Bridging the Modality Gap Through CoT-Enhanced Multimodal Reasoning

Notebook for the ImageCLEF Multimodal Lab at CLEF 2025

Shengjun Deng, Guo Niu*, Xiongfei Yao, Huanlin Mo, Tao Li and Shuaiwei Jiao

Foshan University, Foshan, China

Abstract

This paper proposes a "Question Reconstruction before Answering" (QRA) prompting strategy for the ImageCLEF2025 multimodal reasoning task. The method first completes missing question stems using image information, then guides the language model through step-by-step reasoning and answering, thereby enhancing the model's comprehension and reasoning capabilities.

On the EXAMS-V dataset, through our investigation of different prompts and their impact on accuracy, we found that the QRA prompting demonstrates strong cross-lingual adaptability compared to conventional Chain-of-Thought (CoT) prompting. Experimental results show that this method effectively improves visual question answering performance without requiring OCR or additional fine-tuning, offering a new perspective for multimodal reasoning tasks.

Keywords

Multimodal reasoning, Vision-language models, Prompt engineering, Chain-of-thought

1. Introduction

With the rapid development and widespread application of Visual Language Models (VLMs), they have demonstrated significant potential in cross-modal information fusion and made progress in visual question-answering tasks. Existing methods primarily focus on feature-level alignment between visual encoders and language decoders [1, 2, 3], overlooking the crucial intermediate reasoning process. Although Chain-of-Thought (CoT) prompting has achieved success in unimodal text reasoning [4], it faces two challenges in multimodal scenarios: 1) the "modality gap" caused by relying solely on visual features disrupts language-based reasoning chains; 2) traditional CoT lacks the ability to reconstruct missing question text from visual data. Experiments show that our strategy is effective across multiple languages on EXAMS - V.

In this study, focusing on ImageCLEF25 [5, 6], we propose a novel strategy. We address multimodal reasoning by reconstructing missing question text and integrating them with CoT. On the EXAMS-V dataset [7], our method significantly improves the performance of VLMs in this task by leveraging image features to reconstruct question text, thereby providing more information for CoT reasoning.

2. Related Work

2.1. Visual-Language Models

Visual-Language Models (VLMs) have made significant progress in multimodal understanding tasks in recent years. CLIP[1] established the foundation for multimodal pretraining by constructing a general image-text embedding space through contrastive learning of images and text. BLIP-2[2] introduced a

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

*Corresponding author.

everyonesky@qq.com (S. Deng); sreedomly@163.com (G. Niu); 3084524042@qq.com (X. Yao); 277228089@qq.com (H. Mo); trent0@foxmail.com (T. Li); 2112353048@stu.fosu.edu.cn (S. Jiao)
0009-0003-0089-2651 (S. Deng); 0000-0002-1552-7310 (G. Niu); 0009-0009-8464-8448 (X. Yao); 0009-0006-6802-5521 (H. Mo); 0009-0007-1348-2060 (T. Li); 0009-0000-3805-3237 (S. Jiao)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

lightweight intermediate module to connect a frozen visual encoder with a language model, enhancing image-text question answering and generation capabilities. LLaVA[3] combined CLIP and LLM, adding a projection layer to improve the model's understanding of images through visual instruction fine-tuning, supporting various question answering and dialogue scenarios. VisionLLM[8] optimized the visual attention mechanism based on BLIP-2, achieving more refined image-text alignment. Qwen-VL 2.5[9] further expanded the model's perceptual capabilities by optimizing the projection layer and other methods, demonstrating strong reasoning abilities with excellent performance on multiple benchmarks.

Although these methods have made progress in image-text alignment and language generation, their reasoning processes still perform limitedly under incomplete prompts. Our approach attempts to address this shortcoming by reconstructing question text combined with CoT (Chain-of-Thought) reasoning.

