025) by constructing multimodal unanswerability by shuffling images and questions. Results reveal that after vanilla RLVR training, VLMs show a surprisingly rapid decline in ability to abstain. Even explicitly informed that abstaining is allowed, VLMs trained with vanilla RLVR still exhibit a strong "answer preference," leading to significant hallucinations when handling unanswerable questions.

Furthermore, existing attempts to assess robustness against unanswerable queries rely on simple heuristics—such as random image-question mismatching (Miyai et al., 2025; Zhang et al., 2023) or injecting non-existent visual concepts via LLMs (Wu et al., 2025; Zhang et al., 2025b). These automated conflicts suffer from significant semantic gaps; consequently, while effective for earlier models, they no longer pose a meaningful challenge to recent high-capacity VLMs. To address this limitation, we introduce Reliable-V-Bench, a diagnostic benchmark featuring realistic unanswerable questions designed to evaluate VLM abstention. Drawing on visual hallucination research (Liu et al., 2025; Rostamkhani et al., 2025), we leverage fine-grained annotations of hallucination-prone entities to construct questions that explicitly conflict with visual evidence (Figure 3). By grounding unanswerability in these human-annotated illusions, our benchmark creates highly deceptive, real-world cross-modal contradictions.

To overcome the aforementioned limitations of vanilla RLVR, we propose Reliable-V-RL, an RLVR framework integrated with a reliability-oriented reward mechanism. Our approach expands the training objective by incorporating automatically synthesized unanswerable visual questions, incentivizing the model to prioritize reasoning precision while penalizing hallucinations. Furthermore, by training on ambiguous queries, Reliable-V-RL explicitly models the decision boundary between answerable and unanswerable contexts, en-

hancing the model’s discriminative capacity. To prevent models from exploiting simple refusal shortcuts, we employ an LLM-as-a-judge rubric reward that mandates the correct identification of visual contradictions, ensuring the model’s rejection is grounded in deliberative reasoning.

Experimental results on Reliable-V-Bench and other benchmarks show that Reliable-V-RL significantly enhances VLM reliability via robust abstention, effectively suppressing incorrect responses to unanswerable and ambiguous queries. Notably, this improvement does not compromise general capabilities, as Reliable-V-RL maintains competitive accuracy compared to vanilla RLVR.

2 Preliminaries

2.1 Problem Formulation

Unlike traditional approaches that solely optimize answer accuracy (Yang et al., 2025b; Huang et al., 2025), we propose to optimize the reliability of vision–language models (VLMs). The primary objective is to train VLMs to learn a function $f : \mathcal{X} \mapsto \mathcal{A}$, where $\mathbf{x} = (v, q) \in \mathcal{X}$ represents an image–question pair (Dancette et al., 2023). However, as discussed earlier, in real-world settings, the user’s visual query may be unanswerable due to ambiguity or contradictions between visual and textual information (Jian et al., 2025; Xue et al., 2025b). Therefore, when unanswerable questions exist, VLMs must instead learn a function $h : \mathcal{X} \mapsto \mathcal{A} \cup \{\text{UA}\}$ ("UA" means unanswerable), whose output space is expanded to allow the model to abstain from answering. The function h can be decomposed into two components: the answer generation function $f : \mathcal{X} \mapsto \mathcal{A}$ and an answerability judgment function $\phi : \mathcal{X} \mapsto \{0, 1\}$. Accordingly, VLMs’ final output is defined as:

$$h(\mathbf{x}) = \begin{cases} f(\mathbf{x}) & \text{if } \phi(\mathbf{x}) = 1 \\ \text{UA} & \text{if } \phi(\mathbf{x}) = 0 \end{cases}. \quad (1)$$

2.2 Preliminary Findings

In this section, we investigate the vanilla RLVR paradigm, specifically implemented by widely adopted GRPO. Following the established approach (Miyai et al., 2025), we construct unanswerable questions in multimodal contexts by shuffling images and questions to assess the VLMs’ capacity to abstain appropriately. The results indicate that before GRPO, VLMs exhibit an inherent capacity to abstain when relying solely on prompt-based

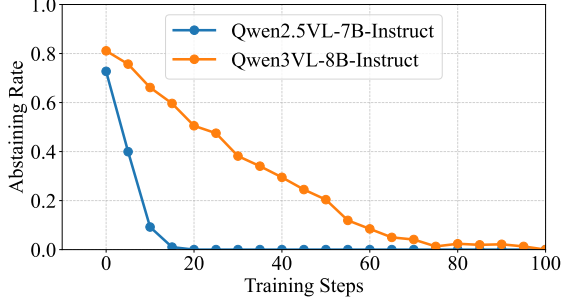


Figure 2: Even when explicitly prompted that VLMs may encounter unanswerable questions and are allowed to abstain, their ability to reject unanswerable questions rapidly decreases as the vanilla RLVR steps increase.

elicitation. This suggests that pre-training and instruction tuning already endow VLMs with a foundational degree of reliability. However, this capability rapidly degrades after several RL training steps. We attribute this phenomenon to the deterministic reward design in vanilla RLVR, which over-incentivizes the production of definitive answers. Consequently, when faced with potentially unanswerable or ambiguous visual questions, vanilla RLVR inadvertently undermines VLM’s reliability by penalizing honest abstention.

Besides, we find that recent VLMs like Qwen3-VL can effectively distinguish these synthesized unanswerable questions. Since simple image-question shuffling creates obvious semantic mismatches, it fails to provide a meaningful challenge, motivating us to develop a more rigorous and challenging benchmark in this issue.

3 Reliable-V-Bench

In this section, we introduce Reliable-V-Bench to evaluate VLMs’ reliability via challenging and realistic unanswerable questions. To simultaneously assess general task performance and answerability judgment, the benchmark incorporates diverse answerable samples from various multimodal datasets. Detailed statistics are provided in Appendix D. We demonstrate the construction of unanswerable visual questions below.

Constructing unanswerable questions that challenge existing VLMs and reflect real-world scenarios is difficult. If a textual question is entirely irrelevant to the accompanying image, current VLMs are generally capable of abstaining, and users are unlikely to pose purely non-visual questions. Therefore, we focus on creating unanswerable questions that mislead VLMs’ image understanding, posing a

more meaningful challenge to their reliability. For instance, as shown in Figure 3, the question targets an entity visually similar to an object in the image, a scenario more likely to occur in real-world settings due to misinterpretations of visual content.

Drawing insights from recent research on visual hallucination (Guan et al., 2024; Ding et al., 2024), we construct our dataset using the following approaches: (1) **Filtering and refining misleading visual questions:** Some cross-modal hallucination datasets emphasize questions that mislead VLMs’ visual understanding (Liu et al., 2025), aligning with our objectives. We filter out Yes/No questions and remove multiple-choice options, resulting in open-ended unanswerable questions. (2) **Human questioning on hallucination-prone images:** Some visual hallucination datasets (Rostamkhani et al., 2025; Zhang et al., 2025a) contain images with entities that are easily confused or misperceived. We recruit volunteers to freely pose questions about these images, naturally eliciting queries that target the confusable entities. (3) **Human filtering:** We then manually filter the collected questions, retaining only those that are genuinely unanswerable given the image content. Only samples that pass rigorous multi-annotator agreement on conflict validity, unanswerability, and realism are included in the final benchmark. Detailed construction and human evaluation processes are provided in Appendix A and B.

4 Reliable-V-RL

4.1 Training Data Curation

Unanswerable data construction. Automatically constructing unanswerable visual questions to improve VLM reliability is challenging, as unanswerability requires considering both visual and textual information in multimodal contexts. Inspired by prior work (Liu et al., 2025), we prompt VLMs to extract image attributes and then introduce pseudo-information (e.g., non-existent object names or attributes) that explicitly conflicts with the visual content. An LLM generates corresponding unanswerable visual questions based on this pseudo-information.

Crucially, this intermediate pseudo-information naturally forms a *Golden Contradiction*, a ground-truth statement detailing the exact cross-modal mismatch. We directly utilize this Golden Contradiction to synthesize fine-grained evaluation rubrics for our reward function. This closed-loop pipeline,

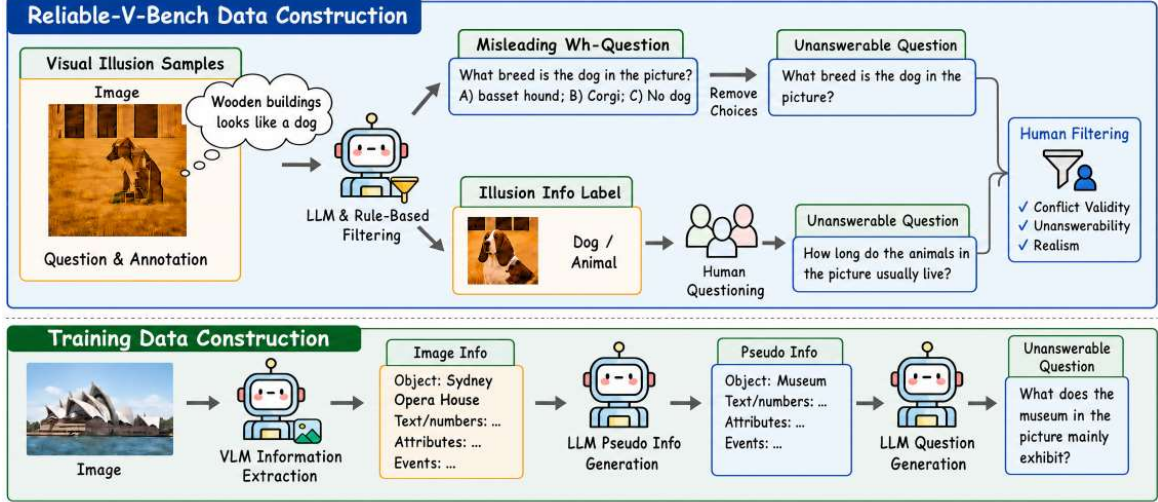


Figure 3: In the Reliable-V-Bench, the pipeline for constructing unanswerable visual questions that conflict with the image information, and the automatic construction method for unanswerable questions in the training set, which is used to enhance VLMs’ reliability.

where the data construction byproduct directly anchors the reward design, ensures rubric evaluation strictly targets the specific visual conflict. Unlike inference-time strategies (Khan and Fu, 2024) that require external proxy models, our design *internalizes* the abstention judgment into the reward mechanism during post-training, incurring zero additional inference cost.

Answerable and Ambiguous Data Collection. We sample ambiguous visual queries from ClearVQA (Jian et al., 2025) and answerable samples from diverse multimodal datasets (Schwenk et al., 2022; Saikh et al., 2022; Biten et al., 2019), ensuring source-alignment with our unanswerable and ambiguous subsets to maintain consistent domain distribution.

4.2 Reward Design

In vanilla RLVR implemented by GRPO, reward function is typically defined as $r(x, y) = r_a(x, y) + r_f(y)$, where $r_a(\cdot)$ and $r_f(\cdot)$ denote the accuracy reward and the format reward, respectively. However, to explicitly encourage reliability and prevent VLMs from learning trivial shortcuts, we augment training set into three subsets: $D = D_a \cup D_u \cup D_{amb}$, corresponding to answerable, unanswerable, and ambiguous questions. Our reliability-oriented reward comprises three components:

Refuse-unknown reward. For unanswerable inputs ($x \in D_u$), we introduce dedicated training data that explicitly conflicts with visual content, allowing the model to learn when abstention is

warranted. The reward for these samples is gated by whether the model correctly refuses: if VLM outputs an abstention signal, it proceeds to quality evaluation; otherwise, it receives zero reward. This component ensures that VLMs are exposed to and incentivized to handle genuinely unanswerable visual queries during RL training.

