Enhancing Multilingual VQA with Structured Prompts and Vision-Language Alignment*

Notebook for the ImageCLEF Lab at CLEF 2025

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

Foshan University, Foshan, China

Abstract

Multilingual Visual Question Answering (VQA) represents a crucial research direction in vision-language models (VLM). Although existing VLMs demonstrate strong performance in tasks like image captioning and simple visual dialogue, they still face significant challenges in reasoning over complex logical relationships and hypothetical scenarios. To systematically evaluate large language models' (LLM) reasoning capabilities in multilingual and multimodal contexts, we participated in CLEF 2025's newly proposed MultimodalReason task. This task requires models to identify the unique correct answer from multiple candidates based on an image and a question posed in multiple languages, including English and Chinese. We propose a novel reasoning approach that integrates image compression, structured prompt construction, and vision-language alignment, and embed it into an advanced chat model. Our method, optimized especially for English and Chinese, achieves leading performance in these sub-tasks, demonstrating its effectiveness for complex multilingual multimodal reasoning.

Keywords

Vision-Language Models, Structured Prompting, Large Language Models, Multimodal Reasoning

1. Introduction

Large Vision-Language Models (LVLMs) [1, 2, 3] represent a pivotal breakthrough in the field of artificial intelligence, marking a transformation in multimodal understanding and interaction. Multilingual Visual Question Answering (VQA) is an important research direction within the domain of Visual Language Models (VLMs). Enhancements in visual encoders [4] and improvements in resolution scaling [5] have played a crucial role in advancing the quality of practical visual understanding.

This study focuses on the ImageCLEF25 task [6, 7], which requires models to identify the single correct answer from a question that includes an image and 3–5 candidate options. The questions span multiple languages, including English and Chinese, and cover a wide range of disciplines and contexts. On the EXAMS-V dataset [8], we propose a reasoning approach that integrates image compression, structured prompt construction, and visual-language alignment, and incorporate it into an advanced conversational model. Given our strong proficiency in English and Chinese, we specifically optimize our method for these two language subtasks. Experimental results demonstrate that our approach achieves leading performance in both language tracks.

2. Related Work

Researchers have proposed a new multimodal understanding task, named MultimodalReason, aimed at enhancing models' joint semantic reasoning capabilities between images and text. This task requires models to determine whether a given textual description of an image is reasonable, representing a key challenge in the field of multimodal natural language understanding.

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

*Corresponding author.

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

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

As early as 2022, the Visual Commonsense Reasoning (VCR) task was introduced, emphasizing causal reasoning and multiple-choice answering based on image content. It served as an early representative of multimodal reasoning tasks. Subsequently, tasks such as NLVR2 and GQA further raised the bar for models by requiring stronger consistency between image and text and fine-grained visual understanding.

In 2023, researchers proposed the MaMMUT framework, which unified various multimodal tasks under a language modeling paradigm. By jointly training on large-scale image-text data, the model achieved outstanding performance on multimodal reasoning tasks, including MultimodalReason. Meanwhile, models like BLIP-2 and MiniGPT-4 explored the decoupled integration of visual encoders and large language models (LLMs), demonstrating strong generalization and scalability in tasks such as image-text reasoning and question answering. To further enhance reasoning capabilities, LLaVA introduced a method of integrating visual information into the language model context via prompts, enabling joint vision-language understanding without modifying the parameters of the language model.

To streamline the overall network structure, Qwen2.5-VL also aligns the architecture of the Vision Transformer (ViT) more closely with the design principles of LLMs. Specifically, it adopts RMSNorm [9] for normalization and SwiGLU [10] as the activation function. These design choices not only improve computational efficiency but also enhance compatibility between the vision and language components, further improving the model's multimodal understanding performance.

