Multi-Prompt Ensemble Reasoning for MultimodalReasoning

Notebook for the ImageCLEF Lab at CLEF 2025

Jianzhong Yan, Leilei Kong*, Qida Wu and Junyi Li

Foshan University, Foshan, China

Abstract

MultimodalReasoning is a task focused on multilingual visual question answering (VQA). Given a question image with 3-5 possible answers, the model is required to identify the only correct answer contained in the image.We proposed a multi-prompt Ensemble Reasoning for Multilingual MultimodalReasoning task(MPER).This method designs multiple prompts, calls the GPT-4.1 model(OpenAI's visual-language model available via API as of April 2025) interface in parallel to obtain multiple answers, and then ensembles them to get the final answer.We achieved an accuracy of 0.5994 on the test set without supervised fine-tuning and zero-shot learning, ranking second in the Multilingual list, significantly better than the official baseline of 0.2701.

Keywords

MultimodalReasoning, Zero-Shot Learning, Prompt Engineering, Ensembling, VQA

1. Introduction

Visual Question Answering (VQA), as a core task for evaluating visual-language cross-modal understanding capabilities, requires the model to have comprehensive capabilities in image semantic parsing, natural language question understanding, and multimodal information fusion reasoning.Unlike traditional image description generation tasks, VQA systems need to complete progressive analysis from visual feature recognition to high-level logical reasoning[1].

Although a single carefully designed prompt can effectively guide LLMs to perform specific tasks, it reveals several limitations in complex VQA tasks:

- `Limited Coverage` : A single prompt may fail to activate the full breadth of knowledge or diverse reasoning paths within large models, struggling to encompass the inherent complexity of VQA tasks.
- `Vulnerability to Error` : Similar to the principles of ensemble learning, leveraging outputs from multiple independent prompts capitalizes on model diversity. This fusion mitigates risks stemming from flaws in a single prompt's design or inherent model randomness, enhancing overall robustness.
- `Narrowed Focus & Style` : Complex VQA reasoning often benefits from multiple viewpoints. A single prompt might constrain the model to a specific task aspect or answer style. In contrast, diverse prompt designs (e.g., varying question phrasing, context, or constraints) encourage the model to approach the problem from different angles, potentially yielding complementary insights or answer components[2, 3].

To overcome the limitations of single-prompt approaches described above, we proposed a multi-prompt Ensemble Reasoning for Multilingual MultimodalReasoning task(MPER). The core idea of this

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

*Corresponding author.

yanmoge520@gmail.com (J. Yan); kongleilei1979@gmail.com (L. Kong); 1362990744wuqida@gmail.com (Q. Wu); m13609768252@163.com (J. Li)

0009-0002-9047-4690 (J. Yan); 0000-0002-4636-3507 (L. Kong); 0009-0002-9532-937X (Q. Wu); 0009-0000-7584-8231 (J. Li)

method lies in systematically constructing and fusing multiple diversely designed prompts to holistically stimulate the intrinsic reasoning capabilities of large models from diverse facets and perspectives. This approach aims to maximize task-specific performance by exploring the combinatorial prompt space.

Breakthroughs in vision-language pre-training technology provide a new paradigm for addressing the above challenges. The self-supervised learning framework based on large-scale multimodal data significantly improves the cross-modal alignment capability of the model: the CLIP model proposed by Radford et al.[4] achieves semantic mapping of open-domain visual concepts through comparative learning of 400 million network image-text pairs; the BLIP framework developed by Li et al.[5] innovatively introduces a noisy data filtering mechanism, and through the collaborative optimization of synthetic description generation and quality discrimination modules, it sets the best performance record at the time in multiple tasks reported in ECCV 2022. It is worth noting that the emergence of general multimodal Transformer models such as GPT-4 indicates that visual reasoning capabilities are gradually being integrated into the general artificial intelligence system. The human-like performance of the model in professional tests verifies the gain effect of parameter scale expansion on complex multimodal reasoning tasks. Under this trend, the Qwen-VL [6] series of models achieves the synergistic improvement of fine-grained visual understanding and multi-round dialogue capabilities in zero-sample scenarios through a multi-stage progressive training strategy.