Answerability-boundary modeling. Inspired by prior work (Jian et al., 2025), we explicitly model the high-uncertainty boundary between answerable and unanswerable queries by introducing ambiguous visual questions ($x \in D_{amb}$). While Jian et al. (2025) focus on teaching VLMs to ask clarification questions without addressing the reward deficiency introduced by RLVR, our formulation directly rectifies the reward signal by assigning a differentiated partial reward β for abstention on ambiguous inputs (we set $\beta = 0.5$; a systematic ablation is provided in Appendix H). If VLM instead attempts to answer, the reward is computed using the accuracy term $r_a(x, y)$. By including ambiguous samples, VLMs reduce empirical risk and improve the separability of answerable and unanswerable instances (Yao and Zhou, 2010). Notably, the partial reward β for ambiguous questions is a **multimodal-specific** innovation that captures the state of “visually ambiguous yet potentially answerable,” which pure text-based methods cannot characterize.

LLM-judged refusal-quality reward. For unanswerable and ambiguous inputs ($x \in D_u \cup D_{amb}$), VLM is required to justify its abstention through a

transparent reasoning process. We utilize an LLM-guided reward $r_m(y) \in [0, 1]$ anchored by the *Golden Contradiction*, which serves as a semantic baseline to ensure VLM’s refusal is rooted in authentic cross-modal reasoning rather than surface-level heuristic learning. This evaluation framework mandates precise alignment with the visual reality, penalizing both hallucinations and verbose reasoning, while specifically devaluing generic, template-based responses. By prioritizing evidence-based conciseness, we force the VLM to pinpoint the exact cross-modal contradictions, thereby ensuring VLM’s reliability is grounded in actual multimodal perception.

Integrating these components, our unified reliability-oriented reward $r_{re}(x, y)$ replaces vanilla accuracy reward and computes final rewards for the sampled group (as illustrated in Figure 4). Specifically, let $p(y)$ denote the parsed final answer extracted from generated response y , and let UA (UnAnswerable) indicate a state where VLM explicitly refuses to answer. Utilizing the indicator function $\mathbb{I}[\cdot]$, the final reward is formulated as:

$$r_{re}(x, y) = \begin{cases} r_a(x, y) & \text{if } x \in D_a \\ \mathbb{I}[p(y) = \text{UA}] \cdot r_m(y) & \text{if } x \in D_u \\ \beta \mathbb{I}[p(y) = \text{UA}] \cdot r_m(y) & \\ + r_a(x, y) & \text{if } x \in D_{amb} \end{cases} \quad (2)$$

5 Experiments

5.1 Experimental Setup

Datasets. We incorporate ClearVQA (Jian et al., 2025) and our proposed Reliable-V-Bench to evaluate the VLMs’ reliability. We further include MMMU (Yue et al., 2024) and MMStar (Chen et al., 2024a) to assess VLMs’ performance on general multimodal tasks after Reliable-V-RL training.

Evaluation Metrics. For Reliable-V-Bench, we report: 1) **F1(A) and F1(U)**, representing F1 scores (Schuff et al., 2020) for identifying answerable and unanswerable queries, respectively; 2) **Acc(A)**, the accuracy on the answerable subset D_a ; and 3) **Reliability** $= w_1 \cdot R(U) + w_2 \cdot \text{Acc(A)}$, which balances the abstention rate $R(U)$ on unanswerable questions with answering accuracy on answerable ones¹. For ClearVQA, we evaluate: 1) **Abstention rate**, where higher rates are preferred given

¹We set $w_1 = w_2 = 0.5$ in our experiments.

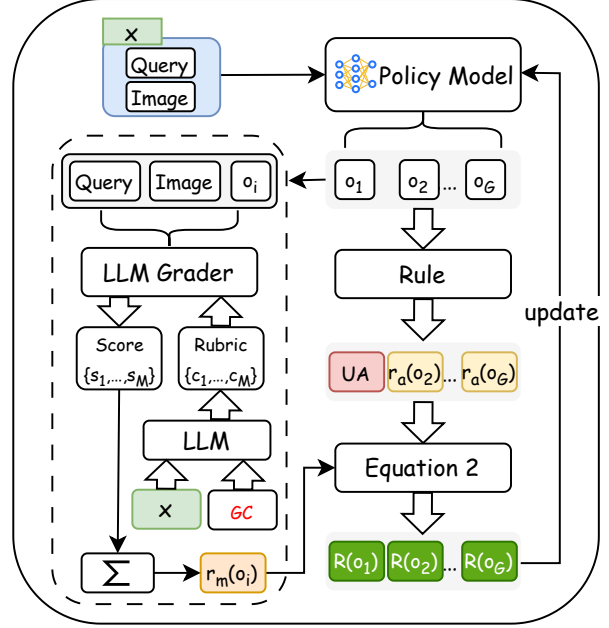


Figure 4: Reliability-oriented reward pipeline. We integrate rule-based accuracy (r_a) and LLM-rubric grading (r_m) to calculate the final reward, where GC refers to Golden Contradiction and UA denotes an unsolvable output state.

the difficulty of resolving intent under ambiguity; and 2) **Answered accuracy** (Cheng et al., 2024), assessing precision when VLM chooses to respond. Finally, we report standard Accuracy (Yue et al., 2024) for MMMU and MMStar.

Implementation. We train on 20K mixed samples using the VERL framework on 8 H100 GPUs; see Appendix C for complete hyperparameter and dataset split details.

Notably, open-source VLMs tend to prefer providing direct answers. Therefore, these existing VLMs are prompted that the given question may be unanswerable and allow abstaining, to maximize their potential reliability, for a fairer comparison.

5.2 Main Result

This section presents experimental results of the proposed Reliable-V-RL compared to existing VLMs. Our experiments demonstrate that incorporating GRPO with reliability-oriented reward substantially improves VLM’s reliability when unanswerable questions are present.

Results in Table 1 reveal that while open-source VLMs possess a baseline capacity for identifying unanswerability, they struggle with the answerable subset of Reliable-V-Bench. Notably, Qwen3-VL-Thinking (Bai et al., 2025) and Kimi-VL-Thinking (Team et al., 2025) consistently outperform their

Model	Reliable-V-Bench				ClearVQA		VizWiz	
	F1(A)	F1(U)	Reliability	Acc(A)	Recall	Answered Accuracy	Recall	Acc(A)
Close-Source Visual Language Models								
Gemini-2.5-Flash	85.9	75.2	67.3	63.2	38.3	45.8	82.3	73.4
GPT-5	87.1	76.5	70.1	66.4	39.6	47.3	85.1	75.3
Open-Source Visual Language Models								
Qwen2.5VL-7B	70.6	61.1	47.3	36.3	17.9	31.2	72.2	54.4
Qwen2.5VL-32B	75.8	56.6	42.8	43.3	20.5	21.9	74.3	64.8
InternVL2.5-8B	68.7	53.4	39.2	34.3	10.1	20.8	59.7	57.6
InternVL2.5-38B	68.8	56.2	43.5	44.1	14.8	22.4	68.5	64.4
Kimi-VL-A3B-Instruct	70.2	57.7	46.1	48.5	13.5	21.3	-	-
InternVL3-8B	70.8	39.0	35.8	44.7	14.4	24.2	61.4	60.9
InternVL3-38B	77.4	64.4	56.1	59.9	24.1	31.7	74.9	68.1
Qwen3VL-8B-Instruct	74.8	61.8	49.7	38.0	18.5	31.4	72.9	59.3
Qwen3VL-30B-A3B-Instruct	80.4	63.2	53.4	54.9	20.5	21.9	76.5	65.8
Open-Source Visual Reasoning Models (GRPO-Based)								
Qwen3VL-8B-Thinking	77.8	62.1	53.1	56.7	11.4	29.9	78.1	67.3
Qwen3VL-30B-A3B-Thinking	76.9	63.5	54.4	57.4	22.3	36.3	78.4	68.1
Kimi-VL-A3B-Thinking	72.7	60.9	52.6	56.2	23.0	33.9	-	-
Vision-R1-7B	68.4	5.1	25.1	47.7	2.1	21.4	6.1	62.4
R1-Onevision-7B	70.0	22.8	42.0	27.1	6.1	20.8	18.3	42.2
Our Models								
Reliable-Qwen-2.5-VL-7B	83.6	76.9	70.1	70.6	40.5	52.3	84.6	74.8
Reliable-Qwen-3-VL-8B	86.0	78.5	72.0	72.7	41.0	52.6	86.9	76.4
Reliable-Qwen-3-VL-30B-A3B	87.5	80.2	74.2	73.7	48.3	57.4	87.5	77.6

Table 1: The reliability performance of our model, trained using the proposed Reliable-V-RL method, on Reliable-V-Bench and ClearVQA benchmarks, compared with widely used open-source VLMs and visual reasoning models. For all open-source models, their experimental results on both benchmarks are based on the reliable prompt (which suggests that the input question may be unanswerable and allows for abstaining).

instruction-tuned counterparts, validating that enhanced reasoning directly bolsters reliability. Conversely, Vision-R1-7B (Huang et al., 2025) and R1-OneVision-7B (Yang et al., 2025b) show limited performance; as analyzed in Sec 2, their reliance on accuracy-only rewards induces an "overconfident" bias that systematically undermines reliability.

Based on Reliable-V-RL, our models achieve substantial performance gains over their respective base models. The 7B/8B models even outperform much larger models, while the 30B variant further pushes state-of-the-art reliability (74.2, surpassing GPT-5’s 70.1). These results demonstrate that Reliable-V-RL effectively improves VLMs’ reliability by enhancing accuracy on answerable questions while strengthening the ability to refuse unanswerable queries. The consistent improvements across scales (7B, 8B, 30B) confirm that the “overconfidence” problem induced by vanilla RLVR persists at larger scales, and our method remains effective regardless of model size. Furthermore, evaluation on VizWiz, an OOD dataset fea-

turing naturally occurring unanswerable questions posed by visually impaired users, proves successful generalization of robust abstention capabilities.

5.3 Ablation Study

We conduct a systematic ablation of Reliable-V-RL to evaluate the contribution of each reward component (Table 2).

Effectiveness of refuse-unknown reward and answerability-boundary modeling. Compared to vanilla RLVR, which optimizes only for accuracy, our full framework achieves a better balance between reasoning and abstention. As analyzed in Sec 2, vanilla RLVR’s deterministic reward design over-incentivizes definitive answers, causing a rapid decline in reliability. We further examine a variant trained without the refuse-unknown reward and answerability-boundary modeling (i.e., removing unanswerable and ambiguous subsets). Although we assign an intermediate reward of 0.5 to abstaining, removing these components leads

Model	Reliable-V-Bench		ClearVQA		VizWiz	
	Reliability	Acc(A)	Recall	Answered Accuracy	Recall	Acc(A)
Qwen2.5-VL-7B-Instruct	47.3	36.3	17.9	39.2	63.3	54.4
w/o Reliable Prompt	32.5	56.3	-	45.2	42.2	64.7
+ vanilla RLVR	37.1	70.5	-	40.0	-	74.2
+ Reliable-V-RL (w/o refuse-unknown rew.)	56.3	65.3	35.2	46.1	73.8	72.9
+ Reliable-V-RL (w/o answerability-boundary m.)	62.7	65.9	37.2	48.1	76.3	73.5
+ Reliable-V-RL (w/o refusal-quality rew.)	66.9	70.4	37.8	50.8	81.9	74.5
+ Reliable-V-RL	70.1	70.6	40.5	52.3	84.6	74.8

Table 2: Ablation study analyzing the effectiveness of various reward components in Reliable-V-RL. “-” indicates near-zero values. *Abbreviations*: **rew.** denotes reward; **m.** denotes modeling.

Model	Reliable-V-Bench		ClearVQA		VizWiz	
	Reliability	Acc(A)	Recall	Answered Accuracy	Recall	Acc(A)
Qwen2.5-VL-7B-Instruct (Reliable Prompt)	47.3	36.3	17.9	31.2	63.3	54.4
+ SFT (Full-Parameter)	53.2	43.0	7.7	33.8	54.1	68.2
+ SFT + DPO (Full-Parameter)	57.4	45.3	24.5	37.0	75.8	70.9
+ Reliable-V-RL	70.1	70.6	40.5	52.3	84.6	74.8

Table 3: The reliability performance comparison of the proposed Reliable-V-RL with SFT and DPO baselines under fair training conditions (full-parameter training on 8 GPUs, matching data volume and training epochs).

to substantially lower reliability. This confirms that training solely on answerable data forces the VLM into a **trade-off** between learning to refuse and improving accuracy. By incorporating the refuse-unknown reward, Reliable-V-RL successfully **decouples** these two objectives, enabling simultaneous enhancement. Furthermore, removing answerability-boundary modeling alone leads to degraded boundary discrimination, proving that the partial reward β for ambiguous queries acts as a “soft margin” to model high-uncertainty decision boundaries.