2.2. Chain-of-Thought Prompt

Chain-of-Thought (CoT) prompting significantly enhances the reasoning capabilities of large language models in complex tasks by guiding the model to generate intermediate reasoning steps [4]. In scenarios such as mathematics and commonsense question answering, CoT helps the model decompose problems step-by-step and generate coherent reasoning chains, thereby improving accuracy. However, applying CoT to multimodal tasks still faces challenges. On one hand, CoT typically relies on explicit textual prompts, but key information in multimodal tasks may exist in visual form, making it difficult for the model to correctly understand the problem. On the other hand, visual features lack clear semantic boundaries, and directly inputting them into the language model often leads to prompt interpretation deviations due to the "modality gap," which in turn affects the completeness and logicality of the reasoning chain.

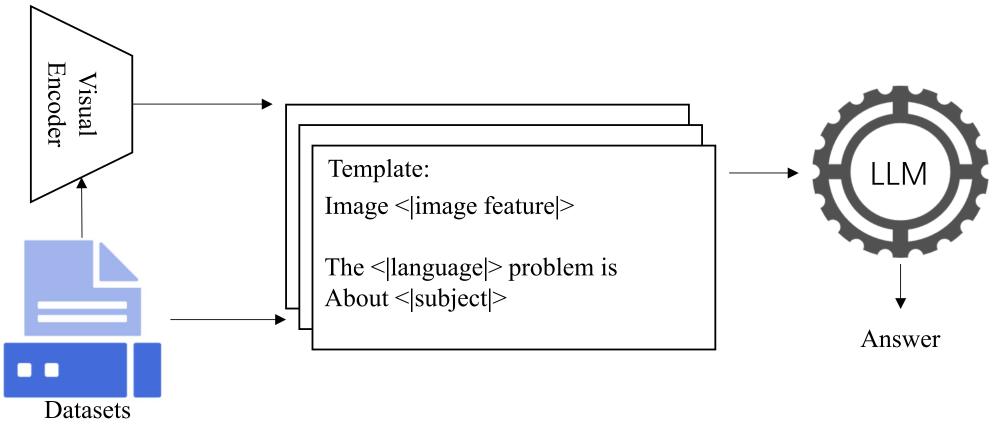


Figure 1: Architecture of our strategy.

3. Method

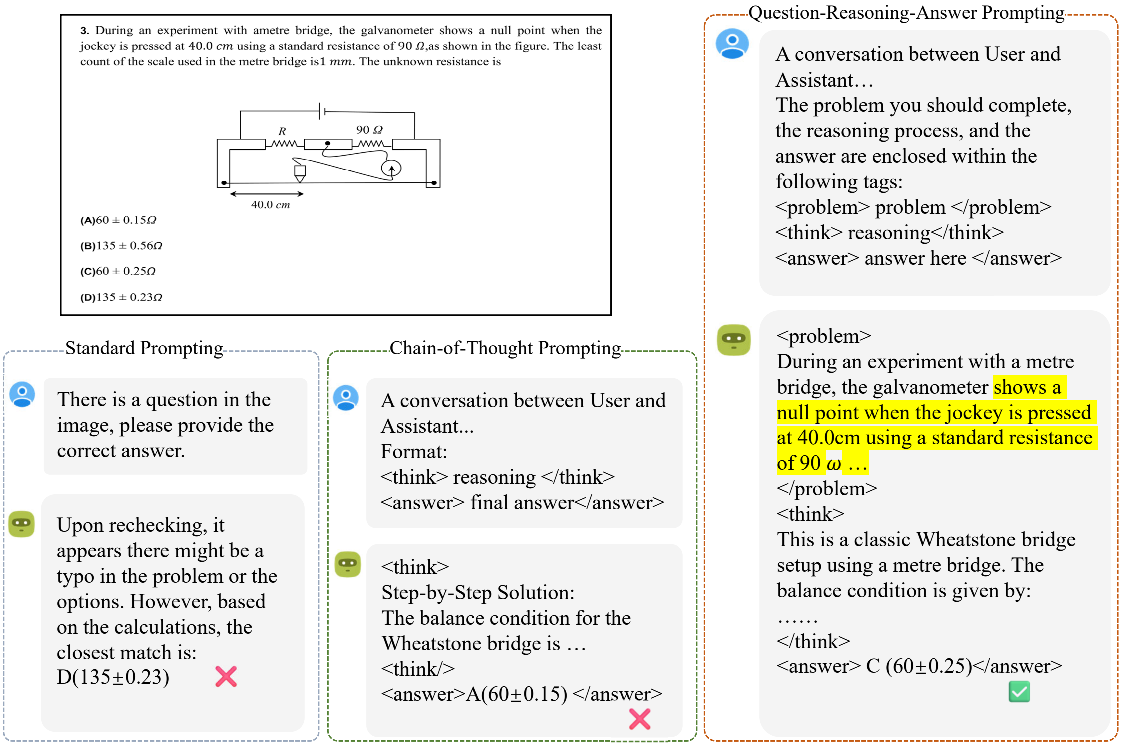
In image-only multimodal question answering tasks,such as the EXAMS-V dataset [7], visual encoders often lead to the loss of certain semantic information when abstractly representing images, particularly the textual cues and detailed content in the images that are relevant to the question stem. This information gap makes it difficult for language models to construct clear reasoning chains. In contrast, when faced with obscured or incomplete questions, humans are usually able to reasonably complete the missing information based on their existing background knowledge and contextual understanding, thereby successfully completing the reasoning task.

