Asking the Missing Piece: Context-Driven Clarification for Ambiguous VQA

Zongwan Cao[1]* **Bingbing Wen**[1]* **Lucy Lu Wang**[1,2]
[1] University of Washington
[2] Allen Institute for AI
{zongwanc, bingbw, lucylw}@uw.edu

Abstract

Visual Question Answering (VQA) can suffer from under-specification, where the same image-question pair may have multiple plausible answers depending on missing external context. Existing research highlights this limitation, but does not provide methods for teaching models to proactively seek for context. In this work, we study the task of open-ended clarification question generation for underspecified VQA. We curate a dataset of ambiguous VQA pairs annotated with human-verified clarification questions that capture cultural, temporal, spatial, or attribute-based uncertainty. To address this task, we develop a reinforcement learning framework, Grounded Reasoning Preference Optimization–Clarification Reasoning (GRPO-CR), which integrates tailored reward functions to ensure generated clarifications are effective at resolving ambiguity. Experimental results show that GRPO-CR enables VLMs to ask clarification questions that more reliably reduce uncertainty. Our work establishes open-ended, context-seeking clarification as a principled pathway toward interactive, trustworthy multimodal systems that know when and what to ask before answering.

1 Introduction

The task of Visual Question Answering (VQA) requires a model to answer a natural language question given an image [Agrawal et al., 2016]. Despite advances in vision-language models [Li et al., 2023, Liu et al., 2023, Bai et al., 2023], these systems often falter when faced with underspecified inputs, where a single image-question pair may have multiple plausible answers depending on the value of missing context [Luo et al., 2024]. For example, the question "Is the driver speeding?" cannot be resolved without knowing the local speed limit. As illustrated in Figure 1, traditional VQA systems usually guess an answer in these cases, while a clarification-based system can proactively identify the missing context and resolve the ambiguity more reliably.

Recent research exposes different sources of ambiguity in VQA. Initial work focuses on visual illusions and perceptual ambiguity: benchmarks like HallusionBench and IllusionVQA demonstrate that models still stumble on misleading visual stimuli [Guan et al., 2023]. Other efforts address underspecification in the language-vision context, such as VQ-FocusAmbiguity, which identifies all plausible referents in an image to resolve focus-based uncertainty [Chen et al., 2025]. Datasets like CODIS and MUCAR evaluate models under context-dependent and multilingual dual-ambiguity scenarios [Luo et al., 2024, Wang et al., 2025]. More recently, robustness-oriented benchmarks like VQA-Rephrasings and visual-corruption challenges have emerged to test answer consistency under linguistic variation and visual noise [Shah et al., 2019, Farhan et al., 2024]. A major step toward

*Equal contribution

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Foundations of Reasoning in Language Models.

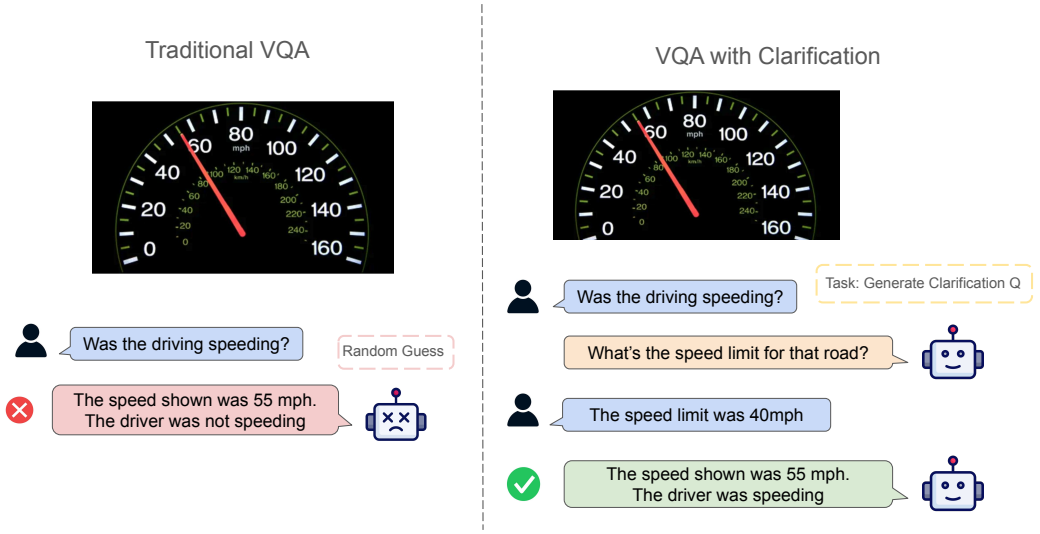


Figure 1: Traditional VQA (left) directly answers ambiguous questions and may guess incorrectly. VQA with clarification (right) asks for missing context (e.g., speed limit) before answering, leading to more accurate and trustworthy responses.

interactive disambiguation comes from ClearVQA, which encourages models to ask clarification questions before answering when the input is ambiguous [Jian et al., 2025].