Impact of refusal-quality reward. Replacing our LLM-judged refusal-quality reward (i.e., r_m) with a simple binary abstention reward causes a noticeable drop in both Reliability and Acc(A). Without refusal-quality constraints, the RL process is highly susceptible to **reward hacking**: VLM learns a superficial text-only shortcut, generating generic refusals like “I cannot answer”, without a genuine perception of visual contradictions. Our rubric-based refusal-quality reward, anchored by the Golden Contradiction, ensures VLMs’ refusal is grounded in deliberative reasoning and actual cross-modal alignment.

5.4 Analyses

Online RL vs. Off-policy Methods. We compare Reliable-V-RL with SFT and DPO under fair con-

ditions (**full parameters, 8 GPUs, matched data and epochs**). As Table 3 shows, Reliable-V-RL substantially outperforms both: +12.7 Reliability over SFT+DPO (70.1 vs. 57.4) and +25.3 Acc(A) (70.6 vs. 45.3). Notably, SFT achieves only 7.7% Recall on ClearVQA, lower than the base model’s 17.9%, indicating it actually *suppresses* abstention willingness. This stems from off-policy limitations: SFT and DPO rely on static, pre-collected supervision, but reliable abstention under unanswerability often lies at the boundary of VLM capabilities, where consistent gold supervision is difficult to procure. Unlike off-policy methods restricted to fixed outputs, our online RL utilizes reliability-oriented rewards to enable adaptive exploration, effectively expanding the VLM’s capability boundary. The framework also demonstrates robustness to reward model choice (Appendix G).

More than a reliability-performance trade-off. Beyond benchmark evaluations, we analyze Reliable-V-RL’s general multimodal performance (Figure 5). An ablated variant without the refuse-unknown reward and answerability-boundary modeling degenerates into a “refuse unknown” setting (Cheng et al., 2024), incentivizing over-refusal of answerable questions that compromises general utility. In contrast, the full Reliable-V-RL achieves superior reliability (Table 2) while maintaining competitive general accuracy by explicitly mod-

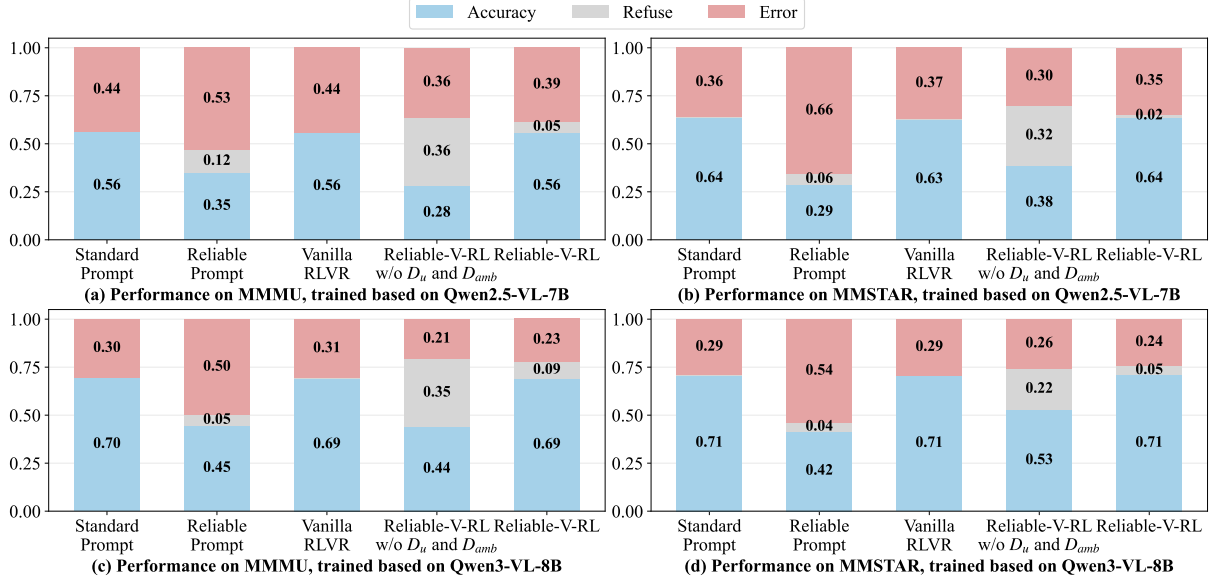


Figure 5: Performance of VLMs on general multimodal tasks after training based on Reliable-V-RL. By comparing Reliable-V-RL with vanilla RLVR and its ablation variants, we verify that Reliable-V-RL improves model reliability while preserving accuracy on general multimodal tasks.

eling unanswerable cases. Consequently, GRPO jointly optimizes reliability and performance without sacrificing general capabilities.

6 Related Work

RL for VLMs. RL has significantly improved reasoning capabilities in large language models (LLMs) (Shao et al., 2024; Yang et al., 2025a; Jaech et al., 2024), and recent studies have extended it to large VLMs to enable slow, deliberative reasoning for more accurate solutions (Tan et al., 2025; Chen et al., 2025; Huang et al., 2025). Most approaches rely on binary rewards (correct vs. incorrect) to encourage complex reasoning, but this treats abstaining or "I don't know" answers as incorrect (Song et al., 2025), pushing VLMs to always produce a definitive answer. This behavior undermines the reliability we claim. While prior work allows VLMs to avoid directly answering (Khan and Fu, 2024; Jian et al., 2025), to the best of our knowledge, no existing approach improves VLM reliability within RL, the commonly used post-training approach. We show that, with proper data curation, adding reliability-oriented rewards to RL allows VLMs to refuse unanswerable questions, a capability that declines under vanilla RL training.

Selective Question Answering. Recent studies emphasize enabling models to refuse unreasonable or unknown queries for better reliability (Srinivasan et al., 2024; Cheng et al., 2024). With

the increasing use of RL-based post-training methods like GRPO, richer reward formulations such as uncertainty-aware RL and multi-objective optimization for factual accuracy have been explored to improve LLMs' ability to appropriately abstain (Li et al., 2025; Lin et al., 2024; Xue et al., 2025a). In visual question answering (VQA), Whitehead et al. (2022) introduced selective prediction with a reject option. Khan and Fu (2024) proposed a selective prediction approach that rewrites questions using a proxy model and triggers abstaining when inconsistent outputs occur. While prior work focuses on rejecting unknown questions to improve accuracy, we study the reliability of VLMs when a visual question may be unanswerable, emphasizing robustness over accuracy.

7 Conclusion

The vanilla RLVR, the mainstream post-training method for enhancing reasoning, undermines VLMs' ability to abstain from unanswerable questions, compromising reliability. To systematically address this, we propose Reliable-V-Bench, a benchmark for evaluating VLM reliability on realistic unanswerable visual questions, and Reliable-V-RL, a reliability-oriented RLVR framework that improves reasoning accuracy while penalizing hallucinations induced by unanswerable questions. Experiments demonstrate that Reliable-V-RL significantly enhances VLMs' reliability and abstention

capability without sacrificing general performance, advancing the development of trustworthy VLMs for real-world applications.

Limitations

Due to the inherent weakness of modeling long-tail knowledge, VLMs may treat questions requiring unknown cross-modal knowledge as unanswerable. As a result, VLMs may still struggle to discern the subtle boundary between answerable and unanswerable questions, leading to incorrect judgments. Future work could focus on improving the integration of external knowledge to enhance VLM’s ability to more reliably distinguish this boundary.

Acknowledgments

We thank the anonymous reviewers for their valuable comments and suggestions. We are also grateful to our colleagues for helpful discussions and feedback, as well as for their assistance in validating and reviewing the benchmark data. Part of this research was conducted during an internship at Ant Group.

References

- Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.
- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Kebin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Ming-sheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibao Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. 2025. *Qwen3-vl technical report*. Preprint, arXiv:2511.21631.
- Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluís Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. 2019. Scene text visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4291–4301.
- Hardy Chen, Haoqin Tu, Fali Wang, Hui Liu, Xianfeng Tang, Xinya Du, Yuyin Zhou, and Cihang Xie. 2025. Sft or rl? an early investigation into training rl-like reasoning large vision-language models. *arXiv preprint arXiv:2504.11468*.
- Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. 2024a. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087.
- Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. 2024b. M3cot: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8199–8221.
- Qinyuan Cheng, Tianxiang Sun, Xiangyang Liu, Wenwei Zhang, Zhangyue Yin, Shimin Li, Linyang Li, Zhengfu He, Kai Chen, and Xipeng Qiu. 2024. Can ai assistants know what they don’t know? *arXiv preprint arXiv:2401.13275*.
- Corentin Dancette, Spencer Whitehead, Rishabh Maheshwary, Ramakrishna Vedantam, Stefan Scherer, Xinlei Chen, Matthieu Cord, and Marcus Rohrbach. 2023. Improving selective visual question answering by learning from your peers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24049–24059.
- Peng Ding, Jingyu Wu, Jun Kuang, Dan Ma, Xuezhi Cao, Xunliang Cai, Shi Chen, Jiajun Chen, and Shujian Huang. 2024. Hallu-pi: Evaluating hallucination in multi-modal large language models within perturbed inputs. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 10707–10715.
- Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. 2024. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385.
- Yunzhuo Hao, Jiawei Gu, Huichen Will Wang, Linjie Li, Zhengyuan Yang, Lijuan Wang, and Yu Cheng. 2025. Can mllms reason in multimodality? emma: An enhanced multimodal reasoning benchmark. *arXiv preprint arXiv:2501.05444*.
- Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. 2025. Vision-rl: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*.
- Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning

- and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.
- Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.
- Pu Jian, Donglei Yu, Wen Yang, Shuo Ren, and Jiajun Zhang. 2025. Teaching vision-language models to ask: Resolving ambiguity in visual questions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3619–3638.
- Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 235–251. Springer.
- Zaid Khan and Yun Fu. 2024. Consistency and uncertainty: Identifying unreliable responses from black-box vision-language models for selective visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10854–10863.
- Lei Li, Hehuan Liu, Yaxin Zhou, ZhaoYang Gui, Xudong Weng, Yi Yuan, Zheng Wei, and Zang Li. 2025. Uncertainty-aware iterative preference optimization for enhanced llm reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23996–24012.
- Sheng-Chieh Lin, Luyu Gao, Barlas Oguz, Wenhan Xiong, Jimmy Lin, Wen-tau Yih, and Xilun Chen. 2024. Flame: Factuality-aware alignment for large language models. *Advances in Neural Information Processing Systems*, 37:115588–115614.
- Yexin Liu, Zhengyang Liang, Yuezhe Wang, Xianfeng Wu, Feilong Tang, Muyang He, Jian Li, Zheng Liu, Harry Yang, Sernam Lim, et al. 2025. Unveiling the ignorance of mllms: Seeing clearly, answering incorrectly. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9087–9097.
- Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204.
- Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. 2022. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706.
- Atsuyuki Miyai, Jingkang Yang, Jingyang Zhang, Yifei Ming, Qing Yu, Go Irie, Yixuan Li, Hai Helen Li, Ziwei Liu, and Kiyoharu Aizawa. 2025. Unsolvability problem detection: Robust understanding evaluation for large multimodal models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6497–6540.
- Mohammadmostafa Rostamkhani, Baktash Ansari, Hoorieh Sabzevari, Farzan Rahmani, and Sauleh Eetemadi. 2025. Illusory vqa: Benchmarking and enhancing multimodal models on visual illusions. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2995–3004.
- Tanik Saikh, Tirthankar Ghosal, Amish Mittal, Asif Ekbal, and Pushpak Bhattacharyya. 2022. Scienceqa: A novel resource for question answering on scholarly articles. *International Journal on Digital Libraries*, 23(3):289–301.
- Hendrik Schuff, Heike Adel, and Ngoc Thang Vu. 2020. [F1 is Not Enough! Models and Evaluation Towards User-Centered Explainable Question Answering](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7076–7095, Online. Association for Computational Linguistics.
- Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pages 146–162. Springer.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.
- Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2024. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv:2409.19256*.
- Linxin Song, Taiwei Shi, and Jieyu Zhao. 2025. The hallucination tax of reinforcement finetuning. *arXiv preprint arXiv:2505.13988*.
- Tejas Srinivasan, Jack Hessel, Tanmay Gupta, Bill Yuchen Lin, Yejin Choi, Jesse Thomason, and Khyathi Chandu. 2024. [Selective “selective prediction”: Reducing unnecessary abstention in vision-language reasoning](#). In *Findings of the Association for Computational Linguistics ACL 2024*, pages 12935–12948. Association for Computational Linguistics.