ImageCLEF-Multimodal Reasoning is a new task first proposed by CLEF 2025[7, 8], which aims to evaluate the generalization ability of VLM in multilingual visual question answering reasoning. Its challenges include language diversity, complex reasoning chains, and the integration of real-world knowledge. We propose a multilingual multimodal reasoning method that integrates multiple prompts. The core of this method is to combine the predictions of multiple prompt-based results to create a stronger predictive model.

2. Methods

This paper proposes a zero-shot multi-language reasoning framework based on multi-prompt integration and ensembling mechanism for the ImageCLEF 2025 multimodal reasoning task. The overall framework is shown in Figure 1.

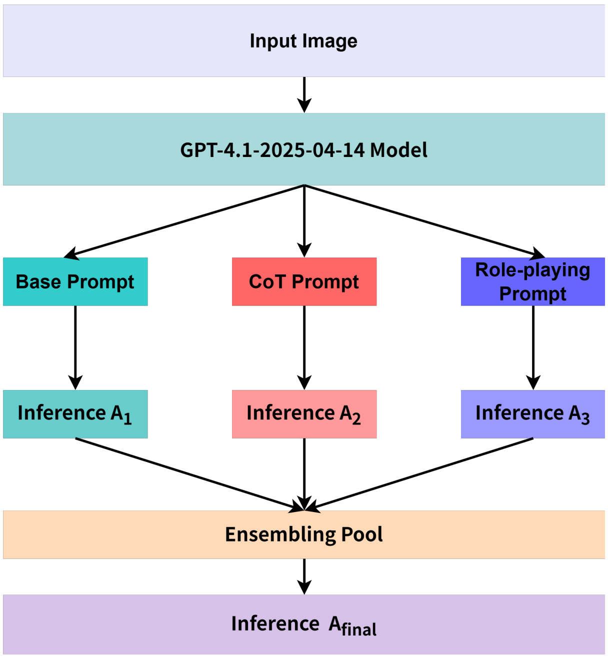

Figure 1: Overall framework

2.1. Prompt Template Construction

To effectively stimulate diverse reasoning pathways within the VLM, this study carefully designed three different types of prompt templates, the Base Prompt, the Chain of Thought (CoT[9]) Prompt, and the Role-playing Prompt,to serve a specific cognitive function. This multifaceted approach is rooted in high-level prompt engineering principles and aims to maximize task-specific performance by exploring the combinatorial prompt space[10].

Base Prompt:

- `Function` : Provides basic direct instructions that present the question and answer options in a straightforward manner. These prompts establish a baseline for the model to understand and respond to the query directly.
- `Intended Strength & Scenario` : Designed to be most effective for instances where the image information is unambiguous and the semantic relationship between visual content and question is straightforward. It primarily relies on the model's pre-trained knowledge base for efficient and direct answering.

Base Prompt

Analyze an image containing a multiple-choice question.
Identifythe question text,all provided answer options (including cases with more than four choices), and any visuals such as graph sortables.
Determine the correct answer using only the image content.
Respond exclusively with the letter of the correct option,without explanation.

Chain of Thought (CoT) Prompt:

- `Function` : Explicitly instructs the model to deduce the answer step-by-step (e.g., "think step by step"). This leverages the CoT technique, which breaks down complex multi-step problems into intermediate logical steps, enhancing reasoning transparency.
- `Intended Strength & Scenario` : Particularly suited for images involving complex visual-linguistic relationships or requiring multi-stage logical deduction. Prior research (e.g., Kojima et al., 2022[11]; Wei et al., 2022[12]) has demonstrated CoT's efficacy in significantly enhancing reasoning capabilities in large models by encouraging explicit, sequential reasoning.

CoT Prompt

You are a visual language model (VLM). Follow these exact instructions for analyzing each image with multiple-choice questions:
1. Identify and transcribe all text accurately (preserve original languages).
2. Clearly identify the question (stem) and list all answer choices in visual order (top to bottom, left to right).
3. Regardless of how options are originally labeled (e.g., numerals, non-Latin letters, or no labels), always relabel them sequentially from A to E according to their visual appearance order (1st→A, 2nd→B, 3rd→C, etc.).
4. Analyze and think any visual elements (graphs, charts, diagrams, images) step by step to determine the correct answer.
5. Output ONLY a single uppercase letter (A-E). Do NOT output original option labels, explanations, spaces, punctuation, or extra characters.
6. If insufficient information makes you uncertain, still choose and output the most likely option from A-E.

