Enhancing Vietnamese VQA-NLE via Learning to Explain with GRPO

Quang-Minh Tran[1]⋆, Phat-Dat To[2]*, Huu-Phuoc Le[4], Duc-Manh Nguyen[3], and Truong-Binh Duong[4]⋆⋆

[1] Vietnam-Korea University of Information and Communication Technology, Vietnam
[2] Ho Chi Minh City University of Technology and Education, Vietnam
[3] University of Transport and Communications, Vietnam
[4] AI VIETNAM Lab, Vietnam

Abstract. Visual question answering with natural language explanations (VQA-NLE) necessitates interpretability in multimodal artificial intelligence. While group relative policy optimization (GRPO) has emerged as an effective post-training paradigm for VQA-NLE tasks, standard formulations often treat internal reasoning and external explanation as a single generation step. We hypothesize that this conflation predisposes the model to hallucinations, where it generates fluent but visually ungrounded justifications. To address this, we investigate an adaptation of GRPO for the Vietnamese VQA-NLE task. Specifically, a novel composite reward mechanism is introduced to separate reasoning from explanation, enforcing clear stages of reasoning, conclusion, and explanation. Crucially, we hypothesize that this structured constraint compels the model to internalize domain knowledge prior to response synthesis. Experiments on the ViVQA-X benchmark demonstrate that this approach significantly outperforms supervised fine-tuning baselines. Notably, an accuracy improvement of 16.05% is achieved on the Vintern-3B backbone, validating the efficacy of structured reasoning in grounding multimodal predictions. We release our code at `https://github.com/T-Sunm/VINLE-GRPO`.

Keywords: VQA-NLE · Reinforcement Learning · GRPO · Low-resource Language · Multimodal Reasoning.

1 Introduction and Related Work

Visual question answering with natural language explanations (VQA-NLE) extends the standard formulation by requiring models to generate an accurate answer accompanied by a human-readable rationale to justify the decision process [9]. This paradigm addresses the interpretability constraints of traditional discriminative models by grounding predictions in explicit causal reasoning. While foundational benchmarks such as VQA-X [9] established the task, the methodology has transitioned toward vision-language models (VLMs). These models leverage large-scale multimodal pre-training to approach VQA-NLE as a generalized generation task.

⋆ Equal contribution.

⋆⋆ Project leader.

To further specialize VLMs for complex reasoning tasks, recent paradigms [10] have increasingly adopted reinforcement learning (RL) post-training algorithms. Notably, group relative policy optimization (GRPO) [11] has emerged as a prominent approach due to its training stability and memory efficiency, achieved by eliminating the critic model. The efficacy of GRPO is inherently tied to the design of the reward function. Consequently, recent studies have prioritized the tailoring of reward mechanisms to specific domains. For instance, MedVLM-R1 [8] integrates clinical metrics to enforce medical correctness, while SVQA-R1 [12] employs view-consistent penalties for spatial reasoning. Similarly, consistency-aware approaches exemplified by TACO [4] address the logical coherence between reasoning chains and final outputs.

Despite the strong alignment ability of GRPO approaches in VQA-NLE tasks, they often treat output generation as a single process and focus on deterministic verification. We observe that in open-ended VQA-NLE, explainability alone is insufficient to establish credibility [2]. Models may first internalize domain knowledge through explicit reasoning before synthesizing explanations. This distinction is critical, as reasoning involves unstructured exploration to derive an answer, whereas explanation requires structured justification to communicate the result. Standard RL objectives, which often conflate these steps, fail to enforce this structure. This is particularly detrimental in low-resource languages such as Vietnamese, where the scarcity of explicit reasoning annotations prevents models from learning grounded logical patterns, leaving them to rely on surface-level statistical shortcuts. To address this, we attempt to adapt GRPO for the Vietnamese VQA-NLE task to tackle data scarcity through structural reinforcement. We employ a composite reward mechanism to decouple reasoning from response generation, enforcing a strict progression of reasoning, conclusion, and explanation. By combining structural constraints with semantic metrics tailored to Vietnamese variability, the framework ensures grounded reasoning rather than surface. The main contributions of this paper are summarized as follows:

1. We investigate an adaptation of GRPO to enforce structured cognitive topology in Vietnamese VQA-NLE.
2. To prioritize faithfulness over surface-level plausibility, we introduce a composite reward mechanism for structural, semantic, and rationale alignment for Vietnamese VQA-NLE task.
3. Experiments conducted on the ViVQA-X [3] dataset show superior accuracy and explanation quality compared to SFT baselines.