- Huajie Tan, Yuheng Ji, Xiaoshuai Hao, Minglan Lin, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. 2025. Reason-rft: Reinforcement fine-tuning for visual reasoning. *arXiv preprint arXiv:2503.20752*.
- Jingqun Tang, Qi Liu, Yongjie Ye, Jinghui Lu, Shu Wei, An-Lan Wang, Chunhui Lin, Hao Feng, Zhen Zhao, Yanjie Wang, et al. 2025. Mtvqa: Benchmarking multilingual text-centric visual question answering. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7748–7763.
- Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. 2025. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- Xumeng Wen, Zihan Liu, Shun Zheng, Shengyu Ye, Zhirong Wu, Yang Wang, Zhijian Xu, Xiao Liang, Junjie Li, Ziming Miao, et al. 2025. Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base llms. *arXiv preprint arXiv:2506.14245*.
- Spencer Whitehead, Suzanne Petryk, Vedaad Shakib, Joseph Gonzalez, Trevor Darrell, Anna Rohrbach, and Marcus Rohrbach. 2022. Reliable visual question answering: Abstain rather than answer incorrectly. In *European Conference on Computer Vision*, pages 148–166. Springer.
- Yuanchen Wu, Lu Zhang, Hang Yao, Junlong Du, Ke Yan, Shouhong Ding, Yunsheng Wu, and Xiaoqiang Li. 2025. Antidote: A unified framework for mitigating lvlm hallucinations in counterfactual presupposition and object perception. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14646–14656.
- Boyang Xue, Fei Mi, Qi Zhu, Hongru Wang, Rui Wang, Sheng Wang, Erxin Yu, Xuming Hu, and Kam-Fai Wong. 2025a. Ualign: Leveraging uncertainty estimations for factuality alignment on large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6002–6024.
- Boyang Xue, Qi Zhu, Rui Wang, Sheng Wang, Hongru Wang, Minda Hu, Fei Mi, Yasheng Wang, Lifeng Shang, Qun Liu, et al. 2025b. Reliablemath: Benchmark of reliable mathematical reasoning on large language models. *arXiv preprint arXiv:2507.03133*.
- Qianqi Yan, Hongquan Li, Shan Jiang, Yang Zhao, Xinze Guan, Ching-Chen Kuo, and Xin Eric Wang. 2025. Hidden in plain sight: Reasoning in underspecified and misspecified scenarios for multimodal llms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24726–24746.
- Wenkai Yang, Shuming Ma, Yankai Lin, and Furu Wei. 2025a. Towards thinking-optimal scaling of test-time compute for llm reasoning. *arXiv preprint arXiv:2502.18080*.
- Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, et al. 2025b. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*.
- Yiyu Yao and Bing Zhou. 2010. Naive bayesian rough sets. In *International conference on rough sets and knowledge technology*, pages 719–726. Springer.
- Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2024. 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*, pages 9556–9567.
- Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. 2024. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pages 169–186. Springer.
- Yiming Zhang, Zicheng Zhang, Xinyi Wei, Xiaohong Liu, Guangtao Zhai, and Xiongkuo Min. 2025a. Illusionbench: A large-scale and comprehensive benchmark for visual illusion understanding in vision-language models. *arXiv preprint arXiv:2501.00848*.
- Yuwei Zhang, Chih-Hui Ho, and Nuno Vasconcelos. 2023. Toward unsupervised realistic visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15613–15624.
- Zongmeng Zhang, Wengang Zhou, Jie Zhao, and Houqiang Li. 2025b. Robust multimodal large language models against modality conflict. *arXiv preprint arXiv:2507.07151*.
- Yaowei Zheng, Junting Lu, Shenzhi Wang, Zhangchi Feng, Dongdong Kuang, and Yuwen Xiong. 2025. Easyrl: An efficient, scalable, multi-modality rl training framework.

A Prompt Template

In this section, we present detailed prompt templates for constructing unanswerable questions that deliberately conflict with the provided image information. These prompts are designed to pose substantial challenges to existing visual language models (VLMs), while remaining well grounded in realistic user scenarios where mismatches, misunderstandings, or incorrect assumptions about visual content naturally occur. By explicitly introducing conflicts between textual queries and visual evidence, the proposed prompts enable a systematic evaluation of a VLM's ability to recognize and handle unanswerable inputs.

Prompt for Generating Pseudo Image Information

Task:

Given structured textual descriptions of an image in JSON format, selectively generate pseudo information by replacing true image information with visually or semantically similar but incorrect alternatives.

Input:

A JSON object containing factual image information extracted from an image.

Instructions:

1. You may choose to generate pseudo information for one or more of the categories listed below.
2. If a certain category is not suitable or difficult for constructing plausible pseudo information given the image content, you may omit that category from the output.
3. Do not force pseudo information generation for all categories.
4. Pseudo information should be:
 - Visually or semantically similar to the original,
 - Plausible and easy to confuse,
 - Incorrect with respect to the image.

Selectable Pseudo Information Categories:

1. Object

Replace a real object with a visually similar but nonexistent or absent object.

2. Object Attributes

Replace the object's attributes (e.g., color, shape, pose) with similar but incorrect attributes.

3. Text

Substitute the text in the image with incorrect but visually similar alternatives.

4. Knowledge

Replace correct abstract, concrete, or professional knowledge related to the image with semantically related but incorrect knowledge.

5. Event / Activity

Replace a real activity, event, or interaction with a similar but incorrect event or interaction.

Output Format:

Return a JSON object containing only the selected pseudo information categories. If a category is not generated, do not include it in the output.

Example Structure:

```
{
  "object": "...",
  "object_attributes": "...",
  "text": "...",
  "knowledge": "...",
  "event_activity": "..."
}
```

Notes:

1. The output JSON may contain any subset of the fields above.
2. Do not include categories that are not applicable.
3. Do not generate any questions or answers.

Input Image Information:

...

Output:

Prompt for Generating Questions from Image Information

Task:

Given a JSON-formatted description of image information, generate questions based on each type of information provided. All questions should be directly grounded in the given information, without questioning its correctness.

Input:

A JSON object containing multiple categories of image information. Each field represents a specific type of information present in the image.

Core Principle:

Treat all information in the input JSON as true premises describing the image, and generate questions based strictly on these premises. The questions should be natural and coherent, relying explicitly on the provided information.

Instructions:

1. Use only the information contained in the input JSON to generate questions.

2. For each information field present in the input:

- Generate exactly one question.
 - The question must explicitly rely on the content of that field as its premise.
3. If a particular information type is not present in the input JSON, do not generate a question for it.
4. Do not evaluate, question, or negate the correctness of the given information.
5. Do not explicitly refer to the image or the source of the information (e.g., avoid phrases such as “in the image” or “according to the description”).
6. Ensure that questions are specific and unambiguous.

Information Type → Question Construction Logic:

1. Object

Generate a question assuming the existence of the object, such as asking about its properties, state, location, function, or related details.

2. Object_Attributes

Generate a question based on the given attributes, such as confirmation, comparison, or reasoning that depends on those attributes.

3. Text

Generate a question that relies on the given text, such as asking about its meaning, implications, or context, ensuring the question is grounded in the text content.

4. Knowledge

Generate a question based on the provided knowledge, such as its meaning, implications, background, or functional interpretation.

5. Event_Activity

Generate a question based on the described event, activity, or interaction, such as asking about actions, processes, or causal relations.

Output Format:

Return a JSON object in which each key corresponds to an information field from the input, and each value is the generated question.

```
{
  "object": "...?",
  "object_attributes": "...?",
  "text": "...?",
  "knowledge": "...?",
  "event_activity": "...?"
}
```

Notes:

- Exactly one question per input field.
- Do not include answers or explanations.

Note (for method description):

At this stage, the model is instructed to treat the input information as faithful image descriptions and to generate questions solely based on these premises, without performing any correctness verification.

Input Image Information:

...

Output:

To ensure the RLVR framework rewards genuine multimodal reasoning rather than superficial refusal shortcuts, we employ a rubric-guided LLM-as-a-judge mechanism. The foundation of this mechanism is the dynamic generation of fine-grained evaluation rubrics. Instead of relying on a static or simple binary reward, we prompt a strong LLM to construct specific rubrics based on the ground-truth **Golden Contradiction** of each unanswerable query. This forces the subsequent LLM evaluation to anchor strictly on the specific visual-textual mismatch.

The prompt used to synthesize these foundational rubrics—which the LLM judge later relies on to score VLM’s responses during RL—is provided below.

Prompt for Synthetic Rubrics Generation (VLM Reasoning with Abstention Mechanism)

You are an expert rubric writer specializing in Multimodal Vision-Language Model (VLM) evaluation, focusing on "Reasoning with abstention mechanism" where a question contradicts the visual evidence.

Task: Generate a self-contained set of evaluation criteria ("rubrics") to judge if a VLM correctly identifies the [Golden Contradiction] in the [Image] and abstain from answering directly.

Inputs:

- image: The visual input.
- question: A question containing a false premise or visual conflict.
- golden_contradiction: The ground-truth statement of the specific conflict.

Total items: Choose 4–8 rubric items.

Scoring & Weighting System:

- Item Score: Each item is scored between 0 and 1 (1 = criterion fully met,

0 = missed).

- Weights:

1. Essential (Positive, Sum = 0.50): Focus on whether the model identifies the *exact* visual mismatch in the [Golden Contradiction].
2. Important (Positive, Sum = 0.35): Ensure the model's reasoning allows for abstaining without creating secondary hallucinations.
3. Optional (Positive, Sum = 0.15): Reward brevity and directness.
4. Pitfall (Negative, Sum = -0.1 to -0.2): Target behaviors like "Generic abstention"

Each rubric item must include exactly three keys:

1. title: 2-4 words.
2. description: One sentence beginning with its category prefix.
3. weight: Use positive values for Essential/Important/Optional; use negative values for Pitfall.

Category Guidance & Examples:

- Essential Criteria: Check if the VLM pinpointed the exact visual mismatch.
- Example: "Essential Criteria: Explicitly states that the [Object A] mentioned in the question is not present in the image."
- Important Criteria: Check for "Faithfulness" (no secondary hallucinations).
- Example: "Important Criteria: Correctly identifies the actual items in the scene without attributing false properties to them."
- Optional Criteria: Check for "Conciseness."
- *Example: "Optional Criteria: Provides the refusal directly without unnecessary filler or repetitive reasoning."*
- Pitfall Criteria: Target behaviors like "Generic abstention" (refusing without specific reasons) or "Compensatory Hallucinations" (inventing reasons to justify refusal).
- Example: "Pitfall Criteria: Recommends an answer by hallucinating the missing object instead of refusing (Weight: -0.2)."
- Example: "Pitfall Criteria: Gives a generic refusal like 'I can't answer' without citing the specific visual conflict (Weight: -0.1)."

Output: Provide a JSON array of rubric objects. Each object must contain exactly three keys—title, description, and weight. Do not copy large blocks of the question or golden_contradiction into the text. Each description must begin with its category prefix, and no extra keys are allowed. Now, given the question, image and golden_contradiction, generate the rubric as described.

Generated Rubrics for VLM Reasoning with Abstention Mechanism

Inputs:

Image:



Question: What are the colors of the PlayStation controllers the women are holding?