Inspired by this, we propose a "complete first, then reason" strategy, the Architecture shown in Figure 1. This strategy first uses image features to guide the language model to complete the missing question information, thereby reconstructing the complete question stem; subsequently, based on the reconstruction results, a Chain-of-Thought (CoT) reasoning mechanism is introduced to enhance the model's cross-modal reasoning ability. This method not only enhances the model's understanding of the task context but also effectively alleviates the semantic disconnection caused by modal differences. Specifically, our method includes two key steps: 1) Question Background Information Prompt Embedding , 2) Question-Reasoning-Answer Prompting.

3.1. Question Background Information Prompt Embedding

In practical multimodal question answering tasks, questions often involve specific languages and subject backgrounds, with language expressions that are highly specialized and context-dependent. Especially in scenarios containing only images, language models, lacking explicit context, are prone to misunderstandings of the question stem.

To address this issue, we introduce question background information embedding. Specifically, we extract the language category (e.g., English, French, etc.) and subject labels (e.g., physics, chemistry, etc.) of the question from the image's metadata and use them as prior knowledge prompt words to guide the language model in context modeling. This approach effectively mitigates semantic ambiguity caused by language specificity, making the model more targeted and accurate when generating completion content.


R
90 Ω
40.0 cm
Standard Prompting
Chain-of-Thought Prompting
Question-Reasoning-Answer Prompting

3. During an experiment with ametre bridge, the galvanometer shows a null point when the jockey is pressed at 40.0 cm using a standard resistance of 90 Ω,as shown in the figure. The least count of the scale used in the metre bridge is1 mm. The unknown resistance is

(A)$60 \pm 0.15\Omega$

(B)$135 \pm 0.56\Omega$

(C)$60 + 0.25\Omega$

(D)$135 \pm 0.23\Omega$

Standard Prompting

There is a question in the image, please provide the correct answer.

Upon rechecking, it appears there might be a typo in the problem or the options. However, based on the calculations, the closest match is: D(135±0.23) ✗

Chain-of-Thought Prompting

A conversation between User and Assistant...
Format:
<think> reasoning </think>
<answer> final answer</answer>

<think>
Step-by-Step Solution:
The balance condition for the Wheatstone bridge is …
<think/>
<answer>A(60±0.15) </answer> ✗

Question-Reasoning-Answer Prompting

A conversation between User and Assistant…
The problem you should complete, the reasoning process, and the answer are enclosed within the following tags:
<problem> problem </problem>
<think> reasoning</think>
<answer> answer here </answer>

<problem>
During an experiment with a metre bridge, the galvanometer shows a null point when the jockey is pressed at 40.0cm using a standard resistance of 90 ω …
</problem>
<think>
This is a classic Wheatstone bridge setup using a metre bridge. The balance condition is given by:
……
</think>
<answer> C (60±0.25)</answer>
✓

Figure 2: Question - Reasoning - Answer Prompting enables large language models to better understand questions, thereby handling complex reasoning without confusion. The emphasized Question process demonstrates the model's attempt to complete the question.

