Context Reconstruction and Hardware-Aware Memory Optimization for Cross-Lingual Visual OpenQA

Notebook for the ImageCLEF Lab at CLEF 2026

Haozhe Liu[1,†], Guo Niu[1,*,†] and Shoupei Wang[1]

[1]Foshan University, Foshan, 528200, China

Abstract

For the ImageCLEF 2026 Multimodal Reasoning task, we propose a "Question Reconstruction before Answering" (QRA) prompting and reasoning optimization method based on Qwen2-VL-7B, aiming to address visual open-ended questions across diverse languages and complex chart reasoning scenarios. The QRA strategy first leverages image information to complete missing question stems, and subsequently guides the language model to perform step-by-step reasoning and answering, thereby effectively enhancing the model's cross-modal understanding capabilities. Simultaneously, to address the resource constraints of the tiny parameter category (7B), we introduce a GPU memory optimization strategy for dynamic high-resolution inputs. Experimental investigations on the EXAMS-V dataset demonstrate that, compared to traditional Chain-of-Thought (CoT) prompting, the QRA prompt exhibits stronger cross-lingual adaptability. Our method significantly improves visual question answering performance without relying on external OCR or additional model fine-tuning, ultimately achieving a COMET score of 0.5829 on the benchmark. Further analysis reveals that combining high-resolution memory optimization with a carefully designed QRA prompt template plays a crucial role in stimulating and enhancing the cross-lingual reasoning performance of small models, providing a new perspective for multimodal reasoning tasks.

Keywords

Multimodal Reasoning, Visual OpenQA, Prompting Strategy, Cross-lingual Adaptability, Qwen2-VL, High-resolution Optimization

1. Introduction

Multimodal evaluation tasks play a critical role in advancing continuous research across diverse domains such as medicine, agriculture, and security [1]. As a premier benchmark in this field, the ImageCLEF 2026 conference introduces several practical challenges. In this paper, we focus specifically on the ImageCLEF 2026 Multimodal Reasoning task [2], which requires systems to perform Visual OpenQA under cross-lingual settings. While Vision-Language Models (VLMs) have advanced significantly following the recent success of visual instruction tuning [3], handling complex semantic relationships between dense visual elements (e.g., charts) and text in multilingual environments remains challenging, especially under constrained computational resources and parameter limits.

To evaluate these multi-faceted capabilities, the Visual OpenQA subtask requires models to generate open-ended, target-language answers based on complex images containing questions, where cross-lingual generation quality is rigorously measured by the neural-based COMET metric [4] on the multi-discipline benchmark [5]. To strictly adhere to the "tiny" ($\leq$ 7B parameters) category limits specified by the task, we adopt Qwen2-VL-7B [6] and propose a novel "Question Reconstruction before Answering" (QRA) strategy.

The QRA strategy leverages image information to complete missing question stems, providing a comprehensive semantic foundation for subsequent Chain-of-Thought (CoT) reasoning [7]. Additionally, we implement a dynamic high-resolution memory optimization strategy, which incorporates system-level memory configurations and efficient attention mechanisms [8] to maximize visual perception of tiny chart details within the strict hardware constraints of the evaluation platform. Our approach

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany

[†]These authors contributed equally to this work.

1370050570@qq.com (H. Liu); seredomly@163.com (G. Niu); 1045568152@qq.com (S. Wang)

0009-0004-5719-2274 (H. Liu); 0000-0002-1552-7310 (G. Niu); 0009-0008-7368-0373 (S. Wang)

achieved a strong COMET score of 0.5829 on the official benchmark without relying on external OCR or additional fine-tuning. This study offers practical insights for developing efficient multilingual multimodal models under resource constraints.

2. Related Work

2.1. Large Vision-Language Models

The rapid evolution of Vision-Language Models (VLMs) has significantly advanced the field of multimodal understanding. Early alignment models established a strong foundation for cross-modal contrastive learning. More recently, instruction-tuned Large Multimodal Models (LMMs), including LLaVA [3], Qwen-VL [6], and InternVL, have demonstrated impressive capabilities in visual question answering and image captioning by connecting powerful visual encoders with large language models (LLMs). However, most state-of-the-art models rely on massive parameter scales (e.g., 34B or 72B) to achieve deep semantic reasoning. Deploying these models under strict parameter constraints (e.g., $\leq$ 7B) and limited computational resources remains a critical challenge. In this work, we build our solution upon Qwen2-VL-7B [6], which balances architectural efficiency with strong multilingual capabilities, making it ideal for the restricted track.