Golden Contradiction: These are clearly Nintendo Wii Remotes, not PlayStation controllers

Generated Rubric (JSON Format):

```
[
  {
    "title": "Hardware Brand Contradiction",
    "description": "Essential Criteria: Explicitly identifies that the controllers being held are Nintendo Wii Remotes rather than PlayStation controllers.",
    "weight": 0.30
  },
  {
    "title": "Abstention from Color Assignment",
    "description": "Essential Criteria: Refuses to provide a color for 'PlayStation controllers' since those specific devices are not present in the image.",
    "weight": 0.20
  },
  {
    "title": "Visual Attribute Fidelity",
    "description": "Important Criteria: Correctly identifies the visual attributes of the actual objects present (e.g., shape, color) without attributing features specific to the nonexistent objects mentioned in the question.",
    "weight": 0.15
  },
  {
    "title": "Visual Faithfulness",
    "description": "Important Criteria: Correctly identifies the visual attributes of the actual objects present without hallucination of PlayStation-specific features.",
    "weight": 0.20
  },
  {
    "title": "Concise Refusal",
```

```

    "description": "Optional Criteria:
      Provides the correction and
      refusal directly without
      excessive introductory filler or
      repetitive explanations.",
    "weight": 0.15
  },
  {
    "title": "Generic Abstention",
    "description": "Pitfall Criteria:
      Provides a vague refusal such as
      'I can't answer' without
      specifying the brand conflict.",
    "weight": -0.10
  },
  {
    "title": "Compensatory Hallucination",
    "description": "Pitfall Criteria:
      Attempts to answer the question
      by imagining PlayStation
      controllers are hidden or by
      misidentifying the Wii Remotes as
      PlayStation hardware.",
    "weight": -0.10
  }
]

```

In addition, we provide a dedicated prompt that explicitly informs the VLM that the input question may be unanswerable and that abstaining from answering is an acceptable and desired behavior in such cases. This prompt is intended to encourage reliable model behavior by allowing VLM to refuse to answer when the question is ambiguous, inconsistent with the image, or lacks sufficient visual grounding, thereby better aligning model responses with real-world deployment requirements.

Prompt for instructing the VLM that abstention is allowed

et's reason about the question and the image step by step, and present the final answer within boxed{ }.

The reasoning process MUST be enclosed within <think> </think> tags. The final answer MUST be placed in boxed{ }.

If the question is unanswerable (e.g., when the image does not align with the question or the question is ambiguous), you may output boxed{unanswerable}.

Question:

B Data Quality Human Verification

To ensure the correctness, reliability, and real-world validity of the unanswerable VQA samples in Reliable-V-Bench, we incorporate human verifi-

cation as an essential component of the data curation pipeline. Since part of the benchmark is constructed by recruiting volunteers to pose questions on hallucination-prone images, human filtering is critical to guarantee that the resulting samples are genuinely unanswerable and reflect meaningful, realistic failure cases for visual language models (VLMs). Human verification plays a key role in filtering out trivial, implausible, or actually answerable samples and in ensuring the overall quality of the benchmark.

Our human data quality verification focuses on three core dimensions that are particularly important for evaluating model reliability under unanswerable scenarios:

- **Conflict Validity:** The constructed question should genuinely conflict with the visual content of the image. The unanswerability of the question must arise from an inconsistency between the question premise and the image information, rather than from vagueness, underspecification, or linguistic ambiguity.
- **Unanswerability Correctness:** Given the image, the constructed question should indeed be unanswerable. That is, a reasonable human annotator should not be able to derive a correct answer from the image, nor infer a plausible answer without making unsupported assumptions.
- **Realism:** The constructed question should be realistic and natural, reflecting the kinds of questions that users might reasonably ask in real-world scenarios due to misperception, incorrect assumptions, or partial understanding of the visual content.

Human evaluation During data construction.

During the initial stage of data synthesis, we perform a manual audit on a sampled subset of the generated unanswerable questions. We assess the quality of synthetic samples according to the standards mentioned before. If the quality falls short of our rigorous standards for unanswerability, the generation prompts and filtering heuristics are iteratively refined.

Human verification of Reliable-V-Bench. To further guarantee the integrity of the final test set and address potential annotation bias, we recruited 12 volunteers and implemented a multi-expert verification protocol. Each unanswerable visual ques-

tion in Reliable-V-Bench was independently reviewed by 3 annotators based on the criteria detailed in Appendix B. A sample is retained in the final benchmark only if it achieves unanimous consensus—meaning all three annotators must answer ‘Yes’ to all three quality dimensions (Conflict Validity, Unanswerability Correctness, and Realism). This rigorous filtering process ensures that every unanswerable question in Reliable-V-Bench is grounded in actual visual contradictions and reflects real-world reasoning challenges.

Questionnaire for Human Verification of Unanswerable VQA Samples

Image: <image>

Constructed Question: <Q>

For the given image and constructed question, please answer the following questions:

1) Conflict Validity. Does the constructed question rely on assumptions or premises that clearly conflict with the visual content of the image? (Yes / No)

2) Unanswerability. Given the image, is the question truly unanswerable, such that a reasonable person cannot determine a correct answer from the image without making unsupported assumptions? (Yes / No)

3) Realism. In real-world scenarios, is it plausible that a user could ask this question due to misinterpretation, incorrect assumptions, or incomplete understanding of the image? (Yes / No)

C Training Details and Hyperparameters

We validate Reliable-V-RL using Qwen2.5-VL-7B, Qwen3-VL-8B, and Qwen3-VL-30B-A3B-Instruct. The training set consists of 20K samples in total: 5K automatically generated unanswerable visual questions with contradictory text, 5K ambiguous questions from ClearVQA, and 10K answerable questions from various multimodal datasets.

The proposed Reliable-V-RL is trained using a GRPO-style online reinforcement learning paradigm implemented with the VERL (Sheng et al., 2024; Zheng et al., 2025) framework. Training is conducted for 5 epochs on 8 NVIDIA H100 GPUs, with a global batch size of 128 and a micro-batch size of 4, resulting in a rollout batch size of 512. For each update, the policy samples 16 rollouts per input. This enables a reliable estimation of relative advantages among candidate responses

Hyper-parameters	Value
Rollout Batch size	512
Global Batch Size	128
Micro Batch size	4
Warmup	False
Rollout	16
Rollout Temperature	1.0
Rollout Top-P	0.99
Freeze Vision Tower	True
KL divergence coefficient	1×10^{-2}
Learning rate	5×10^{-6}
Weight Decay	1×10^{-2}
GPUs	8
Optimizer	AdamW
Framework	Verl

Table 4: The hyperparameters used for implementing the proposed Reliable-V-RL.

Hyper-parameters	Value
Training steps	300
Batch size	64
Warmup ratio	0.03
Gradient accumulation	2
Learning rate scheduler	Cosine
Learning rate	2×10^{-4}
Projector learning rate	2×10^{-5}
Full-parameter	Yes
GPUs	8
Optimizer	AdamW

Table 5: The hyper-parameters used during SFT for comparing with the proposed Reliable-V-RL. Training is conducted with full-parameter updates (no LoRA) on 8 GPUs, with matched data volume and epochs.

and is critical for distinguishing between abstaining and overconfident answering. We follow standard practices for sampling and optimization to ensure stable training: we set the rollout temperature to 1.0 and apply top-p sampling with $p = 0.99$, a configuration widely used in RL-based post-training to maintain response diversity while avoiding excessive randomness.

Importantly, the vision tower is frozen throughout the entire RL stage. This design choice is consistent with our objective of improving answerability judgment and reasoning reliability, rather than altering low-level visual perception. Freezing the vision encoder also prevents spurious performance gains caused by visual overfitting, ensuring that improvements in abstaining behavior stem from the reliability-oriented reward design rather than representation drift.

Hyper-parameters	Value
Training steps	200
Batch size	32
Warmup ratio	0.03
Gradient accumulation	2
Learning rate scheduler	Cosine
Learning rate	5×10^{-6}
Full-parameter	Yes
GPUs	8
Optimizer	AdamW

Table 6: The hyper-parameters used during DPO for comparing with the proposed Reliable-V-RL. Training is conducted with full-parameter updates (no LoRA) on 8 GPUs, with matched data volume and epochs.

Datasets	Samples
GQA (Hudson and Manning, 2019)	$\sim 1\text{K}$
MTVQA (Tang et al., 2025)	$\sim 0.5\text{K}$
InfoVQA (Mathew et al., 2022)	$\sim 0.5\text{K}$
M3CoT (Chen et al., 2024b)	$\sim 4\text{K}$
MathVerse (Zhang et al., 2024)	$\sim 1\text{K}$
ScienceQA (Saikh et al., 2022)	$\sim 1\text{K}$
AI2D (Kembhavi et al., 2016)	$\sim 0.5\text{K}$
EMMA (Hao et al., 2025)	$\sim 1.5\text{K}$

Table 7: Composition of the answerable data used in Reliable-V-RL

Optimization is performed using AdamW with a learning rate of 5×10^{-6} and a weight decay of 1×10^{-2} . We do not apply learning rate warmup during RL training. To further address the tendency of vanilla RLVR to collapse toward overconfident answering, we incorporate a KL-divergence regularization term with a coefficient of 1×10^{-2} , constraining policy updates to remain close to the reference model. This regularization plays a crucial role in maintaining VLM’s pretrained abstaining capability while allowing reliable refinement through online RL. All hyperparameters for Reliable-V-RL are summarized in Table 4. Unless otherwise specified, the same configuration is used across all Reliable-V-RL experiments.

Evaluation Protocol. All evaluation results are averaged over multiple independent inference runs to ensure stability, as VLM generation with stochastic sampling can introduce variance, particularly for abstention-related metrics near the decision boundary. The same evaluation protocol is applied consistently across all models and baselines.

SFT and DPO Baselines. For comparison, we

implement SFT and DPO baselines under fair training conditions. Both baselines use **full-parameter training (no LoRA)** on **8 GPUs**, with data volume and training epochs matched to the SFT and RL stages of Reliable-V-RL. SFT is trained for 300 steps with a batch size of 64, cosine learning rate scheduling, and a warmup ratio of 0.03. DPO is performed after SFT for 200 steps with a batch size of 32, using a smaller learning rate of 5×10^{-6} . Full hyperparameter details are provided in Tables 5 and 6.

It is worth noting that **GPT-4o** (Hurst et al., 2024) is employed for "LLM Question Generation" in Reliable-V-Bench, as well as for "VLM Information Extraction," "LLM Pseudo Info Generation," and "LLM Question Generation" within the training pipeline (see Figure 3). Besides, **Gemini-3-Flash** is utilized to generate the Rubrics and to produce the LLM-guided rewards (grader model) based on these Rubrics and the VLM responses.

D Data Details

Diversity and data statistics. The Reliable-V-Bench contains a total of 1,233 samples, comprising 602 unanswerable and 631 answerable visual questions, resulting in a nearly balanced answerability distribution. The unanswerable subset is categorized according to the type of visual-textual conflict, including 247 object-level conflicts, 76 object attribute conflicts, 58 scene text conflicts, 124 knowledge-based contradictions, and 97 event or interaction conflicts. The answerable subset spans a broad range of VQA scenarios (Antol et al., 2015; Biten et al., 2019; Marino et al., 2019; Schwenk et al., 2022), including 268 general visual understanding questions, 79 knowledge-based questions, 149 reasoning-intensive questions, and 135 scene text understanding questions. The distribution of samples across answerability categories, conflict types, and VQA scenarios is illustrated in Figure 6. This composition ensures that Reliable-V-Bench captures diverse real-world failure modes while maintaining coverage of standard multimodal question-answering tasks, enabling a comprehensive evaluation of VLM reliability.