3.2. Question Reconstruction before Answering Prompting

After completing the question stem, the model still requires strong reasoning capabilities to correctly perform the question-answering task. Traditional Chain-of-Thought (CoT) prompting, which guides language models to generate intermediate reasoning steps, has achieved significant success in textual reasoning tasks. However, directly applying the CoT mechanism to multimodal question-answering tasks involving only images can lead to information confusion or insufficient semantic alignment, resulting in the model's inability to construct coherent and clear reasoning chains.

To address this, we propose a structured "Question Reconstruction before Answering" guided prompting strategy, aiming to explicitly separate the question comprehension process from the reasoning process to enhance the model's ability to build reasoning chains. Specifically, we design a unified prompt template that introduces the <Question>...</Question> tag to guide the model in first understanding the question before engaging in step-by-step thinking and answering. We show the effects of three types of prompts in Figure 2.

Table 1
Overview of the Multilingual results and rankings across all test sets on ImageClef2025 (Multimodal Reasoning).

Rank	Team	Method	ACC
6	**deng113abc (Ours)**	QRA Prompting	**0.5195**
	MSA	-	0.8140
	ymgclef	-	0.5994
	lekshmiscopevit	-	0.5770
	bingezzzleep	-	0.5619
	plutohbj	-	0.5226
	mhh2001	-	0.4418
	yaozihang	-	0.4376
	baseline*	-	0.2701
	elenat	-	0.2188

4. Results

4.1. Comparative Experiments

To validate the effectiveness of our proposed QRA Prompting strategy, we participated in the ImageCLEF2025 multimodal reasoning task and submitted test results for both the Multilingual Track and the English Track. Table 1 lists our performance on the multilingual test set.

In the Multilingual Track, our method ranked 6th among all participating teams, achieving an accuracy of 0.5195. Compared to the official baseline method (accuracy of 0.2701), our approach improved performance by 24.9%, demonstrating the strong competitiveness of our method in practical tasks. This significant improvement indicates that our proposed structured strategy of "first completing the question, then reasoning" has clear advantages in alleviating inter-modal information misalignment and enhancing cross-modal understanding.

Notably, we achieved near-top-tier performance without relying on any additional OCR modules or fine-tuning the model for multilingual tasks. This demonstrates that QRA Prompting possesses strong robustness and excellent transfer generalization capabilities, performing stably and reliably in complex real-world multimodal reasoning scenarios.

In the English Track, we also submitted model predictions based on QRA Prompting, achieving an accuracy of 0.5371 and ranking 6th,As shown in Table 2. Our method consistently delivered strong performance across both tasks, further validating its cross-language consistency.

4.2. Ablation Study

To systematically evaluate the contributions of each key component in QRA Prompting, we conducted ablation experiments on the English validation set of the EXAMS-V dataset. The experiments used Qwen-VL-2.5-32B as the base model, employed a zero-shot setting, and compared against standard Prompting and Chain-of-Thought (CoT) Prompting methods. As shown in Table 3, the standard Prompting method achieved an accuracy of 0.458, demonstrating relatively weak performance. The CoT Prompting method,

Table 2
Overview of the English version results and rankings across all test sets on ImageClef2025 (Multimodal Reasoning).

Rank	Team	Method	ACC
6	**deng113abc (Ours)**	QRA Prompting	**0.5371**
	stormhunter44	-	0.8965
	MSA	-	0.8652
	ayeshaamjad	-	0.8125
	heavyhelium	-	0.8086
	ymgclef	-	0.5938
	bingezzzleep	-	0.5312
	plutohbj	-	0.4922
	mhh2001	-	0.4629
	yaozihang	-	0.4570
	elenat	-	0.2520
	baseline*	-	0.2480

which guides the model through chain-of-thought reasoning, improved accuracy to 0.548. The QRA Prompting strategy further enhanced this performance, achieving an accuracy of 0.582, which represents a 12.4% improvement over the standard method and a 3.4% improvement over the CoT method.

These results indicate that QRA Prompting not only inherits the advantages of chain-of-thought reasoning from CoT but also effectively enhances the language model's understanding of image semantics through explicit question stem completion, significantly boosting the model's performance in complex reasoning tasks.

Table 3
blation results of effect of our QRA strategy.