However, in prior work, ambiguity typically resides in the question phrasing or image content itself, not in the combination of both. This distinction inspires our focus on a complementary setting: underspecified VQA pairs, where the image and question individually are clear, but together yield multiple valid answers depending on external context.

Addressing the ambiguities in the underspecified VQA pairs, the model must learn to explicitly identify and ask for missing context, and not to generically echo or repeat the initial question. While prior work has studied related tasks, they often training models to ask and simulate answers to boolean clarification questions [Jian et al., 2025]. We instead study this in a more open-ended conversational setting. To address this issue, we study the task of generating clarification questions for under-specified VQA. We curate a dataset of 275 image-question pairs, drawn from CODIS and supplemented with manually sourced examples, each annotated with human-verified clarifications capturing cultural, temporal, spatial, or attribute-related ambiguity. We then propose GRPO-CR (Grounded Reasoning Preference Optimization–Clarification Reasoning), a reinforcement learning framework equipped with multiple reward signals covering format, focused relevance, novelty, similarity to reference clarifications, and ambiguity resolution to foster generation of effective clarification questions. Our method consistently outperforms direct prompting and few-shot in-context learning (ICL) baselines in both automatic and human evaluations.

In summary, our contributions are:

- We define the task of VQA clarification question generation in cases of context under-specification. Unlike existing work in VQA clarification, which targets linguistic ambiguity and Yes/No clarification [Jian et al., 2025], we focus on context under-specification and generating open-ended clarification questions.
- We curate a dataset of 275 human-annotated clarification questions targeting VQA contextual ambiguity. Each instance of this dataset contains an ambiguous VQA image-question pair and at least one clarifying question that can be asked to resolve the context ambiguity;
- We develop a tailored reinforcement learning approach (GRPO-CR) that yields high quality, context-seeking clarification questions. We train two base models with GRPO-CR and demonstrate significant improvements over direct-ask and ICL baselines in both automatic metrics and human evaluation.

2 Related Work

Visual Ambiguity Visual ambiguity often arises from incomplete visual cues or distracting noise in a scene [Denison et al., 2018]. Much of the recent work on vision-language models has focused on benchmark creation Yue et al. [2024], Yao et al. [2025], Yang et al. [2025], assessing how well off-the-shelf models can detect and handle ambiguity under different conditions. Early studies examine ambiguities caused by optical illusions [Cui et al., 2023, Fu et al., 2023, Guan et al., 2023]. Later benchmarks such as CODIS [Luo et al., 2024] evaluate whether models interpret images correctly across different contextual settings, while MUCAR [Wang et al., 2025] introduces dual-ambiguity scenarios spanning both textual and visual dimensions. More recently, ClearVQA [Jian et al., 2025] proposes an interactive benchmark where models must ask clarification questions when faced with ambiguous visual questions, and introduces a DPO-based training method to promote clarification-seeking behavior.

Our work is complementary to and differs from ClearVQA [Jian et al., 2025] in three ways. First, in our setting neither the image nor the text is inherently ambiguous; instead, the VQA pair is underspecified, allowing multiple reasonable answers depending on external context. Thus, while both tasks involve clarification question generation, the underlying use cases are distinct. Second, ClearVQA primarily generates Yes/No clarification, whereas our work focuses on open-ended clarification. Third, from a methodological perspective, we adopt a reinforcement learning approach with GRPO, in contrast to the DPO framework [Rafailov et al., 2023] used in ClearVQA.