Role-playing Prompt:

- `Function` : Frames the task by instructing the model to adopt a specific persona or expertise (e.g., "As a master of science exams, please answer..."). This leverages the VLM's ability to assume roles and make role-consistent inferences.
- `Intended Strength & Scenario` : Hypothesized to be beneficial in scenarios involving linguistic ambiguity, requiring cross-cultural understanding within multilingual settings, or needing domain-specific knowledge application. By activating different knowledge subsets or reasoning biases through specific personas, this approach promotes "perspective diversity".

Role-playing Prompt

As a master of science exams (biology/chemistry/physics/mathematics and other science subjects) with global difficulty levels,you are taking a multi-language picture-text multiple-choice exam.
Please answer carefully according to the question image:
Task Instructions
1.Please identify the question stem and all the options in the image, and mark them in visual order as A, B, C, D, E;
2.Please make inferences based on the image information, charts, and formulas, and choose the most likely correct answer;
3.When outputting, only write the capital letter answer you choose (such as A), without outputting any explanation.
Please answer:

Design Rationale & Synergy: These three distinct prompt types are deliberately designed to stimulate the model's reasoning process from complementary perspectives. The Base Prompt offers efficiency on clear-cut cases, the CoT Prompt tackles complex multi-step reasoning, and the Role-playing Prompt addresses ambiguity and leverages contextual/cultural grounding. Their synergistic interaction within the ensemble framework is posited to enhance the overall robustness and accuracy of the final prediction by covering a broader spectrum of reasoning challenges inherent in the VQA task.

All templates support the 13 languages covered by the task to ensure multi-language compatibility.

2.2. Result Ensembling

After the candidate answers are summarized, the final answer A_{final} is determined by the majority ensembling strategy:

$$A_{\text{final}} = \arg \max_{A \in \{A,B,C\}} \sum_{i=1}^{3} \mathbb{I}(A_i = A)$$

When there is a tie (e.g. 1:1:1), the answer with a high confidence mark (e.g. "The correct answer is [A]") is preferred. This heuristic rule uses model self-consistency to resolve ambiguity.

3. Experiments and Evaluation

3.1. Dataset and Evaluation Protocol

The proposed framework is rigorously evaluated on the ImageCLEF 2025 multimodal reasoning task using the Exams-V dataset [13]. The dataset contains 4797 validation set samples and 3565 test set samples, covering 13 languages and involving multidisciplinary knowledge.

To maintain data authenticity and ensure fair evaluation of zero-shot capabilities, no preprocessing or augmentation was performed on the input images. For GPT-4.1 inference, images were converted to Base64 encoded strings and embedded with prompt words according to the OpenAI API specification.

Accuracy, defined as the ratio of correctly predicted samples to the total sample size, is used as the core evaluation metric.

3.2. Parallel Inference

Multithreaded API calls are implemented based on Python's ThreadPoolExecutor, processing three sets of differentiated prompts in parallel for each image. When calling the model (GPT-4.1), we use the default parameters and do not set parameters such as temperature and max_tokens (Note: temperature controls the randomness of the model output, and max_tokens limits the maximum number of tokens generated by the model) . This configuration limits the length of the response while ensuring output determinism. Each set of prompts generates a candidate answer A_i , forming a set $\{A_1, A_2, A_3\}$.

3.3. Comparative Performance Analysis

Experiments on the validation set clearly demonstrate the effectiveness of the multi-prompt ensemble strategy.

Initially, zero-shot reasoning capabilities were compared under a single-cue demonstration model, with GPT-4.1 achieving 48% accuracy, significantly outperforming Qwen-VL-Plus (15%). This establishes GPT-4.1 as a stronger base model.

After integrating three differentiated cue templates and adopting a majority ensembling mechanism, the GPT-4.1 model's accuracy on the validation set is significantly improved to 61%, an absolute improvement of 13% over the single-cue approach. Similarly, Qwen-VL-Plus also significantly improved to 43%, an absolute improvement of 18%.

These results empirically validate the core hypothesis that combining a diversified cue strategy with an ensemble approach can significantly improve the reasoning performance of VLMs, even in the zero-shot setting.

Table 1

On the validation set, the prediction results of the Qwen-VL-Plus and GPT-4.1 models obtained using the single prompt and multi-prompt ensemble methods respectively.Note: The "Single prompt" baseline results reported in this table correspond to the performance achieved using the CoT Prompt template exclusively.

