Multimodal Reasoning in Multilingual Visual Question Answering: A Prompt-Tuned Qwen2.5-vl-plus Approach

Notebook for the ImageCLEF Lab at CLEF 2025

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

Foshan University, Foshan, China

Abstract

We propose a prompt-tuning approach based on Qwen2.5-vl-plus for the MultimodalReason task at ImageCLEF 2025, which involves answering multiple-choice questions grounded in images across multiple languages and complex reasoning scenarios. Our method achieves an accuracy of 0.4418 on the benchmark, representing a 63% improvement over the baseline SmolVLM (0.2701). Further analysis indicates that well-designed prompt templates play a crucial role in enhancing the model's cross-lingual reasoning performance.

Keywords

Multilingual, Multimodal Reasoning, Vision-Language Models, Prompt Engineering, ImageCLEF 2025

1. Introduction

Multimodal reasoning has become a key research focus in the field of artificial intelligence, particularly due to its wide-ranging applications in tasks that integrate vision and language [1]. In recent years, although large multimodal models have achieved significant progress in image-text understanding tasks [2, 3], they still face considerable challenges in modeling the complex semantic relationships between images and text in real-world multilingual environments [4, 5].

To systematically evaluate models' comprehensive capabilities in multilingual and multimodal contexts, CLEF 2025 introduced the MultimodalReason task [6, 7], which centers on Multilingual Visual Question Answering (VQA). In this task, models are required to understand an image containing a question along with four candidate answers and accurately identify the single correct option. This setting demands the integration of image understanding, multilingual text processing, and logical reasoning, closely reflecting real-world scenarios involving cross-language and cross-modal information processing [8, 9].

In this study, we adopt Qwen2.5-VL-Plus as our primary model. This model is capable of handling both image and multilingual text inputs and has demonstrated strong performance in various multimodal benchmarks [10]. Compared to the baseline model SmolVLM, which uses a single system prompt, we further introduce a hybrid prompt design that combines system instructions with exemplar-based few-shot prompting [11, 12] to better activate the model's reasoning capabilities.

Experimental results show that Qwen2.5-VL-Plus performs well in multilingual visual question answering tasks, particularly in integrating visual cues with multilingual expressions. Our approach achieved excellent results in the competition; however, there is still room for improvement when handling more complex cross-modal reasoning samples. We hope this research provides practical experience and theoretical insights for the development of multilingual multimodal models and serves as a valuable reference for future work.

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

mhl\protect1_100@foxmail.com (H. Mo); sreedomly@163.com (G. Niu); 1137922877@qq.com (S. Deng); 3084524042@qq.com (X. Yao); trent0@foxmail.com (T. Li); 2112353048@stu.fosu.edu.cn (S. Jiao)

0009-0006-6802-5521 (H. Mo); 0000-0002-1552-7310 (G. Niu); 0009-0003-0089-2651 (S. Deng); 0009-0009-8464-8448 (X. Yao); 0009-0007-1348-2060 (T. Li); 0009-0000-3805-3237 (S. Jiao)

2. Related Work

2.1. Multimodal Vision-Language Models

Recent advances in vision-language models (VLMs) have significantly improved performance on tasks that require understanding both visual and textual inputs, such as image captioning, visual question answering (VQA), and visual entailment. Foundational models like CLIP [13], Flamingo [2], and BLIP [3] have demonstrated the effectiveness of joint vision-language pretraining. More recent models such as MiniGPT-4 [9] and LLaVA [8] combine large language models (LLMs) with vision encoders to enable open-ended multimodal reasoning.

However, most of these models are primarily trained on English-centric datasets and often rely on pattern recognition rather than deep reasoning. Their performance on complex logical inference, especially in multilingual and real-world settings, remains limited.

2.2. Multilingual Visual Question Answering

Multilingual VQA aims to evaluate a model's ability to understand and reason about images and text across different languages. Prior work in this area is relatively sparse compared to English-only VQA benchmarks such as VQAv2 [14] or GQA [15]. Some efforts, such as MaXM [4], explore multilingual alignment, but many VLMs still underperform on low-resource or morphologically rich languages.

The MultimodalReason task introduced by CLEF 2025 provides a more realistic and challenging multilingual setting by requiring models to process questions presented in various languages (e.g., English, Chinese, Spanish) while reasoning over visual content and selecting one correct answer from multiple options.