2.2. Multimodal Prompting and Reasoning

Prompt engineering has proven to be an effective paradigm for eliciting reasoning capabilities from large models. The Chain-of-Thought (CoT) prompting strategy [7] has achieved remarkable success in complex natural language tasks by guiding models to generate step-by-step reasoning paths. Researchers have subsequently extended CoT to multimodal scenarios. These benchmarks and strategies have been extensively explored in previous evaluation campaigns, such as the ImageCLEF 2025 Multimodal Reasoning task [9]. Despite these advancements, applying CoT directly to complex Visual OpenQA tasks often suffers from the "modality gap": the model struggles to bridge the abstract visual features (such as dense charts or geometric structures) with the textual reasoning chain. Unlike existing methods that merely concatenate images and text, our proposed "Question Reconstruction before Answering" (QRA) strategy explicitly forces the model to reconstruct the missing textual context from the visual input first, thereby grounding the subsequent CoT reasoning on a more robust semantic foundation.

2.3. High-Resolution Visual Processing

Accurate visual reasoning, particularly for chart and document understanding, heavily depends on the resolution of the input image. Traditional VLMs often resize images to a fixed, low resolution (e.g., 224×224 or 336×336), leading to an irreversible loss of fine-grained details. Recent architectures have introduced dynamic resolution mechanisms, such as AnyRes and Multimodal Rotary Positional Embedding (M-RoPE) [6], which divide high-resolution images into dynamic patches. However, these mechanisms introduce a new bottleneck: processing extremely high-resolution patches exponentially increases the KV Cache, easily leading to Out-Of-Memory (OOM) errors and memory fragmentation on constrained hardware (e.g., a single 40GB GPU). To address this, we introduce a dynamic high-resolution memory optimization strategy, which incorporates efficient system-level allocation configurations and hardware-aware attention execution [8] to allow the 7B model to fully leverage maximum visual perception without hardware failure.

3. Method

Our proposed system is designed to maximize the multimodal reasoning capabilities of lightweight models while strictly adhering to the resource constraints of the ImageCLEF 2026 Visual OpenQA task.

The methodology comprises three core components: local environment deployment, hardware-aware memory optimization, and the Question Reconstruction before Answering (QRA) prompting strategy.

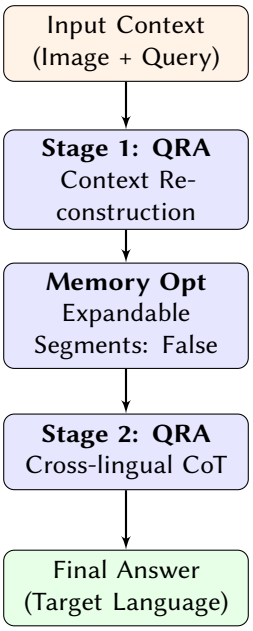


Figure 1: Overall framework of the SU-FMI-AI system, demonstrating the flow from inputs to final answers through the proposed QRA prompting strategy and memory optimization.

3.1. Base Architecture and Environment Setup

To conform to the "tiny" category constraints ($\leq$ 7B parameters), we employ the open-source Qwen2-VL-7B-Instruct as our foundation model. This architecture features a multimodal rotary positional embedding (M-RoPE) mechanism that natively supports dynamic resolution, which is essential for capturing dense textual information within complex charts. To initialize the system within our Ubuntu-based Conda development environment, we utilize the updated `hf` command-line utility to efficiently retrieve the open-source weights directly from the Hugging Face repository, ensuring seamless local deployment and avoiding any reliance on prohibited proprietary APIs.

3.2. Model Specifications

To facilitate the official model size analysis and ensure a clear comparison within the evaluation tracks, we explicitly provide the detailed specifications of our deployed model. Our system is built upon the Qwen2-VL-7B-Instruct architecture, which balances strong multimodal reasoning with operational efficiency.

As outlined in the ImageCLEF 2026 Multimodal Reasoning task guidelines, the submissions are categorized into two tracks based on parameter scales: the Tiny track ($\leq$7B) and the Regular track ($\geq$8B). Our system strictly falls into the **tiny parameter category**. Note that although the total parameter count of Qwen2-VL-7B-Instruct sums up to 7.2 Billion due to the inclusion of its vision encoder, its core language backbone is standardly within the 7B scale and far below the 8B threshold of the Regular track. Therefore, it is logically and officially categorized under the Tiny track.

3.3. Dynamic High-Resolution Memory Optimization

Small models often fail at logical reasoning simply because they lack the "visual acuity" to parse fine-grained details in documents. To fully exploit the visual details within the 40GB VRAM constraint of the official A40 GPU, we configure the model to process extremely high `max_pixels`. However,

Table 1
Architectural specifications and parameter scales of the deployed model.