Method	Qwen-VL-Plus	GPT-4.1
`Single prompt`	0.15	0.48
`MPER(Our)`	0.43	0.61

The table 1 provides direct quantitative evidence of the performance gains achieved by the multi-prompt ensemble strategy. It clearly demonstrates the causal impact of the ensemble approach by showing the significant accuracy gains from single prompt to multi-prompt plus ensembling for GPT-4.1 and Qwen-VL-Plus. This table is a cornerstone for verifying academic claims such as enhanced self-consistency and perspective diversity.

3.4. Overall Test Set performance

The positive improvement trend observed on the validation set is maintained on the unseen test set. Our framework achieves an accuracy of 0.5994, relying entirely on GPT-4.1 + multi-prompt ensemble solution.

This performance ranks second overall in the ImageCLEF 2025 multimodal reasoning evaluation, with an absolute improvement of 32.93% over the official baseline of 0.2701. Although there is still a gap with the championship team's accuracy of 81.40%, our method shows significant competitive advantages and robust generalization ability.

The table 2 directly demonstrates the competitive advantage of the proposed method over the official baseline on the final test set. It quantifies the overall impact of the framework and provides strong evidence for its practicality and competitive position in real-world benchmarks.

Table 2
At the final submission stage, the results of the baseline and this study's multi-cue integrated multilingual multimodal reasoning method on the test set.

Method	Multilingual
`Baseline`	0.2701
`MPER(Our)`	0.5994

3.5. Multilingual Generalization Analysis

To further evaluate the robustness of the framework with respect to language diversity, this study presents detailed results for 13 languages in the ImageCLEF 2025 task.

As shown in Table 3, the performance varies across languages, but generally demonstrates strong multilingual generalization, with accuracies ranging from 0.3941 (Urdu) to 0.7750 (Bulgarian). This highlights the framework's ability to handle challenges such as language diversity, complex reasoning chains, and integration of real-world knowledge, which are inherent challenges in the ImageCLEF task.

Table 3
Results achieved by our multi-prompt integrated multilingual multimodal reasoning approach in rankings across languages at the final submission stage.

Language	**MPER(Our)**	**Baseline**
English	**0.5938**	0.2480
Bulgarian	**0.7750**	0.2450
Chinese	**0.5283**	0.2678
German	**0.7403**	0.3101
Arabic	**0.4324**	0.2703
Italian	**0.6010**	0.2414
Spanish	**0.6696**	0.3156
Urdu	**0.3941**	0.3011
Serbian	**0.5468**	0.2365
Hungarian	**0.6518**	0.2348
Croatian	**0.5764**	0.2709
Polish	**0.7181**	0.2934
Kazakh	**0.5350**	0.2738

3.6. Discussion of Experimental Findings

Empirical results clearly confirm the significant performance gains achieved by the multi-prompt ensemble strategy. The significant accuracy improvements observed for GPT-4.1 (13% absolute improvement on validation set accuracy) and Qwen-VL-Plus (18% absolute improvement on validation set accuracy) highlight the value of this approach.

In-depth analysis shows that the multi-prompt ensemble strategy is particularly valuable for reasoning tasks in complex scenarios. When faced with questions containing semantic ambiguity or long text descriptions, different prompt templates effectively guide the model to focus on information features at different levels of abstraction. For example, in questions involving multi-entity spatial relations, the ensemble system stably outputs the correct answer through a ensembling mechanism, while the predictions of a single prompt show large fluctuations.

This improvement directly confirms Wang et al.'s [14] theory on enhancing self-consistency, which states that introducing diversity in reasoning paths can effectively reduce the probability of accidental errors. In addition, the findings of this study echo the "perspective diversity" theory recently proposed by the Trad team [3], which argues that different prompt templates are equivalent to building a multi-dimensional thinking entry point for the model, thereby activating the model's potential multimodal reasoning ability.

4. Limitations and Future Work

Despite the significant performance improvements achieved by the MPER framework, key limitations warrant acknowledgment. Primarily, the substantial cost associated with utilizing the commercial GPT-4.1 API constrained our ability to conduct thorough ablation studies. This limitation prevents a deeper quantitative analysis of the individual contribution of each prompt type (Base, CoT, Role-playing) and their various combinations to the overall ensemble performance. Additionally, the current prompt designs are primarily empirical, and performance variation persists across languages.