Clarification Question Generation Beyond benchmark construction, a growing body of research explores how models can proactively ask clarification questions to resolve uncertainty. In NLP, clarification question generation has been studied in dialogue and QA settings, where models are trained to request missing information rather than produce incomplete or incorrect answers [Rao and Daumé, 2018, Aliannejadi et al., 2019, Wen et al., 2025]. Recent large-scale language model work has extended this idea, using preference optimization or reinforcement learning to encourage clarification seeking in conversational agents [Zhang et al., 2024, Li et al., 2025, Zhang and Choi, 2023, Wen et al., 2023].

In multimodal domains, work remains limited. ClearVQA [Jian et al., 2025] is one of the first to study clarification in VQA, while our approach focuses on context missing VQA pairs and introduces a GRPO-based reinforcement learning framework to optimize the format of the questions, relevance, novelty and the potential for ambiguity resolution.

3 Dataset

Source Our dataset is derived from the CODIS benchmark [Luo et al., 2024], which contains 222 images paired with 395 VQA questions. Every CODIS question is deliberately underspecified, designed to admit multiple plausible answers depending on external context (e.g., cultural background, temporal framing, or spatial setting). While this property makes CODIS valuable for studying ambiguity, not all pairs are equally suitable for clarification: some questions are overly vague, while others require reasoning over multiple missing factors simultaneously.

From the 395 CODIS pairs, we retain 275 after discarding 30% judged to be low quality or unsuitable for focusing on a single missing context. In addition, because some CODIS images are visually weak or poorly aligned with their paired questions, we replace 10 CODIS images with carefully matched ambiguous images collected from the Internet. These replacements preserve the same questions and contexts but provide clearer visual grounding, making the overall reasoning more coherent.

Annotation To construct clarification questions, we use GPT-4o as an initial generator. Each CODIS image, original question, and its associated contexts are provided as input, with a prompt instructing GPT-4o to output one concise clarification question that targets the missing context without revealing it. The generated candidates are then manually reviewed against the CODIS contexts. Annotators apply three operations:

- Discard: 30% of CODIS VQA pairs (120) are removed due to vague questions or contexts requiring multiple clarifications.
- Accept without change: 40% of the remaining 275 pairs use GPT-4o outputs directly.

Split	# Samples	Percentage
Training	191	70%
Validation	42	15%
Test	42	15%
Total	275	100%

Table 1: Dataset split of 275 VQA pairs with verified clarification questions.

- Modify: 60% of the 275 required human intervention, including 20% where the original CODIS VQA question itself is rewritten to ensure it depends on only one missing fact.

For example, CODIS provides the question "Is my brother's behavior legal?" with contexts specifying both age and country. We revise the VQA to "My brother is 20 years old. Is his behavior legal?", leaving only the country unspecified. This allows a clarification such as "Where was this image taken?" to directly resolve the ambiguity.

Table 1 shows the dataset statistics, with 275 VQA pairs divided into training, validation, and test splits following a 70/15/15 ratio.

4 Methodology

4.1 Baseline Setup

To provide comparison points for our proposed approach, we consider two baseline methods:

Direct-Ask. The model is given only the image and the original ambiguous question, without any additional instructions or demonstrations. This baseline tests whether the model can spontaneously recognize missing context and ask for clarification on its own.

Few-Shot In-Context Learning (ICL). The model is given several demonstration pairs of VQA and their corresponding clarifications before being asked to generate a clarification for a new input. This setup evaluates whether providing exemplar clarifications improves the model's ability to detect and query for missing context.

4.2 Our Approach: GRPO-Clarification Reasoning

4.2.1 Problem Formulation

We frame clarification generation as a reinforcement learning problem, where the goal of the policy is to actively seek out missing context by asking a clarifying question.

Formally, let the input be an image x_{image} and a question x_{question}, where there is some contextual ambiguity associated with the question. The policy π_θ defines a distribution over clarification questions q conditioned on $(x_{\text{image}}, x_{\text{question}})$. After a clarification q is generated, the environment supplies a response r that represents the missing context. Together with $(x_{\text{image}}, x_{\text{question}})$, this response enables the production of a final answer. A scalar reward $R(q)$ is computed to measure whether the clarification is well formed and whether it helps to resolve the ambiguity of the initial question. The learning objective is to maximize the expected reward.