2.3. Multimodal Reasoning and Prompt Engineering

Deep reasoning in multimodal contexts remains a major challenge. While recent LLM-augmented VLMs (e.g., GPT-4V, Qwen-VL [10]) demonstrate better reasoning performance than early models, they still struggle with tasks that involve hypothetical scenarios, abstract logic, or long-range dependencies between visual and textual elements.

Prompt engineering has emerged as an effective technique to steer model behavior without fine-tuning. In-context learning via exemplars or task-specific instruction formatting can significantly enhance performance on reasoning tasks [12, 11]. In multimodal settings, hybrid prompting strategies that combine visual inputs with structured textual cues (e.g., few-shot examples, multilingual instructions) have shown promise, but their impact in multilingual VQA is still under-explored.

3. Method

3.1. Baseline System

The official baseline for this task employs the SmolVLM model, a lightweight vision-language model optimized for inference efficiency.

Using only a default system prompt,which can be seen in Figure 1, this model achieved an overall accuracy of 16% on the development set. The default prompt includes only minimal instruction (e.g., "You are a helpful assistant") and lacks task-specific context or examples.

3.2. Our Approach: Prompt Engineering with Exemplars and Model Upgrade

To improve performance, we adopt a two-pronged strategy:

- **Model Upgrade:** We replace the baseline SmolVLM with the more capable Qwen2.5-VL-Plus model, which has demonstrated stronger reasoning capabilities across multiple vision-language tasks.

```
baseline_prompt= [
        {
          "type": "text",
          "text": """You are a sophisticated Vision-Language Model (VLM) capable of analyzing images
           containing multiple-choice questions, regardless of language. To guide your analysis, you may
           adopt the following process:
          1. Examine the image carefully for all textual and visual information.
          2. Identify the question text, even if it's in a different language.
          3. Extract all answer options (note: there may be more than four).
          4. Look for additional visual elements such as tables, diagrams, charts, or graphs.
          5. Ensure to consider any multilingual content present in the image.
          6. Analyze the complete context and data provided.
          7. Select the correct answer(s) based solely on your analysis.
          8. Respond by outputting only the corresponding letter(s) without any extra explanation.""",
        }
```

Figure 1: Baseline prompt used in the official system.

- **Hybrid Prompt Design:** We introduce a hybrid prompt structure that combines:
  1. A system prompt that defines the model's role and multilingual capabilities (Figure 2).
  2. One or more in-context examples ("sample prompts") drawn from the training set (Figure 3). Each includes an image, a multilingual question, five candidate answers (A–E), and the correct answer.

```
SYSTEM_PROMPT_TEXT = (
  "A conversation between User and Assistant. The user asks a question, and
   the Assistant solves it. "
  "The assistant first repeats the problem within <problem></problem> tags
   according by image, "
  "then thinks about the reasoning process within <think></think> tags, "
  "and finally provides the answer within <answer></answer> tags. "
  "i.e., <problem> the user' s question here </problem><think> reasoning
   process      here </think><answer> answer here </answer>"
)
```

Figure 2: System prompt design

```
sample_prompt = (
        f"The problem language is {item['language']} "
        f"and subject is {item['subject']}. "
        f"Only one option is true."
        "Provide the final answer (option) in <answer></answer> tags, "
        "<answer>your option</answer>,like <answer>A</answer>. Don't
         answer  any other text and don't answer with other letters."
        "Option must be A B C D E upercase."
        "Your response:"
    )
```

Figure 3: sample prompt

3.3. System Architecture

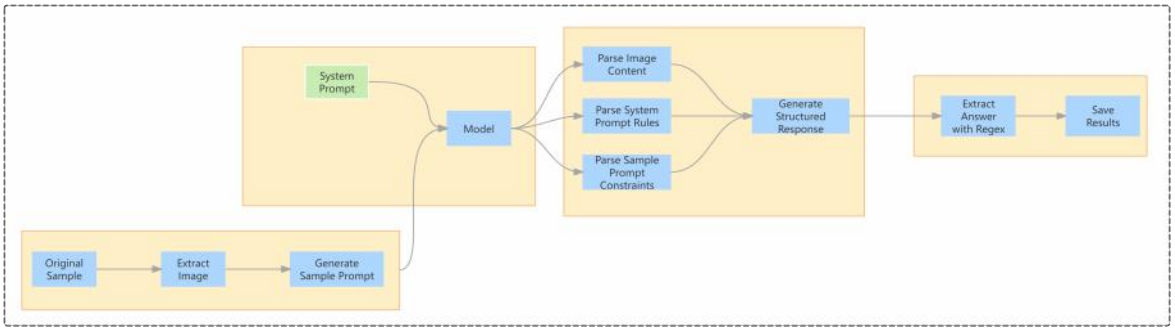

Figure 4: Overall pipeline architecture for hybrid prompting

As shown in Figure 4, we design a pipeline centered around the upgraded Qwen2.5-VL-Plus model. The original image-question sample is processed to extract key information and generate the corresponding sample prompt. Together with the system prompt, this is fed into the model.