Future work will prioritize conducting comprehensive ablation experiments to rigorously quantify the impact of each prompt design choice and ensemble strategy component on model performance. This deeper analysis will provide crucial insights for further refining the approach. Further exploration of method optimization is also warranted.

5. Conclusion

This paper proposes a multi-language multimodal reasoning method based on multi-prompt integration of GPT-4.1 for the ImageCLEF 2025 MultimodalReasoning task. We compared the performance of the Qwen-VL-Plus and GPT-4.1 models under single prompt and multi-prompt integration, and finally selected the GPT-4.1 model and multi-prompt integration strategy to improve the accuracy. Experimental results show that this strategy has achieved significant gains on both the validation set and the test set, and finally won the second place in the competition. Our work shows that in complex multi-language reasoning tasks, the use of prompt engineering and integration methods can fully tap the capabilities of VLM. Future work can consider methods such as reinforcement learning, VLM knowledge fusion[15], knowledge graphs[16], and hybrid integration to further improve the multimodal reasoning performance of the model.

Acknowledgments

This work is supported by the Natural Science Platforms and Projects of Guangdong Province Ordinary Universities (Key Field Special Projects) (No. 2023ZDZX1023)

Declaration on Generative AI

During the preparation of this work, the author(s) used DeepSeek in order to: check and improve grammar, spelling, and language fluency. 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] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, D. Parikh, Vqa: Visual question answering, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 2425–2433.

[2] G. Marvin, N. Hellen, D. Jjingo, J. Nakatumba-Nabende, Prompt engineering in large language models, in: International conference on data intelligence and cognitive informatics, Springer, 2023, pp. 387–402.

[3] F. Trad, A. Chehab, To ensemble or not: Assessing majority voting strategies for phishing detection with large language models, in: ISPR, 2024. URL: https://api.semanticscholar.org/CorpusID:274436749.

[4] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, 2021. URL: https://api.semanticscholar.org/CorpusID:231591445.

[5] J. Li, D. Li, C. Xiong, S. C. H. Hoi, Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, in: International Conference on Machine Learning, 2022. URL: https://api.semanticscholar.org/CorpusID:246411402.

[6] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, J. Zhou, Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. URL: https://api.semanticscholar.org/CorpusID:261101015.

[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, 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.

[8] 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.

[9] Y. Wang, S. Wu, Y. Zhang, S. Yan, Z. Liu, J. Luo, H. Fei, Multimodal chain-of-thought reasoning: A comprehensive survey, ArXiv abs/2503.12605 (2025). URL: https://api.semanticscholar.org/CorpusID:277065932.

[10] W. Li, X. Wang, W. Li, B. Jin, A survey of automatic prompt engineering: An optimization perspective, ArXiv abs/2502.11560 (2025). URL: https://api.semanticscholar.org/CorpusID:276408554.

[11] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa, Large language models are zero-shot reasoners, ArXiv abs/2205.11916 (2022). URL: https://api.semanticscholar.org/CorpusID:249017743.

[12] J. Wei, X. Wang, D. Schuurmans, M. Bosma, E. H. Chi, F. Xia, Q. Le, D. Zhou, Chain of thought prompting elicits reasoning in large language models, ArXiv abs/2201.11903 (2022). URL: https://api.semanticscholar.org/CorpusID:246411621.

[13] R. Das, S. Hristov, H. Li, D. Dimitrov, I. Koychev, P. Nakov, EXAMS-V: A multi-discipline multilingual multimodal exam benchmark for evaluating vision language models, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 7768–7791. URL: https://aclanthology.org/2024.acl-long.420/. doi:10.18653/v1/2024.acl-long.420.

[14] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. H. Chi, D. Zhou, Self-consistency improves chain of thought reasoning in language models, ArXiv abs/2203.11171 (2022). URL: https://api.semanticscholar.org/CorpusID:247595263.

[15] F. Wan, X. Huang, D. Cai, X. Quan, W. Bi, S. Shi, Knowledge fusion of large language models, ArXiv abs/2401.10491 (2024). URL: https://api.semanticscholar.org/CorpusID:267061245.

[16] Y. Wang, M. Yasunaga, H. Ren, S. Wada, J. Leskovec, Vqa-gnn: Reasoning with multimodal knowledge via graph neural networks for visual question answering, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 21582–21592.