We optimize the policy using *Group-Relative Policy Optimization* (GRPO) [Shao et al., 2024], a reinforcement learning method that removes the need for a value function baseline. For each input $(x_{\text{image}}, x_{\text{question}})$, the policy π_θ samples a group of K clarification questions $\{q_1, \ldots, q_K\}$. Each q_i receives a scalar reward $R(q_i)$. The group mean reward serves as the baseline:

$$\bar{R} = \frac{1}{K} \sum_{i=1}^{K} R(q_i).$$

The advantage for each candidate is computed as:

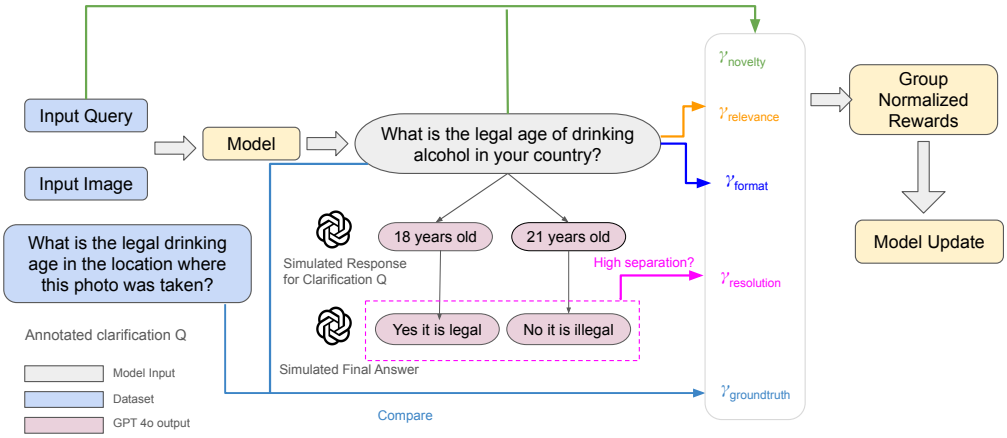


Figure 2: Training pipeline for our method GRPO-CR. The model generates clarification questions from image-query pairs, which are evaluated by multi-dimensional rewards and updated via group-normalized reinforcement.

$$A(q_i) = R(q_i) - \bar{R}.$$

The GRPO objective then maximizes the weighted log-likelihood of the sampled questions:

$$J(\theta) = \frac{1}{K} \sum_{i=1}^{K} A(q_i)\ \log \pi_\theta(q_i \mid x_{\text{image}}, x_{\text{question}}).$$

4.2.2 Training Pipeline

Our training follows the GRPO-CR framework. For each ambiguous VQA input, the model generates candidate clarification questions. These outputs are scored by a set of reward functions that enforce structural validity, contextual relevance, novelty, similarity to curated references and ambiguity resolution potential. The aggregated reward is then used to update the policy with preference optimization. This process encourages the model to consistently produce clarifications that are well-formed, targeted, and useful for resolving uncertainty.

4.2.3 Reward Design

Our reward design is inspired by prior work that demonstrates that teaching models to ask good questions requires balancing structural validity and informativeness. For example, ALFA shows that aligning LLMs to ask clarification questions benefits from rewards that capture both format compliance and clinical utility [Li et al., 2025]. Similarly, GRIT demonstrates that multimodal reasoning tasks require explicit rewards to encourage models to think with images [Fan et al., 2025]. Unlike earlier work, which focused on text only or general reasoning, we introduce a tailored set of multimodal rewards: format, focused relevance, novelty, ground-truth similarity, and ambiguity resolution, specifically for context underspecification in VQA. This combination ensures that clarifications are not only well-formed, but also directly reduce uncertainty by eliciting the missing context.

Format reward (γ_{format}). We introduce a question format reward to ensure that clarification questions in a consistent and usable form. It does not evaluate the content of the question, but enforces basic structural rules. A generated question receives a positive reward if it is a single line that ends with exactly one '?', falls within a range of 4 to 35 words. Otherwise, it is penalized. By acting as a gatekeeper, the format reward ensures that subsequent rewards for relevance, novelty and ambiguity resolution are only applied when the output is a valid question.

Focused Relevance reward ($\gamma_{\text{relevance}}$). The focused relevance reward measures whether a clarification is both on-topic and specific. Relevance is checked via keywords linked to the ambiguity category (e.g., temporal, cultural, attribute), while targetedness is identified by patterns that narrow the question to a concrete detail. Questions satisfying both receive the highest reward, those meeting only one get a smaller reward, and irrelevant or unfocused questions are penalized. This encourages clarifications that are not just related but directly address the source of ambiguity.