2 Methodology

2.1 Preliminaries

In the context of VQA-NLE, let the multimodal input be denoted as $x = \{v, q\}$, comprising an image v and a textual question q. Based on this input, the policy π_θ, parameterized by learnable weights θ, samples a group of G outputs $\{y_1, \ldots, y_G\}$ from the old policy $\pi_{\theta_{\text{old}}}$. Subsequently, the advantage $\hat{A}_i$ for each candidate $i \in \{1, \ldots, G\}$ is derived by normalizing the reward r_i against group statistics:

$$\hat{A}_i = \frac{r_i - \text{mean}(\{r_1, r_2, \ldots, r_G\})}{\text{std}(\{r_1, r_2, \ldots, r_G\}) + \epsilon}, \tag{1}$$

where ϵ is a small constant for numerical stability. To optimize the policy π_θ, the objective $\mathcal{J}_{\text{GRPO}}$ is formulated to maximize the expected return, incorporating a clipped surrogate term and a Kullback-Leibler (KL) divergence penalty:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \frac{1}{G}\sum_{i=1}^{G}\left(\min\left(\rho_i \hat{A}_i, \text{clip}(\rho_i, 1-\delta, 1+\delta)\hat{A}_i\right) - \beta D_{\text{KL}}(\pi_\theta || \pi_{\text{ref}})\right). \tag{2}$$

Here, ρ_i denotes the probability ratio between the current and the old policies. The clipping mechanism restricts ρ_i to the interval $[1-\delta, 1+\delta]$, while β controls the divergence from the reference policy π_{ref}. This mechanism promotes the generation of sequences that outperform other samples within the group.

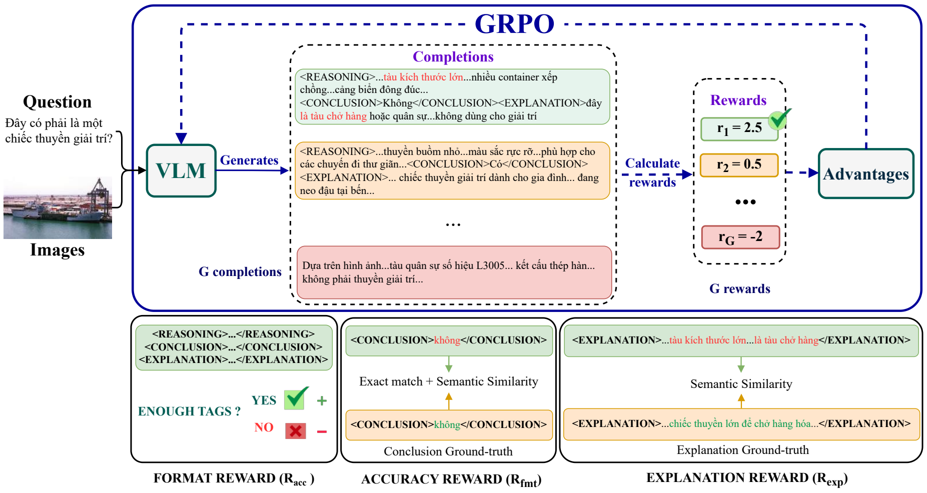
GRPO
Question
Đây có phải là một chiếc thuyền giải trí?
Images
VLM
Generates
Completions
<REASONING>...tàu kích thước lớn...nhiều container xếp chồng...cảng biển đông đúc... <CONCLUSION>Không</CONCLUSION><EXPLANATION>đây là tàu chở hàng hoặc quân sự...không dùng cho giải trí
<REASONING>...thuyền buồm nhỏ...màu sắc rực rỡ...phù hợp cho các chuyến đi thư giãn...<CONCLUSION>Có</CONCLUSION> <EXPLANATION>... chiếc thuyền giải trí dành cho gia đình... đang neo đậu tại bến...
...
Dựa trên hình ảnh...tàu quân sự số hiệu L3005... kết cấu thép hàn... không phải thuyền giải trí...
G completions
Calculate rewards
Rewards
r1 = 2.5
r2 = 0.5
•••
rG = -2
G rewards
Advantages
<REASONING>...</REASONING>
<CONCLUSION>...</CONCLUSION>
<EXPLANATION>...</EXPLANATION>
ENOUGH TAGS ?
YES ✓ +
NO ✗ −
<CONCLUSION>không</CONCLUSION>
Exact match + Semantic Similarity
<CONCLUSION>không</CONCLUSION>
Conclusion Ground-truth
<EXPLANATION>...tàu kích thước lớn...là tàu chở hàng</EXPLANATION>
Semantic Similarity
<EXPLANATION>...chiếc thuyền lớn để chở hàng hóa...</EXPLANATION>
Explanation Ground-truth
FORMAT REWARD (Racc)
ACCURACY REWARD (Rfmt)
EXPLANATION REWARD (Rexp)

Fig. 1. Illustration of the GRPO adaptation for the Vietnamese VQA-NLE task. The VLM samples G completions evaluated via composite rewards $(R_{fmt}, R_{acc}, R_{exp})$. Normalized advantages update the policy directly without a critic model.