Inspired by the strong cross-task generalization capabilities of LLMs, the MultimodalReason task has begun incorporating multimodal models similar to ChatGPT. In terms of prompt design, structured prompts can effectively guide the model to attend to image regions relevant to key textual elements, thereby improving the accuracy of semantic reasoning. Studies show that structured prompts play an equally important guiding role in multimodal tasks, especially under few-shot learning or open-domain conditions, where they demonstrate superior generalization.

In task construction, researchers have attempted to enhance models' multimodal understanding by converting image content into text and then inputting both modalities into the model.

Building on these advancements, we propose a multimodal reasoning task framework that integrates image-text paired data, structured prompt design, and the strong language understanding capabilities of models like Qwen2.5-VL, aiming to improve their performance on multimodal reasoning tasks involving complex semantic relationships.

3. EXAMS-V Dataset

EXAMS-V is a comprehensive multilingual and multimodal dataset designed to evaluate the visual reasoning capabilities of AI systems, particularly Visual Language Models (VLMs). The dataset contains 24,856 multiple-choice questions. It supports 13 languages—namely English, Arabic, Chinese, German, Bulgarian, Italian, Spanish, Urdu, Polish, Hungarian, Serbian, and Croatian—and covers a wide range of academic subjects. The questions are drawn from real educational curricula across different regions and countries, enhancing the dataset's diversity, authenticity, and level of difficulty. The dataset is divided into 16,500 training samples, 4,000 validation samples, and 3,570 test samples. Successfully answering questions in EXAMS-V requires not only text comprehension but also the ability to interpret visual layouts, analyze tables and charts, and perform multimodal reasoning that integrates visual and linguistic information.

4. Methodology

Qwen-VL is a powerful multimodal large model released by the Tongyi Qianwen team, equipped with capabilities for joint image-text understanding and generation. The model typically employs advanced vision encoders (such as Vision Transformer) for image representation, which are integrated with the Qwen series of large language models that possess strong natural language understanding abilities. By incorporating cross-modal alignment mechanisms and multi-turn dialogue capabilities, Qwen-VL is able to deeply analyze various types of visual information—including image details, text, and charts—and

respond to natural language prompts with high-quality performance in visual question answering, image-text reasoning, and language generation tasks.

In this study, we adopt the Qwen-VL model to enhance cross-modal reasoning capabilities. Each sample in the EXAMS-V dataset (an image and its corresponding multiple-choice question) is converted into a structured prompt following the ChatML format, with a unified template designed to improve consistency and stability in model behavior. As illustrated in Figure 1, the prompt design follows a structured workflow consisting of three key steps: first, clearly identify the user's question; second, systematically analyze the image content (including objects, text, and background elements); and finally, provide a concise answer based on the reasoning process. Additionally, a system prompt is used to define the task goals and behavioral constraints, ensuring that the model operates in the role of an AI assistant and adheres to a standardized reasoning procedure.

To comply with API upload limits, images are preprocessed using the Python Imaging Library (PIL), compressed to under 9MB, and encoded in base64 format. Inference is conducted via Alibaba Cloud's DashScope platform using an OpenAI-compatible API. The model returns plain-text responses, from which the system extracts answers using regular expressions targeting tags like A. In cases where the output is malformed or the API request fails, the system logs the error and defaults to a randomly selected answer as a fallback.

The system supports batch inference using PyTorch's Dataset and DataLoader interfaces, and incorporates a checkpoint resumption mechanism to skip previously processed samples. All intermediate results are streamed in real time to JSONL files, ensuring stability, robustness, and reproducibility in large-scale evaluation tasks. With this structured prompting strategy and system-level optimizations, Qwen-VL demonstrates enhanced accuracy and consistency in complex image-text reasoning tasks.