Ambiguity resolution reward ($\gamma_{\text{resolution}}$). The ambiguity resolution reward evaluates whether a clarification meaningfully reduces uncertainty in the original question. A clarification is rewarded if different user responses to it would lead to different answers, and penalized if all responses converge to the same outcome. For example, for "Is the driver speeding?", asking "What is the legal speed limit for this road?" enables different final interpretations depending on the reply. We operationalize this by prompting GPT-4o to simulate multiple responses and checking whether the resulting answers diverge.

Novelty reward(γ_{novelty}) The novelty reward penalizes trivial rephrasings of the original ambiguous question and encourages clarifications that introduce genuinely new information. If the generated question closely mirrors the original phrasing, it receives a negative reward; partial overlap is mildly penalized; and questions with sufficient divergence are rewarded. This discourages echo-like outputs and promotes clarifications that add meaningful value beyond the original query.

Ground truth check reward ($\gamma_{\text{groundtruth}}$). We introduce a ground truth similarity reward as a soft guide. For each VQA pair, a curated clarification question is used as reference. The model's output is compared to this reference using GPT-4o, yielding a similarity score between 0 and 1. Rather than penalizing divergence, the reward provides positive signal when the output aligns with the reference, helping to stabilize training and encourage human-like clarifications.

Rewards Aggregation. The individual rewards are combined into a single scalar value for each rollout. The final reward is normalized across the batch and used as the advantage estimate in GRPO.

4.3 Implementation Details

We initialize experiments with publicly available checkpoints of Qwen2.5-VL-3B-Instruct-Instruct [Qwen et al., 2024] and InternVL3-2B-Instruct-Instruct [Zhu et al., 2025]. Training is conducted using LoRA fine-tuning for parameter efficiency. Optimization is performed with the DeepSpeed ZeRO-2 framework to handle large batch sizes across multiple GPUs. Models are trained with a learning rate of 2×10^{-6}, batch size of 2 per device, and gradient accumulation of 8 steps. We set a maximum prompt length of 1000 tokens and allow generations up to 1000 tokens, although clarifications are typically much shorter.

For each update step, the model produces two candidate generations per input. Rewards are computed for each candidate, and the GRPO objective is used to adjust the policy accordingly. Training proceeds for one epoch over the 191 training examples, with evaluation conducted every 50 steps on the validation set. Human-readable outputs are logged and inspected throughout training to monitor quality and detect potential reward hacking behaviors.

All experiments are conducted on two NVIDIA GPUs per run: Qwen2.5-VL-3B-Instruct on A100-80GB and InternVL3-2B-Instruct on A40-48GB. Training required roughly 12 hours in each case. We used the train/validation/test split of 191/42/42 examples described in Section 3. Results are reported on the held-out test set, with validation used for early monitoring.

4.4 Evaluation Metrics

We evaluate model outputs using two complementary approaches.

Reward based. All reward functions introduced during training are reapplied to the test set outputs. This provides an automatic, fine-grained assessment of structural validity and usefulness.

Human evaluation. We manually assess each clarification question along two dimensions: (i) propensity to ask, measuring whether the model identifies the need for clarification rather than attempting to answer directly, and (ii) question quality, judging whether the clarification question meaningfully resolves ambiguity in the original query. One author assessed generated questions

Model	Method	Format	Focused Rel.	Novelty	GT Sim.	Ambig. Res.
Qwen2.5-VL-3B	Direct-Ask	-0.48	0.00	0.00	0.04	0.01
	Few-Shot ICL	0.40	0.13	-0.07	0.09	0.08
	GRPO-CR	**0.50**	**0.17**	**0.07**	**0.32**	**0.33**
InternVL3-2B	Direct-Ask	-0.50	0.00	0.00	0.00	0.01
	Few-Shot ICL	-0.12	0.06	0.03	0.05	0.06
	GRPO-CR	**0.50**	**0.18**	**0.08**	**0.29**	**0.33**

Table 2: Automatic reward-based evaluation results

Model	Method	Propensity to Ask (%)	Question Quality (%)
Qwen2.5-VL-3B	Direct-Ask	0	0
	Few-Shot ICL	100	14.29
	GRPO-CR (ours)	**100**	**47.62**
InternVL3-2B	Direct-Ask	0	0
	Few-Shot ICL	100	9.52
	GRPO-CR (ours)	**100**	**45.24**

Table 3: Human evaluation results

from all model settings (2 base models x 3 methods) for all 42 questions in the test split (252 total questions), providing binary judgments along these two dimensions.