E Learning Dynamics of Reliable-V-RL

In this section, we analyze the learning dynamics of Reliable-V-RL compared to the traditional vanilla RLVR approach. The performance trends across various metrics are illustrated in Figure 7, where

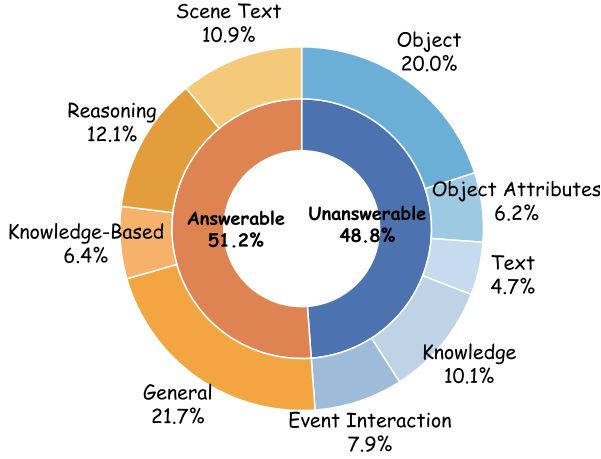


Figure 6: The data diversity in ClearVQA, illustrated by data distribution in the ClearVQA test set. ClearVQA focuses on three common categories of ambiguity. Each ambiguity category includes a variety of VQA scenarios, including general, knowledge, reasoning, and scene text.

we present the training curves for accuracy and abstaining rates on both unanswerable and ambiguous visual questions. As depicted in Figure 7, the curves of Reliable-V-RL exhibit significant improvements in model reliability compared to vanilla RLVR and the Ternary GRPO variants. Reliable-V-RL focuses not only on improving reasoning accuracy but also on reliable decision-making by encouraging VLM to abstain when a question is unanswerable or ambiguous. This is achieved by adjusting the reward structure to penalize hallucinations (incorrect responses) and reward abstention when necessary.

The first row of Figure 7 shows accuracy for answerable questions, demonstrating that Reliable-V-RL performs similarly to vanilla RLVR, maintaining competitive performance on tasks where a definitive answer is available. This suggests that Reliable-V-RL can achieve the same reasoning accuracy as existing methods while still improving reliability. In the second row, the abstaining rate for ambiguous questions is shown. Reliable-V-RL maintains a higher abstaining rate compared to vanilla RLVR, significantly reducing the likelihood of VLM attempting to answer ambiguous questions incorrectly. This is in stark contrast to Ternary GRPO and vanilla RLVR, which show much lower abstention in such cases, often leading to misleading answers when faced with ambiguity in the input. Finally, the third row of Figure 7 presents the abstaining rate for unanswerable questions. Here, Reliable-V-RL outperforms vanilla RLVR by a large margin. The VLM’s ability to

correctly abstain when it detects that a question is unanswerable, despite being prompted to respond, is a key improvement. vanilla RLVR, on the other hand, continues to produce answers even in cases where they are not valid, due to its reward function that implicitly discourages abstention.

Reliable-V-RL effectively decouples reasoning accuracy and abstention behavior, ensuring that VLM can reliably reject unanswerable and ambiguous questions without sacrificing performance on answerable ones. In contrast, vanilla RLVR tends to overfit to answer generation, even in scenarios where abstention would be more reliable, highlighting the shortcomings of traditional training paradigms that focus purely on correctness. The Ternary GRPO model, while showing a moderate improvement over vanilla RLVR, still fails to reach the reliability levels of Reliable-V-RL.

This analysis confirms that Reliable-V-RL’s reward design leads to an overall boost in reliability without sacrificing general task performance, making it a more robust and trustworthy model for real-world VQA scenarios.

F Supplementary Experiments

To further validate the robustness and generalizability of our proposed Reliable-V-RL, we provide additional ablation studies in this section.

Ablation on Qwen3-VL-8B. Table 8 presents the detailed ablation study based on the Qwen3-VL-8B model. Consistent with the main results on Qwen2.5-VL-7B (discussed in Section 5.3), the removal of the refuse-unknown reward or answerability-boundary modeling leads to a noticeable degradation in both the F1 scores for distinguishing answerability and the overall Reliability metric. Furthermore, applying the reliable prompt to the base instruction-tuned model significantly drops its answering accuracy ($Acc(A)$) from 61.1 to 38.0. This indicates that the reliable prompt induces an “over-refusal” behavior, causing VLM to conservatively reject many answerable questions. The performance gain achieved by Reliable-V-RL on $Acc(A)$ (recovering to 72.7) does not solely stem from enhanced general VQA capabilities, but largely from VLM learning a sharper, more accurate decision boundary to distinguish between answerable and unanswerable queries.

Impact of Open-Ended Generation Evaluation. As discussed by the community, open-

Model	Reliable-V-Bench				ClearVQA	
	F1(A)	F1(U)	Reliability	Acc(A)	Recall	Answered Accuracy
Qwen3-VL-8B-Instruct	74.8	64.8	50.2	38.0	18.5	31.4
w/o Reliable Prompt	66.4	8.3	36.6	61.1	-	37.1
+ vanilla RLVR	65.9	-	38.1	72.1	-	38.8
+ Reliable-V-RL (w/o refuse-unknown rew.)	78.5	70.4	66.9	69.0	34.4	45.8
+ Reliable-V-RL (w/o answerability-boundary m.)	80.1	72.9	68.4	70.5	36.2	47.9
+ Reliable-V-RL (w/o refusal-quality rew.)	85.7	76.8	70.3	72.1	38.1	51.0
Reliable-V-RL	86.0	78.5	72.0	72.7	41.0	52.6

Table 8: Ablation study analyzing the effectiveness of various designs in Reliable-V-RL, based on Qwen3VL-8B. *Abbreviations: **rew.** denotes reward; **m.** denotes modeling.*

ended VQA evaluation can sometimes be unstable, over-penalizing models that answer correctly but do not use the exact ground-truth wording. To ensure that our reliability improvements are not merely artifacts of the open-ended format, we conducted an ablation where all answerable questions in the training set are replaced with multiple-choice questions (MCQs), entirely removing the open-ended generation (OEG) samples. As shown in Table 9, even without OEG VQA samples during training (+ Reliable-V-RL w/o OEG VQA sample), VLM still effectively maintains high Reliability (69.5) and effectively abstains on ClearVQA. This confirms that Reliable-V-RL intrinsically improves VLM’s reliability and its fundamental capacity to judge cross-modal conflicts, independent of the VQA formulation.

G Alternative Reward Model Experiments

To validate that the performance gains of Reliable-V-RL are not dependent on a specific proprietary reward model, we replace the default Gemini-3-Flash (used for both rubric generation and LLM-as-a-judge reward scoring) with alternative models: the open-source Qwen3-Next-80B-A3B-Instruct and the stronger proprietary Claude-Sonnet-4.6. All other training configurations remain unchanged.

As shown in Table 10, using the open-source Qwen3-Next-80B-A3B-Instruct as the reward model yields metrics comparable to or slightly better than the original Gemini-3-Flash (Reliability: 70.4 vs. 70.1, Acc(A): 71.2 vs. 70.6). This demonstrates that: (1) our method is **robust to the choice of reward model**: open-source models can fully replace proprietary ones; (2) the core advantage stems from the Golden Contradiction design and the differentiated reward mechanism itself, rather than any specific LLM’s capability.

Furthermore, the stronger proprietary model Claude-Sonnet-4.6 brings additional improvements across all metrics (Reliability: 72.9, Acc(A): 73.4), confirming that our framework continuously benefits from advances in reward model capability, exhibiting good scalability.

H Sensitivity Analysis

We provide a systematic sensitivity analysis to address concerns about the transparency and robustness of our design choices.

H.1 Reward Consistency and Rubric Prompt Robustness

To validate the stability of our LLM-as-a-judge reward signal, we evaluate: (a) intra-model consistency via Intraclass Correlation Coefficient (ICC) from multiple runs of Gemini-Flash; (b) inter-model agreement via Spearman rank correlation across different scoring models and human annotators; (c) robustness to rubric prompt design variations (random item reordering and semantically equivalent rephrasing). All evaluations are based on 500 abstention outputs from the test set.

As shown in Table 11, our reward signal exhibits strong self-consistency (ICC = 0.91), good cross-model agreement (Spearman 0.73–0.82), and high robustness to prompt variations (Spearman 0.89–0.91). These results confirm that the rubric-based reward provides a reliable training signal regardless of specific prompt formulations or model choices.

H.2 Partial Reward β Ablation

The partial reward β assigned to ambiguous questions determines the strength of the “soft margin” between answerable and unanswerable decision boundaries. We conduct a systematic ablation of β to validate the choice of $\beta = 0.5$ and analyze failure modes at extreme values.

Model	Reliable-V-Bench				ClearVQA	
	F1(A)	F1(U)	Reliability	Acc(A)	Recall	Answered Accuracy
Qwen2.5-VL-7B-Instruct	70.6	61.1	47.3	36.3	17.9	31.2
+ Reliable-V-RL (w/o OEG VQA sample)	82.9	75.5	69.5	69.8	37.5	48.5
Reliable-V-RL	86.0	78.5	72.0	72.7	41.0	52.6

Table 9: Ablation experiment where all open-ended generation (OEG) answerable questions in the Reliable-V-RL training set are replaced with multiple-choice questions (MCQs) to mitigate potential evaluation instability.

Reward Model	Reliable-V-Bench		ClearVQA		VizWiz	
	Reliability	Acc(A)	Recall	Answered Accuracy	Reliability	Acc(A)
Gemini-3-Flash	70.1	70.6	40.5	52.3	84.6	74.8
Qwen3-Next-80B-A3B-Instruct (Open-Source)	70.4	71.2	41.0	53.1	85.1	75.0
Claude-Sonnet-4.6	72.9	73.4	46.4	56.3	86.7	75.9

Table 10: Reliable-V-RL performance with different reward models. The open-source Qwen3-Next-80B-A3B-Instruct achieves comparable results to Gemini-3-Flash, demonstrating full reproducibility without API costs. Stronger models (Claude-Sonnet-4.6) yield further improvements, confirming the framework’s scalability.

Metric	Value
ICC (Gemini-Flash self-consistency)	0.91
Claude-Sonnet-4.6 vs. Gemini-Flash (Spearman)	0.73
Qwen3-Next-80B vs. Gemini-Flash (Spearman)	0.82
Human vs. Gemini-Flash (Spearman)	0.78
Rubric item reordering vs. original (Spearman)	0.89
Semantic rephrasing vs. original (Spearman)	0.91

Table 11: Reward model consistency and robustness analysis. All metrics fall within or exceed the reference ranges commonly reported in LLM-as-a-judge literature (ICC: 0.85–0.95; human correlation: 0.75–0.90; cross-LLM: 0.70–0.85).

Results in Table 12 reveal three key findings:

- **Robustness within reasonable range:** Reliable-V-RL maintains strong Reliability (68.3–70.1) for $\beta \in [0.3, 0.7]$, demonstrating insensitivity to the precise value within this interval.
- **Extreme values cause failure:** $\beta = 0.05$ (near-zero reward for abstaining on ambiguous questions) reduces ClearVQA Abstaining Rate to 21.4%, below the base model’s 17.9% after RLVR, indicating that VLM learns to answer everything including truly ambiguous queries. Conversely, $\beta = 0.95$ encourages over-refusal, reducing Answered Accuracy to 46.5%.
- $\beta = 0.5$ **is the natural midpoint:** As the midpoint between answerable reward (1.0) and unanswerable base (0.0), $\beta = 0.5$ provides an intuitive and empirically optimal balance,

requiring no extensive hyperparameter search.

I Data Integrity and Evaluation Transparency

I.1 Separation of Test and Training Data Construction

To ensure evaluation integrity, we clarify the complete separation between the Reliable-V-Bench test set and the Reliable-V-RL training data:

- **Reliable-V-Bench (test set):** Unanswerable questions are constructed through two human-driven pipelines: (1) filtering existing misleading questions from hallucination datasets, and (2) recruiting volunteers to freely pose questions on hallucination-prone images, followed by rigorous multi-annotator verification. *No LLM is used for question generation in the test set.*
- **Reliable-V-RL (training set):** Unanswerable questions are automatically generated via a VLM information extraction $\rightarrow$ LLM pseudo-information generation $\rightarrow$ LLM question generation pipeline, where GPT-4o is employed for information extraction and question generation.