Method	Model	ACC
standard Prompting	Qwen-VL-2.5-32B	0.458
CoT Prompting	Qwen-VL-2.5-32B	0.548
QRA Prompting	Qwen-VL-2.5-32B	0.582

5. Conclusion

In this study, we addressed the Multimodal Reasoning task of the ImageCLEF2025 Multimodal Lab. By employing the QRA strategy, we enhance the inference accuracy of models in visual question answering tasks. Our approach involves constructing QRA prompt templates and integrating contextual information. These strategies effectively address two key challenges faced by traditional Chain-of-Thought (CoT) in multimodal scenarios: they alleviate the "modality gap" problem caused by relying solely on visual features and enhance the ability to reconstruct missing question text from visual data.

Evaluation results demonstrate the feasibility and effectiveness of our method, achieving an accuracy of 0.5195 on the multilingual version of the EXAMS-V test set. These findings indicate that our approach provides a viable solution for visual question answering tasks that use only visual features, contributing to the field of multimodal reasoning.

Acknowledgments

This work is supported by the Research Projects of OrdinaryUniversities in Guangdong Province under Grant2023KTSCX133, the Guangdong Basic and Applied Basic Research Foundation under Grant 2022A1515140103.

Declaration on Generative AI

During the preparation of this work, the author(s) used deepseek-v3 in order to: Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, PmLR, 2021, pp. 8748–8763.

[2] J. Li, D. Li, S. Savarese, S. Hoi, Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, in: International conference on machine learning, PMLR, 2023, pp. 19730–19742.

[3] H. Liu, C. Li, Q. Wu, Y. J. Lee, Visual instruction tuning, 2023. URL: https://arxiv.org/abs/2304.08485. arXiv:2304.08485.

[4] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al., Chain-of-thought prompting elicits reasoning in large language models, Advances in neural information processing systems 35 (2022) 24824–24837.

[5] B. Ionescu, H. Müller, D.-C. Stanciu, A.-G. Andrei, A. Radzhabov, Y. Prokopchuk, Ştefan, Liviu-Daniel, M.-G. Constantin, M. Dogariu, V. Kovalev, H. Damm, J. Rückert, A. Ben Abacha, A. García Seco de Herrera, C. M. Friedrich, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, C. S. Schmidt, T. M. G. Pakull, B. Bracke, O. Pelka, B. Eryilmaz, H. Becker, W.-W. Yim, N. Codella, R. A. Novoa, J. Malvehy, D. Dimitrov, R. J. Das, Z. Xie, M. S. Hee, P. Nakov, I. Koychev, S. A. Hicks, S. Gautam, M. A. Riegler, V. Thambawita, P. Halvorsen, D. Fabre, C. Macaire, B. Lecouteux, D. Schwab, M. Potthast, M. Heinrich, J. Kiesel, M. Wolter, B. Stein, Overview of imageclef 2025: Multimedia retrieval in medical, social media and content recommendation applications, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the 16th International Conference of the CLEF Association (CLEF 2025), Springer Lecture Notes in Computer Science LNCS, Madrid, Spain, 2025.

[6] D. Dimitrov, M. S. Hee, Z. Xie, R. Jyoti Das, M. Ahsan, S. Ahmad, N. Paev, I. Koychev, P. Nakov, Overview of imageclef 2025 – multimodal reasoning, in: CLEF 2025 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Madrid, Spain, 2025.

[7] R. J. Das, S. E. Hristov, H. Li, D. I. Dimitrov, I. Koychev, P. Nakov, Exams-v: A multi-discipline multilingual multimodal exam benchmark for evaluating vision language models, 2024. URL: https://arxiv.org/abs/2403.10378. arXiv:2403.10378.

[8] W. Wang, Z. Chen, X. Chen, J. Wu, X. Zhu, G. Zeng, P. Luo, T. Lu, J. Zhou, Y. Qiao, J. Dai, Visionllm: Large language model is also an open-ended decoder for vision-centric tasks, 2023. URL: https://arxiv.org/abs/2305.11175. arXiv:2305.11175.

[9] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, J. Lin, Qwen2.5-vl technical report, 2025. URL: https://arxiv.org/abs/2502.13923. arXiv:2502.13923.