5 Results & Discussion

Automatic Evaluation Results Our results show that GRPO-CR substantially outperforms direct-ask and few-shot ICL, producing clarifications that are more valid, novel, and effective at reducing ambiguity. As shown in Table 2, the direct-ask baseline performs poorly across all automatic metrics, often producing no valid clarification and defaulting to near-zero scores. Few-shot ICL achieves moderate gains, especially on *focused relevance* (0.13 for Qwen2.5-VL-3B and 0.06 for InternVL3-2B), but these improvements mostly reflect surface-level echoes or rephrasings of the original question, which explains its weak *novelty* (negative for Qwen2.5-VL-3B, marginal for InternVL3-2B) and low *ambiguity resolution* (0.08 and 0.06).

In contrast, GRPO-CR achieves the best performance across nearly all metrics: the highest *format compliance* (0.50 for both models), stronger *novelty* (0.07 and 0.08), and clear gains in *ground truth similarity* (0.32 and 0.29). Most importantly, it delivers the strongest *ambiguity resolution* (0.33 for both models), showing that its clarifications are most effective at recovering missing context. The only partial exception is *focused relevance* on Qwen2.5-VL-3B, where ICL (0.13) comes close to GRPO-CR (0.17), reflecting ICL's tendency to mimic the input wording without fully addressing the ambiguity.

Human Evaluation Results Our human evaluation shows that GRPO-CR produces substantially higher-quality clarifications than both baselines, while direct-ask rarely asks a question and ICL mainly echoes the input. These findings suggest that explicit rewards for relevance, novelty, and ambiguity resolution are crucial. Simply prompting or providing demonstrations is not enough to elicit high-quality clarifications.

As shown in Table 3, Direct-Ask almost never produces a clarification, which is unsurprising since the model receives no instruction to ask. Few-shot ICL asks nearly every time, but the quality is very low: most questions are shallow rephrasings that do little to reduce uncertainty.

By contrast, GRPO-CR consistently generates clarifications that annotators judged as useful, with quality scores above 45% for both backbones. This demonstrates that reinforcement learning with tailored rewards not only enforces structural validity but also helps the model identify missing context.

Qualitative inspection reinforces this distinction. For example, when asked "Based on the image, what time is it now?" with two clocks showing different times, ICL often outputs vague echoes like "What is the current time?", which fails to resolve the ambiguity. GRPO-CR instead asks "Is the

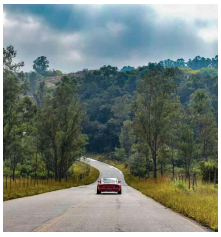
VQA Image	VQA Question and Model-generated Outputs (with commentary)
	**Original question:** I am driving on the left lane. Is it allowed? **Ground truth clarification:** That depends on the country's traffic laws. In which country was this photo taken? **Direct-Ask:** Yes, it is allowed to drive on the left lane. *(unwarranted inference)* **Few-Shot ICL:** What is the speed limit on this road? *(partially relevant but does not ask about the key disambiguator)* **GRPO-CR:** Is the road in the image a two-way or one-way road? *(okay: resolves some ambiguity)*
	**Original question:** I want to buy two bottles of milk. Is my gesture appropriate? **Ground truth clarification:** In what country or region is this gesture being used?" **Direct-Ask:** Your gesture is not appropriate for buying milk. It is a peace sign. *(unwarranted inference)* **Few-Shot ICL:** Is the gesture being made with the right hand? *(does not resolve the ambiguity of the original question)* **GRPO-CR:** Are you making a peace sign or a hand gesture that could be interpreted as a sign of agreement or approval? *(partially relevant but does not resolve the main ambiguity)*
	**Original question:** The clock is on 12-hour format. What time is it on a 24-hour schedule? **Ground truth clarification:** Is this photo from the morning or late at night? **Direct-Ask:** The time on the clock is 10:00. *(unwarranted inference)* **Few-Shot ICL:** What is the current time on a 24-hour clock? *(repeats original question, no additional information gained)* **GRPO-CR:** Is the clock showing the time in the morning or afternoon? *(good: resolves AM/PM)*
	**Original question:** Is the subway going in the same direction as I face? **Ground truth clarification:** Is it approaching or moving away? **Direct-Ask:** The subway is going in the opposite direction as you face. *(unwarranted inference)* **Few-Shot ICL:** Is the train moving towards the platform or away from it? *(good: resolves direction)* **GRPO-CR:** Is the train moving towards you or away from you? *(good: resolves direction)*