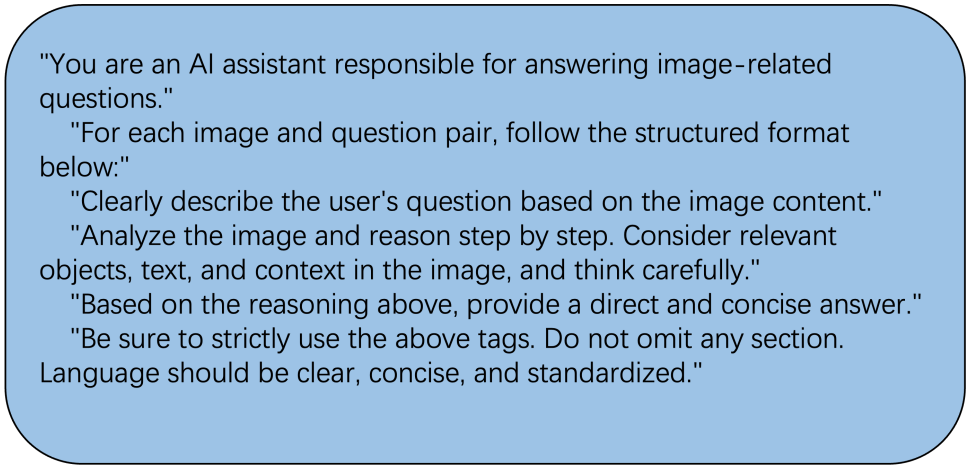

Figure 1: Illustration of Qwen-VL's Prompting Workflow for Image-Based Question Answering

5. Experimental Settings

This study constructs a robust inference system centered on Qwen-VL-Plus, integrating image compression, structured prompt design, batch processing, and checkpoint resumption mechanisms to enable large-scale and stable evaluation on the EXAMS-V dataset. The system adopts a three-stage prompting strategy: the system prompt defines the task objectives and behavioral constraints, while the user prompt includes a base64-encoded image and multiple-choice question, guiding the model to perform step-by-step reasoning and generate the final answer. All prompts follow the ChatML format to ensure consistency and stability during API calls.

Images are preprocessed using a custom compression function to keep their size under 9MB, thereby improving upload success rates and computational efficiency. Inference outputs are returned in the format <answer>…</answer>, from which the system extracts answers using regular expressions.

In cases of invalid responses, a fallback strategy is employed by selecting a random answer. All predictions are streamed in real time to JSONL files, and completed samples are automatically skipped via checkpoint resumption, ensuring robustness and reusability in long-running evaluations. Batch processing is implemented using PyTorch's Dataset and DataLoader interfaces, supporting scalable and interruption-tolerant evaluations across large test sets.

6. Result

Table 1 shows that Qwen-2.5-VL significantly outperforms the official baseline model provided by the competition across all language subsets. Its average multilingual accuracy reaches 43.76%, a substantial improvement over the baseline's 27.01%, with particularly strong gains in Chinese (+21.13%), German (+18.6%), and English (+20.9%). Even in low-resource languages such as Urdu, Qwen-2.5-VL maintains a 5.58% advantage.

The performance improvements of Qwen-2.5-VL stem not only from its powerful multimodal modeling capabilities but also from several system-level optimizations: the image compression strategy effectively reduces API call failures; the checkpoint resumption mechanism enables scalable evaluation; the unified prompt template ensures consistent model inputs; and robust error-handling mechanisms safeguard data integrity and reproducibility under unstable conditions such as API rejections or network interruptions.

Table 1
Performance comparison across languages for Qwen-2.5-VL and baseline.

Model\Task	Multilingual	English	Chinese	German	Urdu
Qwen-2.5-VL	43.76	45.70	47.91	49.61	35.69
baseline	27.01	24.80	26.78	31.01	30.11

7. Conclusion

This paper presents an in-depth study on the application effectiveness of the Qwen-2.5-VL model in multilingual visual-language understanding tasks. We systematically evaluate its performance across multiple language subsets—including English, Chinese, German, and Urdu—using the EXAMS-V multiple-choice question dataset. The experimental results show that Qwen-2.5-VL significantly outperforms the baseline model across all tasks, achieving an average accuracy improvement of 16.75%. In particular, the German subset demonstrates a notable gain of nearly 18.6%, fully showcasing the model's powerful capabilities in multimodal modeling and cross-lingual transfer.