The system prompt provides macro-level instructions, while the sample prompt offers task-specific context. The model parses the image content, task rules, and exemplar structure, and generates a response. Finally, the answer is extracted via regular expressions and saved.

This design ensures accurate and efficient multimodal reasoning and output generation.

3.4. Analysis of Prompt Strategies

Prompt design plays a vital role in multimodal visual question answering tasks. We analyze the limitations of the baseline prompt and advantages of our hybrid prompt strategy below.

Limitations of Baseline Prompt

- **Generalization Issues:** The baseline prompt lists general analysis steps without enforcing output format. This lack of structure often leads to noisy or incomplete answers, especially in multilingual contexts.
- **No Structured Reasoning Guidance:** The prompt fails to explicitly guide reasoning or require intermediate steps. Thus, even when the model arrives at the right answer, it's unclear whether it followed a logical path or guessed .

3.5. Implementation Details

We preprocess all images to a resolution of 448×448 pixels. For text input, we use the official tokenizer and image processor from the Qwen2.5-VL-Plus repository. Prompts are inserted in a zero-shot format unless otherwise specified. The model response is decoded as free text, and the final prediction is determined by matching it to one of the five answer choices (A–E).

3.6. Evaluation Methodology

The evaluation of the MultimodalReason task is centered around a straightforward yet crucial metric: accuracy. Given that the task requires participants to identify the single correct answer from a set of four options presented in an image - based question, accuracy serves as the primary indicator of a model's performance. It directly reflects the proportion of correctly answered questions out of the total number of questions in the dataset.

4. Results

4.1. Dataset

Our dataset, the cornerstone of the MultimodalReason task, is accessible via "Exams-V" [16]. It is partitioned into 16,724 training instances and 4,208 development/validation instances. The test data will be released subsequently.

The EXAMS-V dataset is a meticulously curated, multi-disciplinary, multimodal, and multilingual benchmark. It encompasses 20,932 multiple-choice questions from 20 disciplines, spanning natural science, social science, and fields like religion, fine arts, and business.

EXAMS-V stands out with its rich multimodal features, including text, images, tables, graphs, charts, maps, scientific symbols, and equations. Questions are presented in 11 languages from 7 language families.

Unlike typical benchmarks, EXAMS-V is assembled from school exam questions across various countries and educational systems. This diverse origin endows the dataset with complexity, requiring models to navigate language barriers, understand question nuances, and apply region-specific knowledge for reasoning.

Here is a snapshot of the dataset's statistics (languages ranked from high- to low-resource):

Table 1
Dataset statistics

Language	Language Family	Grade	# of Subjects	# of Questions	# of Multimodal Questions	# of Text - only Questions
English	Germanic	11, 12	4	724	181	543
Chinese	Sino - Tibetan	8–12	6	2,635	1,991	644
French	Romance	12	3	439	50	389
German	Germanic	12	5	819	144	675
Italian	Romance	12	11	1,645	292	1,353
Arabic	Semitic	4–12	6	823	117	706
Polish	Slavic	12	1	2,511	422	2,089
Hungarian	Finno - Ugric	12	6	3,801	495	3,306
Bulgarian	Slavic	4, 12	4	2,132	435	1,697
Croatian	Slavic	12	13	3,969	700	3,269
Serbian	Slavic	12	11	1,434	259	1,175

4.2. Experimental Results

The analysis across multiple languages strikingly demonstrates that our approach has achieved remarkable enhancements in multilingual performance. Specifically, the "mhl2001 Score" for the multilingual evaluation has soared from 0.2701 to 0.4418, marking a significant improvement. This showcases the effectiveness of our system in handling diverse languages simultaneously.