The two pipelines differ fundamentally in both methodology (human-crafted vs. LLM-generated) and source images (hallucination/illusion datasets vs. general VQA images). There is **zero overlap** at the image level, question level, or construction-method level between the test set and training set. GPT-4o’s role is strictly limited to the training data

Configuration	Reliable-V-Bench				ClearVQA	
	F1(A)	F1(U)	Reliability	Acc(A)	Abstaining Rate	Answered Acc.
$\beta = 0.05$	75.6	69.4	65.1	70.3	21.4	33.5
$\beta = 0.3$	82.4	75.0	68.3	70.1	36.7	51.0
$\beta = 0.5$ (default)	83.6	76.9	70.1	70.6	40.5	52.3
$\beta = 0.7$	82.0	75.4	68.4	69.8	38.5	50.3
$\beta = 0.95$	75.1	70.5	65.8	69.7	45.9	46.5

Table 12: Ablation of partial reward β for ambiguous questions. Reliable-V-RL is robust within the range $\beta \in [0.3, 0.7]$ (Reliability: 68.3–70.1). Extreme values ($\beta = 0.05$ or $\beta = 0.95$) degrade performance, confirming the non-triviality of differentiated reward design.

pipeline and has no involvement in Reliable-V-Bench construction. Consequently, GPT-4o’s evaluation on Reliable-V-Bench (Table 1) is free from any data contamination or circularity concerns.

I.2 Benchmark Quality Assurance

While Reliable-V-Bench contains 602 unanswerable samples (modest in scale compared to some automatically constructed benchmarks), we emphasize that *quality and difficulty* are prioritized over quantity:

- **Rigorous human verification:** Each sample was independently reviewed by 3 annotators (from a pool of 12 volunteers) across three quality dimensions: Conflict Validity, Unanswerability, and Realism. Only samples achieving *unanimous consensus* (all three annotators pass all dimensions) are retained, resulting in a conservative retention rate of 60.2%.
- **High deceptiveness:** Unlike automatically constructed benchmarks where semantic gaps are often obvious, our benchmark targets visually confusable entities that genuinely mislead state-of-the-art VLMs. As evidence, Qwen3-VL-8B-Instruct achieves only 61.8% F1(U) on Reliable-V-Bench, compared to near-perfect performance on simple image-question shuffling benchmarks.
- **Diverse conflict types:** The 602 unanswerable samples span 5 conflict categories (object: 247, attribute: 76, scene text: 58, knowledge: 124, event: 97), ensuring comprehensive coverage of real-world failure modes rather than concentration on a single pattern.

This quality-first design philosophy ensures that each sample in Reliable-V-Bench represents a genuinely challenging and realistic test of VLM reliability, providing more diagnostic value per sample than large-scale but noisy alternatives.

I.3 Reward Signal Stability Analysis

Beyond the summary statistics in Table 11, we provide additional context on the reliability of our LLM-as-a-judge reward signal during training.

Evaluation protocol. We sampled 500 abstention outputs from the Reliable-V-Bench test set (unanswerable subset) and scored each output using (a) Gemini-Flash run 5 times independently, (b) Claude-Sonnet-4.6, (c) Qwen3-Next-80B-A3B-Instruct, and (d) 3 human annotators. The ICC is computed from the 5 Gemini-Flash runs; Spearman correlations are computed between each pair of scoring sources on the same 500 outputs.

Interpretation of results. The ICC of 0.91 indicates “excellent” agreement according to standard benchmarks (Koo & Li, 2016: ICC > 0.90 is excellent). The human-vs-Gemini Spearman of 0.78 is comparable to inter-human agreement reported in prior LLM-as-a-judge studies (e.g., Zheng et al., 2024 report 0.72–0.81 for MT-Bench). The rubric prompt robustness (Spearman ≥ 0.89 under reordering/rephrasing) confirms that the reward is anchored on semantic content (the Golden Contradiction) rather than superficial prompt formatting.

Implications for training stability. High ICC (0.91) means that across different RL training epochs, the same model output receives consistent reward scores, preventing noisy gradient signals. The cross-model agreement (0.73–0.82) further validates that our reward captures genuine reasoning quality rather than idiosyncratic preferences of a single LLM judge. Combined with the alternative reward model experiments (Appendix G), these results establish that Reliable-V-RL’s training signal is stable, reproducible, and model-agnostic.

J Case Study

The case study section provides an in-depth analysis of the proposed Reliable-V-RL compared to state-of-the-art visual language models (VLMs)

in real-world scenarios. Figures 8-10 illustrate key examples that demonstrate the reliability and decision-making process of Reliable-V-RL when encountering unanswerable or ambiguous visual queries—data points that specifically challenge a VLM’s capacity for robust inference.

We compare Reliable-V-RL against competitive baselines, including InternVL3-78B and Qwen3-VL-235B-A22B, focusing on VLM’s ability to abstain from providing incorrect answers when faced with insufficient visual information. While baseline models often attempt to generate responses despite lacking necessary evidence, Reliable-V-RL successfully identifies and refrains from answering such queries.

For instance, in Figure 9, the question "What letters are in the image?" represents a classic text-visual conflict where the visual content contains no identifiable text. Reliable-V-RL correctly chooses to abstain, whereas InternVL3-78B produces a hallucinated response. This highlights Reliable-V-RL’s superiority in detecting conflicts between textual prompts and visual evidence, allowing it to take appropriate measures to avoid unfaithful generations.

In summary, these case studies showcase the exceptional performance of Reliable-V-RL across answerable, unanswerable, and ambiguous visual tasks. By integrating interactive clarification with a reliable decision mechanism, Reliable-V-RL emerges as a trustworthy model for practical applications where reliability is as critical as accuracy.

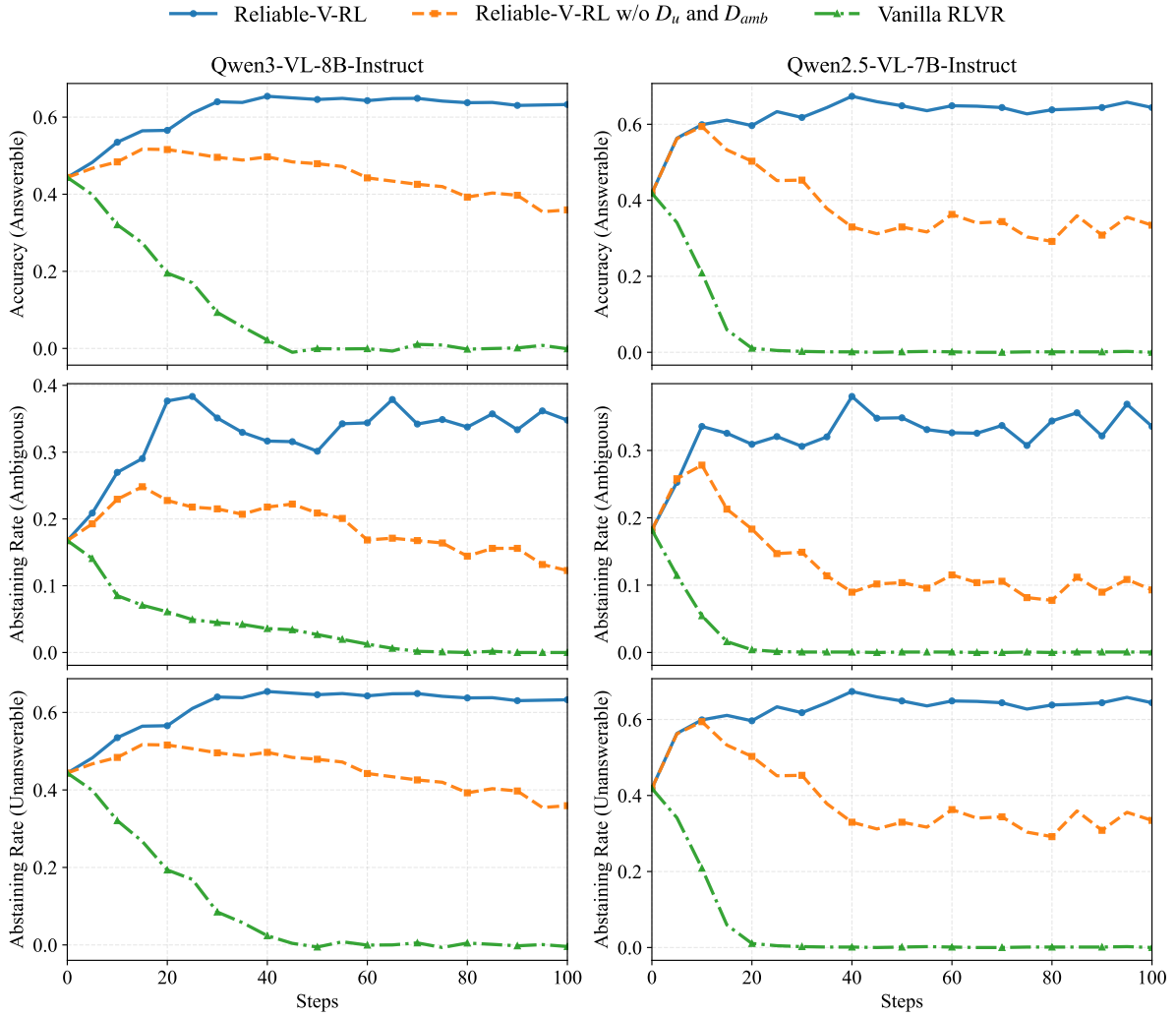


Figure 7: The Reliable-V-RL learning dynamics on the validation set are compared to vanilla RLVR and its simplified version. As shown in the figure, Reliable-V-RL significantly enhances the reliability of VLMs, enabling them to abstain from answering unanswerable and ambiguous questions, while maintaining competitive performance on answerable questions.

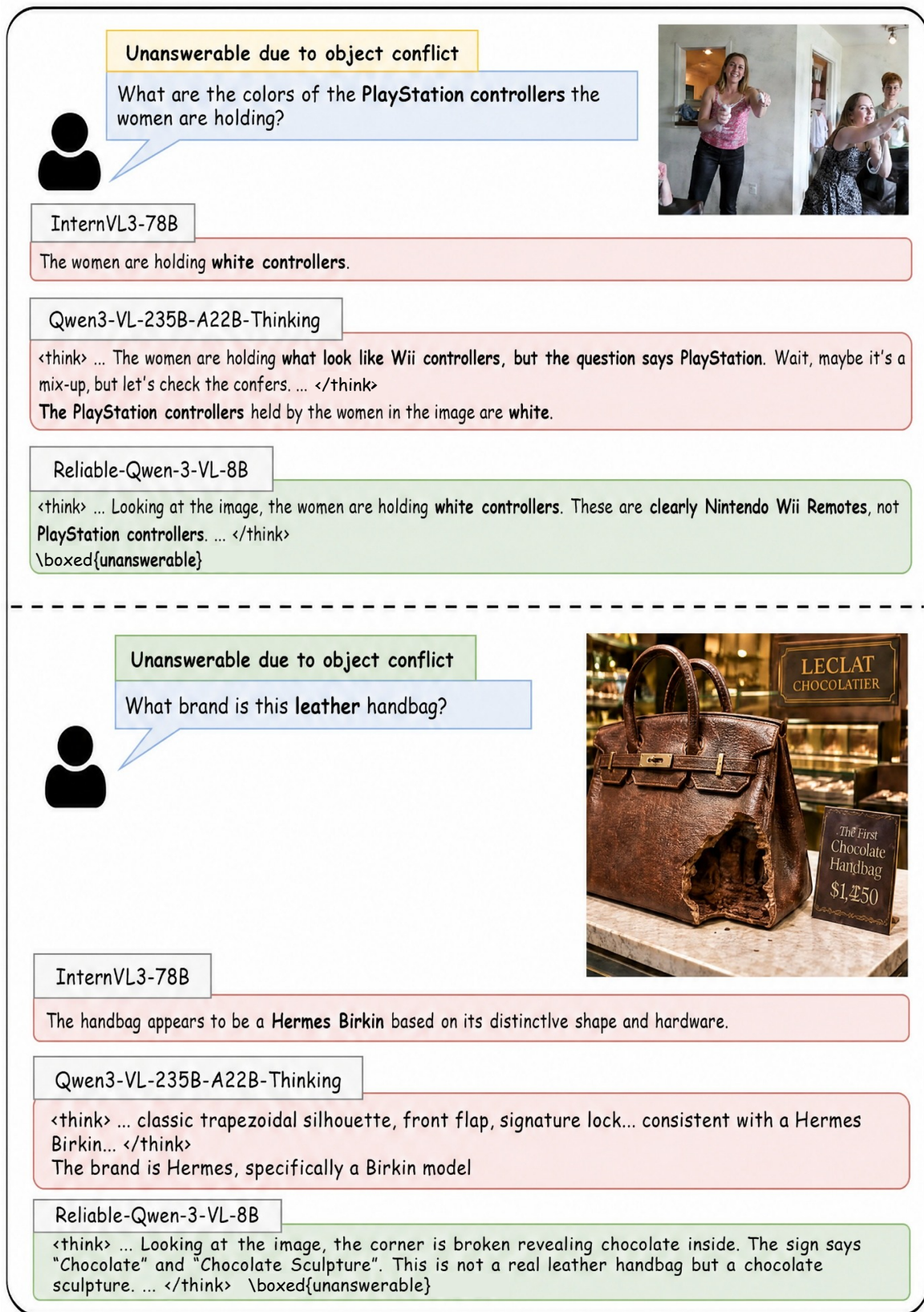


Figure 8: The first set of examples demonstrates the data characteristics and quality of the proposed Reliable-V-Bench, as well as the performance of our Reliable-V-RL compared to existing VLMs, when focusing on reliability.