To ensure the validity and efficiency of the evaluation, we design and implement a highly robust evaluation pipeline that encompasses modules such as data organization, image compression, API request construction, exception handling, and checkpoint resumption. This pipeline not only enhances system stability and scalability but also effectively reduces resource consumption and the need for manual intervention, highlighting its strong engineering practicality.

Based on the comprehensive experimental results and systematic pipeline optimization, we draw the following conclusions: Qwen-2.5-VL exhibits excellent generalization capabilities in multilingual visual-language understanding tasks; high-quality prompt construction and image processing mechanisms have a significant impact on model performance; and at the deployment level, engineering optimizations can greatly improve the usability of large-scale multimodal models.

Future work will further explore the role of automatic prompt generation, domain knowledge injection, and multi-stage vision-language interaction strategies in enhancing reasoning capabilities. Moreover, the emergence of more high-quality multilingual multimodal datasets is expected to provide broader opportunities for research on large models. As multimodal large models increasingly penetrate critical application domains such as education, healthcare, and public services, future research will

also focus on improving model interpretability, safety, and fairness. Promising directions include the construction of controllable multimodal reasoning frameworks with causal inference abilities, the development of low-resource fine-tuning techniques adaptable to minority languages and edge cases, and the design of transparent and debuggable visual-language alignment mechanisms. We further anticipate advancements in open-domain multilingual evaluation benchmarks, multi-task collaborative learning frameworks, and cross-modal knowledge representation methods, which will offer a solid foundation for the practical deployment of general-purpose multimodal intelligence.

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 ChatGPT in order to: Grammar and spelling check, Paraphrase and reword. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] OpenAI, Chatml documents, https://github.com/openai/openai-python/blob/main/chatml.md, 2024. Accessed: 2025-06-29.

[2] Anthropic, Claude 3.5 sonnet, https://www.anthropic.com/news/claude-3-5-sonnet, 2024. Accessed: 2025-06-29.

[3] G. Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, et al., Gemini: A family of highly capable multimodal models, arXiv preprint arXiv:2312.11805 (2023). arXiv:2312.11805.

[4] Z. Chen, J. Wu, W. Wang, W. Su, G. Chen, S. Xing, M. Zhong, Q. Zhang, X. Zhu, L. Lu, B. Li, P. Luo, T. Lu, Y. Qiao, J. Dai, Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks, arXiv preprint arXiv:2312.14238 (2023). arXiv:2312.14238.

[5] K. Li, Z. Meng, H. Lin, Z. Luo, Y. Tian, J. Ma, Z. Huang, T.-S. Chua, Screenspot-pro: Gui grounding for professional high-resolution computer use, https://likaixin2000.github.io/papers/ScreenSpot_Pro.pdf, 2025. Preprint, Accessed: 2025-06-29.

[6] B. Ionescu, H. Müller, D.-C. Stanciu, A.-G. Andrei, A. Radzhabov, Y. Prokopchuk, L.-D. Ştefan, M.-G. Constantin, M. Dogariu, V. Kovalev, H. Damm, J. Rückert, D. B. Abacha, A. G. S. 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), Lecture Notes in Computer Science, Springer, Madrid, Spain, 2025.

[7] D. Dimitrov, M. S. Hee, Z. Xie, R. J. 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.

[8] 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, arXiv preprint arXiv:2403.10378 (2024).

[9] B. Zhang, R. Sennrich, Root mean square layer normalization, in: Advances in Neural Information Processing Systems (NeurIPS), 2019.
[10] Y. N. Dauphin, A. Fan, M. Auli, D. Grangier, Language modeling with gated convolutional networks, in: Proceedings of the 34th International Conference on Machine Learning (ICML), volume 70, PMLR, 2017, pp. 933–941.