2.2 Reward Functions

While the original GRPO [11] demonstrated effectiveness in mathematical reasoning, its reliance on deterministic verification is inadequate for the open-ended nature of VQA-NLE. Especially in Vietnamese, high lexical diversity renders rigid outcome verification insufficient. To address this, we propose a composite reward mechanism (R_{total}) that enforces a sequential workflow, requiring explicit internal reasoning before deriving conclusions and synthesizing explanations. As illustrated in Fig. 1, given a model response y, optimization is decoupled into three complementary components:

$$R_{\text{total}}(y) = R_{\text{fmt}}(y) + R_{\text{acc}}(y) + R_{\text{exp}}(y), \tag{3}$$

where R_{fmt}, R_{acc}, and R_{exp} target format adherence, answer precision, and explanation faithfulness, respectively.

Format reward ($\boldsymbol{R}_{\textbf{fmt}}$). This component enforces the structural separation of reasoning and communication. The output must be strictly encapsulated within <REASONING>, <CONCLUSION>, and <EXPLANATION> tags:

$$R_{\text{fmt}}(y) = \mathbb{I}_{\text{struct}}(y) \cdot w_{\text{base}} - \max(0, N_{\text{tags}} - N_{\text{req}}) \cdot w_{\text{pen}}, \tag{4}$$

where $\mathbb{I}_{\text{struct}}$ represents an indicator function returning 1 for valid structures. Hyperparameters w_{base} and w_{pen} regulate the reward and penalty magnitudes, while N_{tags} and N_{req} denote the generated and required tag counts.

Accuracy reward ($\boldsymbol{R}_{\textbf{acc}}$). To accommodate the linguistic variability discussed above where affirmative answers appear as distinct tokens such as "Đúng" or "Có", we replace exact matching with a hybrid metric combining semantic similarity and lexical overlap:

$$R_{\text{acc}}(y) = \begin{cases} -1.0 & \text{if } y_{\text{ans}} = \emptyset \text{ or } S_{\text{hybrid}} < \tau, \\ S_{\text{hybrid}} & \text{otherwise} \end{cases} \tag{5}$$

where $S_{\text{hybrid}} = \alpha \cdot \text{BS}(y_{\text{ans}}, g_{\text{ans}}) + (1-\alpha) \cdot \text{ROUGE}(y_{\text{ans}}, g_{\text{ans}})$. Here, y_{ans} and g_{ans} denote the normalized predicted and ground-truth answers, respectively, while α governs the trade-off. Specifically, BS utilizes BERTScore [14] with a PhoBERT backbone [7] to capture semantic alignment, ensuring robustness against synonym usage. Meanwhile, ROUGE-L [6] is integrated to guarantee lexical precision, penalizing hallucinations that may be semantically close but contextually inappropriate. Finally, a threshold τ filters low-confidence predictions.

Explanation reward ($\boldsymbol{R}_{\textbf{exp}}$). This function quantifies the reasoning quality by measuring the semantic alignment between the generated explanation y_{exp} and the ground-truth rationale g_{exp}:

$$R_{\text{exp}}(y) = \text{BERTScore}(y_{\text{exp}}, g_{\text{exp}}). \tag{6}$$

Sim_{sem} employs BERTScore [14] to capture contextual embedding alignment, prioritizing underlying logic preservation over surface-level lexical matching.

3 Experimental Results

3.1 Dataset

Experiments are conducted on the ViVQA-X dataset [3], a Vietnamese adaptation of VQA-X derived from MS COCO. This dataset comprises 32,886 question-answer pairs and 41,817 explanations, serving as a critical resource for post-training alignment and a comprehensive benchmark for multimodal reasoning.

3.2 Evaluation Metrics

The performance of the proposed adaptation is evaluated based on answer accuracy and explanation quality. Answer correctness, denoted as Acc, employs synonym-aware matching to accommodate Vietnamese lexical variations. Furthermore, the SMILE metric [5], configured with a PhoBERT backbone [7], is utilized to capture semantic depth via references generated by LLMs. Regarding explanation quality, BERTScore, denoted as BS [14], is adopted to quantify semantic alignment in the embedding space rather than lexical overlap.

3.3 Experimental Setup

Our experiments utilize two backbones, which are Vintern-3B-R-beta [1], a Vietnamese-centric model, and InternVL3.5 [13], an open-source multimodal framework. Experiments are performed on an NVIDIA RTX A5000 GPU using 4-bit QLoRA optimization. The training process includes 1000 steps with a LoRA rank of 32, an alpha of 64, and a learning rate of 1×10^{-5}. The GRPO configuration employs a group size of $G = 4$, a temperature of 0.9, and a KL-penalty coefficient of 0.04.