Notably, among individual languages, Chinese and German have witnessed substantial progress. In Chinese, the score has leaped from 0.2678 to 0.5553, a remarkable 107% increase, while in German, it has risen from 0.3101 to 0.4922, a 58.7% increase. These gains are attributed to the enhanced language modeling capabilities of Qwen2.5 - VL - Plus and the meticulously crafted prompts that capture the structural intricacies of multiple - choice reasoning questions.

Even for relatively low - resource languages like Hungarian, our model still exhibits a notable performance boost, with the score advancing from 0.2348 to 0.3563. This indicates the model's proficiency in cross - lingual generalization without the need for language - specific fine - tuning.

Overall, these results unequivocally prove that our system not only elevates accuracy but also provides a more robust and scalable solution for multimodal reasoning tasks across a wide spectrum of linguistic contexts.

Language	mhl2001 Score	Baseline Score
Multilingual	0.4418	0.2701
English	0.4629	0.2480
Chinese	0.5553	0.2678
German	0.4922	0.3101
Arabic	0.4775	0.2703
Hungarian	0.3563	0.2348

Table 2
Comparison of mhl2001 and baseline scores across different languages in the ImageCLEF Multimodal Reasoning task.

5. Conclusion

This paper proposes a simple yet effective approach to the CLEF 2025 MultimodalReason task by combining a stronger vision-language model, Qwen2.5-VL-Plus, with a carefully crafted hybrid prompt that integrates multilingual system instructions and exemplar-based few-shot learning. Without any task-specific fine-tuning, our method significantly improves overall accuracy from 0.2701 to 0.4418, achieving consistent gains across all 11 languages in EXAMS-V, including notable improvements in

low-resource languages like Hungarian. The results confirm that model capability and prompt design jointly play a crucial role in enhancing multilingual multimodal reasoning.

While our method performs well on multiple-choice VQA tasks, challenges remain in handling complex images with dense text, domain-specific knowledge questions, and languages beyond the EXAMS-V set. Future work will explore dynamic exemplar selection, step-by-step rationale generation, lightweight parameter tuning (e.g., LoRA), and knowledge grounding via external resources, aiming to further boost performance in challenging multilingual settings.

Acknowledgments

This work is supported by the Research Projects of Ordinary Universities in Guangdong Province under Grant 2023KTSCX133, 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] T. Baltrušaitis, C. Ahuja, L.-P. Morency, Multimodal machine learning: A survey and taxonomy, IEEE Transactions on Pattern Analysis and Machine Intelligence 41 (2018) 423–443.

[2] J.-B. Alayrac, et al., Flamingo: a visual language model for few-shot learning, arXiv preprint arXiv:2204.14198 (2022).

[3] J. Li, D. Li, C. Xiong, S. C. Hoi, Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, Proceedings of the International Conference on Machine Learning (ICML) (2022).

[4] J. Li, et al., Maximizing multilingual multimodal learning with prompt engineering, arXiv preprint arXiv:2306.05450 (2023).

[5] Z.-Y. Dou, et al., Coarse-to-fine vision-language pre-training with fusion in the backbone, Advances in Neural Information Processing Systems 35 (2022) 16650–16663.

[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] 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, H. M. Shan, 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.

[8] H. Liu, C. Zhang, et al., Visual instruction tuning, arXiv preprint arXiv:2304.08485 (2023).

[9] D. Zhu, et al., Minigpt-4: Enhancing vision-language understanding with advanced large language models, arXiv preprint arXiv:2304.10592 (2023).

[10] B. Inc., Qwen-vl: A multimodal foundation model for language, vision, and more, arXiv preprint arXiv:2403.09047 (2024).
[11] T. Kojima, et al., Large language models are zero-shot reasoners, arXiv preprint arXiv:2205.11916 (2022).
[12] T. B. Brown, et al., Language models are few-shot learners, in: Advances in Neural Information Processing Systems (NeurIPS), 2020.
[13] 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.
[14] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, D. Parikh, Making the v in vqa matter: Elevating the role of image understanding in visual question answering, in: CVPR, 2017, pp. 6904–6913.
[15] D. A. Hudson, C. D. Manning, Gqa: A new dataset for real-world visual reasoning and compositional question answering, in: CVPR, 2019, pp. 6700–6709.
[16] R. J. Das, S. E. Hristov, H. Li, D. I. Dimitrov, Ivan, Exams-v: A multi-discipline multilingual multimodal exam benchmark for evaluating vision language, 2024. arXiv:2403.10378.