Table 4: Qualitative examples show that training improves model ability to ask clarifying questions, especially in the GRPO-CR setting.

clock on the left showing the correct time?", directly pinpointing the missing context. Similarly, for "The clock is on 12-hour format. What time is it on a 24-hour schedule?", ICL asks "What is the current time on a 24-hour clock?", whereas GRPO-CR produces "Is the clock showing the time in the morning or afternoon?", making the clarification specific and actionable.

Qualitative Examples Table 4 shows sample outputs: direct-ask answers directly, ICL often echoes the input, while GRPO-CR generates clarifications that pinpoint the missing context.

6 Conclusion

We present the task of clarification question generation for ambiguous VQA due to missing context, supported by a dataset of 275 human verified examples covering diverse ambiguity types. Our reinforcement learning framework, GRPO-CR, uses tailored rewards to produce clarifications that are both well formed and useful, outperforming direct-ask and few-shot ICL baselines. These results show that reward design is key for teaching models when and how to seek context clarification, moving toward safer and more collaborative multimodal AI.

7 Limitation and Future work

While our results demonstrate the promise of GRPO-CR, the study is limited by the relatively small scale of our dataset and evaluation, as well as by the design of our Direct-Ask baseline, which did not explicitly instruct the model to ask for clarification. Future work will extend training to additional backbone models and combine reinforcement learning with supervised fine-tuning. A key challenge is handling settings where ambiguous and unambiguous VQA questions coexist. This will require coupling clarification generation with ambiguity detection or uncertainty estimation. We also plan to evaluate on larger benchmarks such as MME and MMA to assess generalization, and to explore multi-turn interactions where clarification unfolds over multiple rounds.

Acknowledgments and Disclosure of Funding

We gratefully acknowledge gift funding from Google and the Allen Institute for AI (Ai2).

References

Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. Vqa: Visual question answering. *arXiv:1505.00468 [cs]*, 10 2016. URL `https://arxiv.org/abs/1505.00468`.

Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W. Bruce Croft. Asking clarifying questions in open-domain information-seeking conversations. *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 07 2019. doi: 10.1145/3331184.3331265.

Yulong Bai, Xiaozhi Bai, Ailun Chen, et al. Qwen-vl: A versatile vision-language model for understanding, generating, and instruction following. *arXiv preprint arXiv:2308.12966*, 2023.

Chongyan Chen, Yu-Yun Tseng, Zhuoheng Li, Anush Venkatesh, and Danna Gurari. Acknowledging focus ambiguity in visual questions, 2025. URL `https://arxiv.org/abs/2501.02201`.

Chenhang Cui, Yiyang Zhou, Xinyu Yang, Shirley Wu, Linjun Zhang, James Zou, and Huaxiu Yao. Holistic analysis of hallucination in gpt-4v(ision): Bias and interference challenges, 2023. URL `https://arxiv.org/abs/2311.03287`.

Rachel N. Denison, William T. Adler, Marisa Carrasco, and Wei Ji Ma. Humans incorporate attention-dependent uncertainty into perceptual decisions and confidence. *Proceedings of the National Academy of Sciences*, 115:11090–11095, 10 2018. doi: 10.1073/pnas.1717720115.

Yue Fan, Xuehai He, Diji Yang, Kaizhi Zheng, Ching-Chen Kuo, Yuting Zheng, Sravana Jyothi Narayanaraju, Xinze Guan, and Xin Eric Wang. Grit: Teaching mllms to think with images, 2025. URL `https://arxiv.org/abs/2505.15879`.

Ishmam Md Farhan, Ishmam Tashdeed, Talukder Asir Saadat, Md Hamjajul Ashmafee, Kamal, and Md Azam Hossain. Visual robustness benchmark for visual question answering (vqa), 2024. URL `https://arxiv.org/abs/2407.03386`.