3.4 Results

Table 1. Performance comparison on the ViVQA-X test set. Best scores are in **bold**. Second-best scores are underlined.

Method	**BS**	**SMILE**	**Acc**
Backbone: Vintern-3B-R-beta [1]			
Base	51.90	56.00	54.83
SFT	**53.69**	51.45	46.60
GRPO (DeepSeek Reward)	52.20	57.07	56.15
Ours	52.81	**60.42**	**62.65**
Backbone: InternVL3.5 [13]			
Base	52.10	**69.45**	55.28
SFT	52.20	69.00	56.20
GRPO (DeepSeek Reward)	52.14	69.14	54.98
Ours	**52.24**	65.47	**61.23**

Table 2. Impact of reasoning and explanation components.

Method	**Acc**	**SMILE**	**BS**
Baselines			
Base (Direct)	46.2	51.3	52.5
Base (CoT)	54.8	56.0	51.9
Ours			
w/o Reasoning	42.8	54.7	53.9
w/o Explanation	47.4	56.7	50.7
Full	**62.7**	**60.4**	**52.8**

Quantitative results on the ViVQA-X test set are presented in Table 1. Constrained to a 1000-step fine-tuning budget, SFT demonstrates limited adaptability. It results in a performance regression on the Vintern backbone and yields only marginal improvements on InternVL, suggesting that the model generalizes poorly under the supervised objective. This is because zero-shot baseline is prompted to proceed “think” via CoT while SFT trained on data lacking reasoning annotations is forced to bypass this reasoning process. In contrast, RL-based methods use group sampling to explore various reasoning paths within the same training budget. While the standard GRPO baseline already surpasses SFT in accuracy, our methodology further amplifies these gains by tailoring the reward mechanism to the specific requirements of VQA-NLE. This approach achieves the highest accuracy spanning both backbones, reaching **62.65%** on Vintern and **61.23%** on InternVL. These results indicate that the policy optimization prioritizes maximizing response accuracy, thereby guiding the model toward logically grounded conclusions.

4 Ablation Study

Table 2 investigates the impact of reasoning components, with all variants trained for 1000 steps. Explicit reasoning proves essential; its integration improves accuracy by 8.6% over direct inference. Conversely, removing CoT from GRPO reduces performance to **42.80%**, confirming that reasoning is a vital step for visual interpretation. Furthermore, the *w/o Explanation* variant, which aligns reasoning with ground-truth, underperforms the full framework. This suggests reasoning entails unstructured exploration, while explanation requires structured justification. Decoupling these objectives prevents constriction of the search space, allowing independent optimization of accuracy and communication quality.

5 Conclusion

In this paper, a GRPO adaptation for the Vietnamese VQA-NLE task is presented. By utilizing a composite reward mechanism to enforce structured reasoning, the method outperforms supervised baselines by 16.05% on the ViVQA-X benchmark. Additionally, ablation analyses validate the necessity of decoupling internal reasoning from explanation generation. Future works will investigate granular visual alignment metrics and extend this paradigm to other underrepresented languages.

References

1. Doan, K.T., et al.: Vintern-1b: An efficient multimodal large language model for vietnamese. arXiv preprint arXiv:2408.12480 (2024)
2. Du, M., et al.: Learning credible deep neural networks with rationale regularization. In: 2019 IEEE International Conference on Data Mining (ICDM). IEEE (2019)
3. Duong, T.B., et al.: An automated pipeline for constructing a vietnamese vqa-nle dataset. In: Proceedings of the International Conference on Intelligent Systems and Networks (ICISN). Springer (2025)
4. Kan, Z., et al.: Taco: Think-answer consistency for optimized long-chain reasoning and efficient data learning via reinforcement learning in lvlms. arXiv preprint arXiv:2505.20777 (2025)
5. Kendre, S., et al.: Smile: A composite lexical-semantic metric for question-answering evaluation. arXiv preprint arXiv:2511.17432 (2025)
6. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. Association for Computational Linguistics (2004)
7. Nguyen, D.Q., Nguyen, A.T.: Phobert: Pre-trained language models for vietnamese. In: Findings of the Association for Computational Linguistics: EMNLP 2020. Association for Computational Linguistics (2020)
8. Pan, J., et al.: Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer (2025)
9. Park, D.H., et al.: Multimodal explanations: Justifying decisions and pointing to the evidence. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (2018)
10. Rafailov, R., Sharma, A., et al.: Direct preference optimization: Your language model is secretly a reward model. Advances in neural information processing systems (2023)
11. Shao, Z., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024)
12. Wang, P., Ling, H.: Svqa-r1: Reinforcing spatial reasoning in mllms via view-consistent reward optimization. arXiv preprint arXiv:2506.01371 (2025)
13. Wang, W., et al.: Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025)
14. Zhang, T., et al.: Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675 (2019)