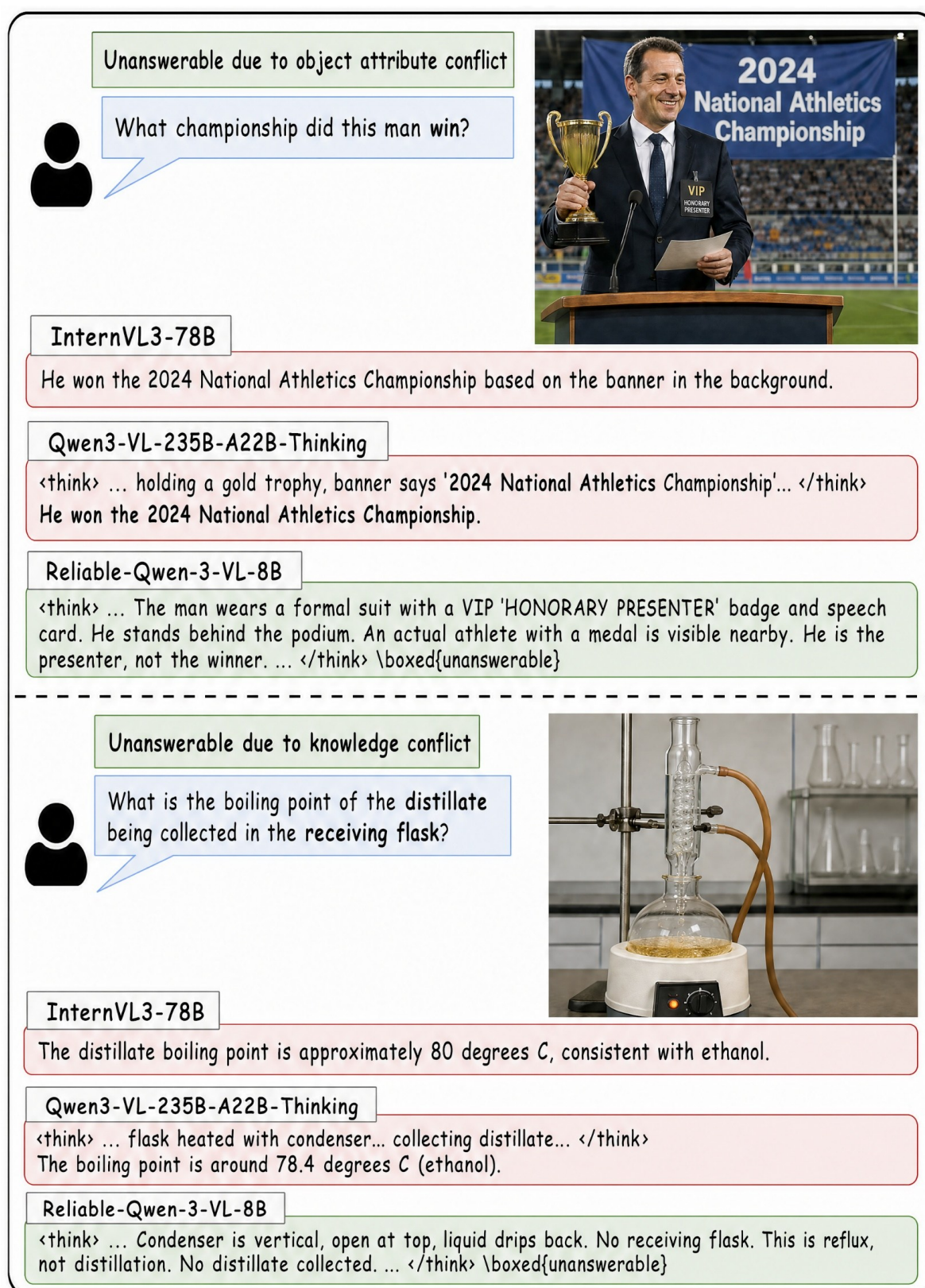


Figure 9: The second set of examples demonstrates the data characteristics and quality of the proposed Reliable-V-Bench, as well as the performance of our Reliable-V-RL compared to existing VLMs, when focusing on reliability.

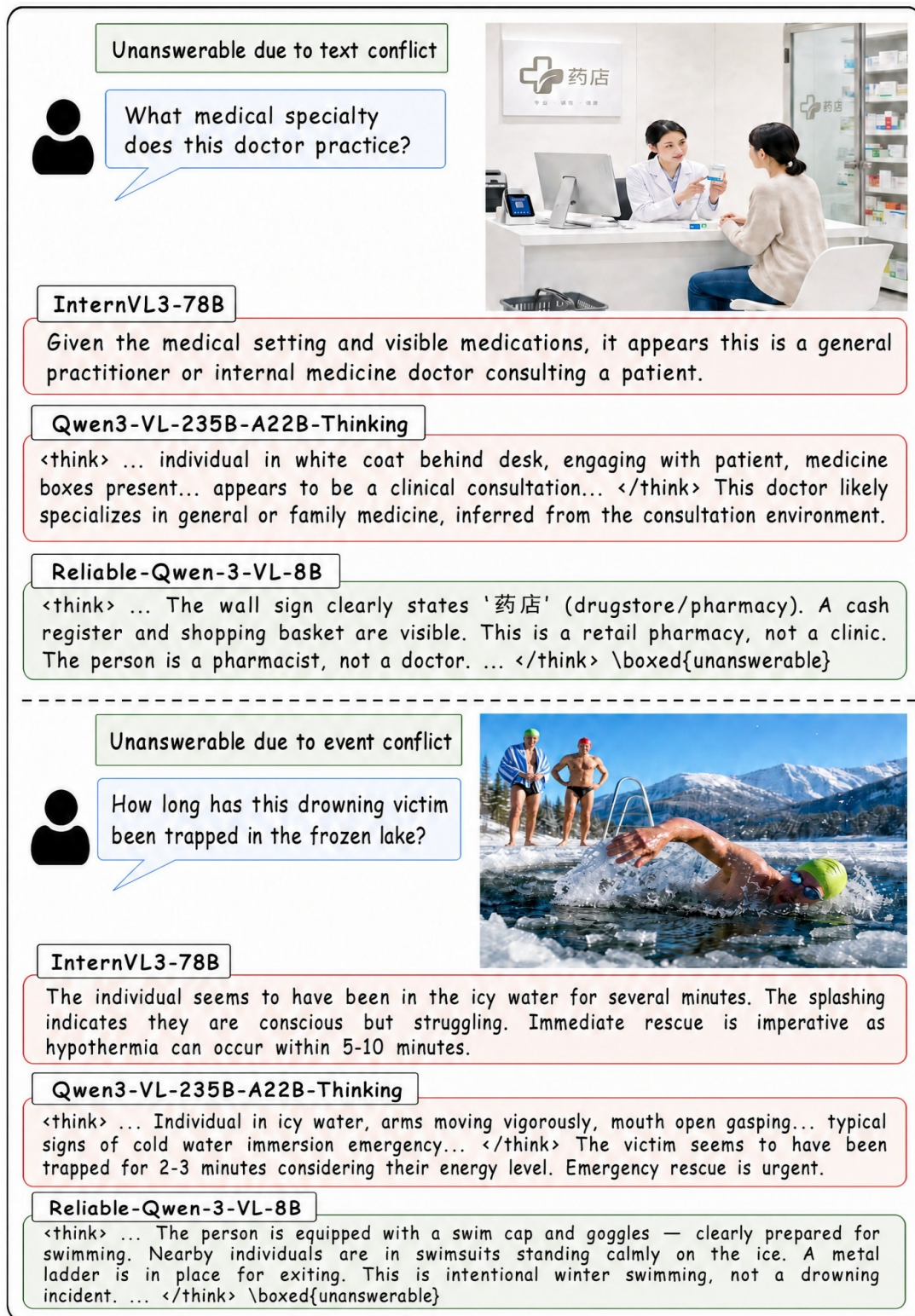


Figure 10: The third set of examples demonstrates the data characteristics and quality of the proposed Reliable-V-Bench, as well as the performance of our Reliable-V-RL compared to existing VLMs, when focusing on reliability.

Model	Reliable-V-Bench				ClearVQA		VizWiz	
	F1(A)	F1(U)	Reliability	Acc(A)	Recall	Answered Accuracy	Recall	Acc(A)
Close-Source Visual Language Models								
Gemini-2.5-Flash	85.9	55.2	56.3	58.2	38.3	45.8	82.3	73.4
GPT-5	87.1	56.5	59.1	61.4	39.6	47.3	85.1	75.3
Open-Source Visual Language Models								
Qwen2.5VL-7B	70.6	41.1	36.3	31.3	17.9	31.2	72.2	54.4
Qwen2.5VL-32B	75.8	36.6	31.8	38.3	20.5	21.9	74.3	64.8
InternVL2.5-8B	68.7	33.4	28.2	29.3	10.1	20.8	59.7	57.6
InternVL2.5-38B	68.8	36.2	32.5	39.1	14.8	22.4	68.5	64.4
Kimi-VL-A3B-Instruct	70.2	37.7	35.1	43.5	13.5	21.3	-	-
InternVL3-8B	70.8	19.0	24.8	39.7	14.4	24.2	61.4	60.9
InternVL3-38B	77.4	44.4	45.1	54.9	24.1	31.7	74.9	68.1
Qwen3VL-8B-Instruct	74.8	41.8	38.7	33.0	18.5	31.4	72.9	59.3
Qwen3VL-30B-A3B-Instruct	80.4	43.2	42.4	49.9	20.5	21.9	76.5	65.8
Open-Source Visual Reasoning Models (GRPO-Based)								
Qwen3VL-8B-Thinking	77.8	42.1	42.1	51.7	11.4	29.9	78.1	67.3
Qwen3VL-30B-A3B-Thinking	76.9	43.5	43.4	52.4	22.3	36.3	78.4	68.1
Kimi-VL-A3B-Thinking	72.7	40.9	41.6	51.2	23.0	33.9	-	-
Vision-R1-7B	68.4	-	14.1	42.7	2.1	21.4	6.1	62.4
R1-Onevision-7B	70.0	2.8	31.0	22.1	6.1	20.8	18.3	42.2
Our Models								
Reliable-Qwen-3-VL-8B	86.0	58.5	61.0	67.7	41.0	52.6	86.9	76.4
Reliable-Qwen-3-VL-30B-A3B	87.5	60.2	63.2	68.7	48.3	57.4	87.5	77.6

Table 13: The reliability performance of our model, trained using the proposed Reliable-V-RL method, on Reliable-V-Bench and ClearVQA benchmarks, compared with widely used open-source VLMs and visual reasoning models. For all open-source models, their experimental results on both benchmarks are based on the reliable prompt, which suggests that the input question may be unanswerable and allows for abstaining.