Chaoyou Fu, Renrui Zhang, Zihan Wang, Yubo Huang, Zhengye Zhang, Longtian Qiu, Gaoxiang Ye, Yunhang Shen, Mengdan Zhang, Peixian Chen, Sirui Zhao, Shaohui Lin, Deqiang Jiang, Di Yin, Peng Gao, Ke Li, Hongsheng Li, and Xing Sun. A challenger to gpt-4v? early explorations of gemini in visual expertise. *arXiv (Cornell University)*, 01 2023. doi: 10.48550/arxiv.2312.12436.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models, 2023. URL https://arxiv.org/abs/2310.14566.

Pu Jian, Donglei Yu, Wen Yang, Shuo Ren, and Jiajun Zhang. Teaching vision-language models to ask: Resolving ambiguity in visual questions. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3619–3638, 2025. doi: 10.18653/v1/2025.acl-long.182. URL https://aclanthology.org/2025.acl-long.182/?utm_source=chatgpt.com.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven CH Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2023.

Shuyue Stella Li, Jimin Mun, Faeze Brahman, Pedram Hosseini, Bryceton G Thomas, Jessica M Sin, Bing Ren, Jonathan S Ilgen, Yulia Tsvetkov, and Maarten Sap. Alfa: Aligning llms to ask good questions a case study in clinical reasoning, 2025. URL https://arxiv.org/abs/2502.14860.

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

Fuwen Luo, Chi Chen, Zihao Wan, Zhaolu Kang, Qidong Yan, Yingjie Li, Xiaolong Wang, Siyu Wang, Ziyue Wang, Xiaoyue Mi, Peng Li, Ning Ma, Maosong Sun, and Yang Liu. Codis: Benchmarking context-dependent visual comprehension for multimodal large language models. *arXiv preprint arXiv:2402.13607*, 2024.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2024. URL https://arxiv.org/abs/2412.15115.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 05 2023. URL https://arxiv.org/abs/2305.18290.

Sudha Rao and Hal Daumé. Learning to ask good questions: Ranking clarification questions using neural expected value of perfect information. *Meeting of the Association for Computational Linguistics*, 07 2018. doi: 10.18653/v1/p18-1255.

Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. Cycle-consistency for robust visual question answering, 2019. URL https://arxiv.org/abs/1902.05660.

Zhihong Shao, Runxin Xu, Peiyi Wang, Qihao Zhu, Junxiao Song, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

Xiaolong Wang, Zhaolu Kang, Wangyuxuan Zhai, Xinyue Lou, Yunghwei Lai, Ziyue Wang, Yawen Wang, Kaiyu Huang, Yile Wang, Peng Li, and Yang Liu. Mucar: Benchmarking multilingual cross-modal ambiguity resolution for multimodal large language models, 2025. URL https://arxiv.org/abs/2506.17046.

Bingbing Wen, Zhengyuan Yang, Jianfeng Wang, Zhe Gan, Bill Howe, and Lijuan Wang. Infovisdial: An informative visual dialogue dataset by bridging large multimodal and language models. *arXiv preprint arXiv:2312.13503*, 2023.

Bingbing Wen, Jihan Yao, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe, and Lucy Lu Wang. Know your limits: A survey of abstention in large language models. *Transactions of the Association for Computational Linguistics*, 13:529–556, 2025. doi: 10.1162/tacl_a_00754. URL https://aclanthology.org/2025.tacl-1.26/.

Yiwei Yang, Chung Peng Lee, Shangbin Feng, Dora Zhao, Bingbing Wen, Anthony Zhe Liu, Yulia Tsvetkov, and Bill Howe. Escaping the spuriverse: Can large vision-language models generalize beyond seen spurious correlations? In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. URL https://openreview.net/forum?id=es2NkPKFCB.

Jihan Yao, Yushi Hu, Yujie Yi, Bin Han, Shangbin Feng, Guang Yang, Bingbing Wen, Ranjay Krishna, Lucy Lu Wang, Yulia Tsvetkov, et al. Mmmg: a comprehensive and reliable evaluation suite for multitask multimodal generation. *arXiv preprint arXiv:2505.17613*, 2025.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 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, 2024.

Michael Zhang and Eunsol Choi. Clarify when necessary: Resolving ambiguity through interaction with lms, 2023. URL https://arxiv.org/abs/2311.09469.

Michael Zhang, Knox W Bradley, and Eunsol Choi. Modeling future conversation turns to teach llms to ask clarifying questions, 2024. URL https://arxiv.org/abs/2410.13788.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025. URL https://arxiv.org/abs/2504.10479.