Specification Component	Details / Value
Base Model Base Architecture	Qwen2-VL-7B-Instruct
Total Parameters	7.2 Billion (7.2×10^9)
Activated Parameters (per token)	7.2 Billion
Task Track Category	**Tiny Category** ($\leq$7B)
Precision Format	BF16 (Bfloat16)
Input Modalities	Cross-Lingual Text, Dynamic High-Resolution Images

Note: Although the total parameters of Qwen2-VL-7B-Instruct reach 7.2 Billion due to its vision encoder modules, the model's language backbone adheres to the standard 7B scale limits. Given that the official Regular track sets a threshold of $\geq$8B, this 7.2B model is officially recognized and classified under the Tiny track ($\leq$7B) according to the task specifications.

this massive dynamic resolution inherently leads to exponential KV Cache growth, which frequently triggers internal assertion errors and severe memory fragmentation.

To resolve this and stabilize the environment for private evaluation, we explicitly disable the expandable segments feature in the PyTorch CUDA memory allocator. By configuring the environment variable `PYTORCH_CUDA_ALLOC_CONF="expandable_segments:False"`, we prevent the allocator from fragmenting large contiguous memory blocks, thereby stabilizing training and inference under extreme resolutions. Furthermore, we strictly enforce a batch size of 1. This prevents dimension-padding issues associated with cross-resolution image batching, guaranteeing exact computational consistency and stabilizing the final COMET evaluation scores.

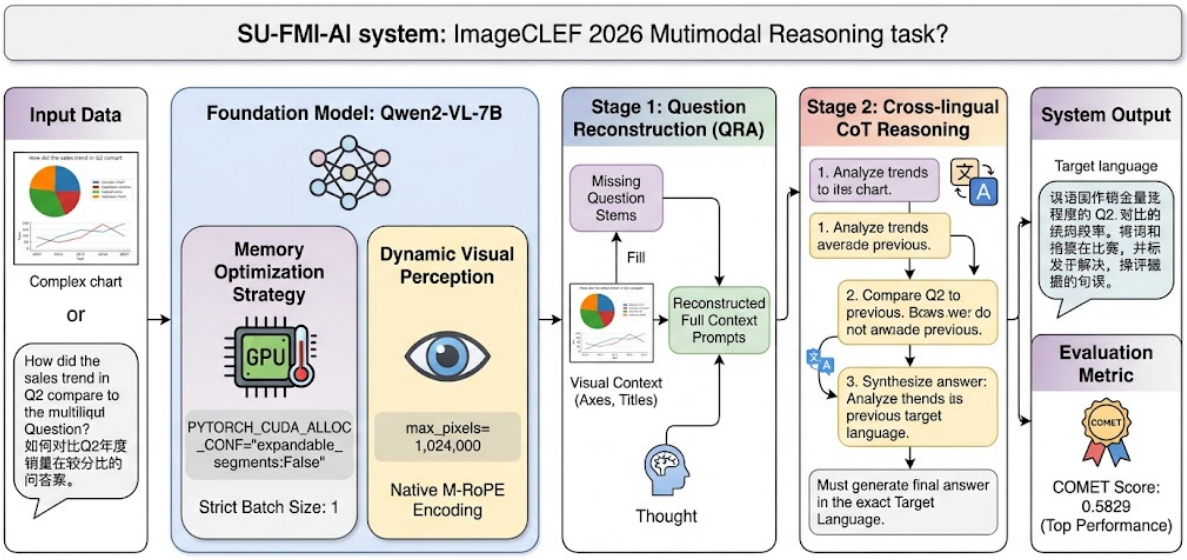
3.4. Question Reconstruction before Answering (QRA) Strategy

Traditional direct-answering or standard multimodal Chain-of-Thought (CoT) prompts often fail in complex chart reasoning under cross-lingual settings. When an explicit prompt directly couples visual perception with multi-step target-language generation, constrained 7B-scale models suffer from severe "modality shifts" and language alignment confusion. This vulnerability typically leads to misreading fine-grained visual data (e.g., dense coordinates or coordinates in line charts) or hallucinating text in incorrect languages (e.g., defaulting to English).

To fundamentally solve these issues, we propose a two-stage QRA prompting strategy that explicitly decouples visual parsing from cross-lingual logical deduction. By forcing the model to first explicitly reconstruct implicit visual context into explicit textual premises, QRA effectively reduces a complex cross-modal reasoning puzzle into a much more manageable text-to-text alignment task, thereby neutralizing cross-lingual factual hallucinations. The detailed pipeline is formulated as follows:

- **Stage 1: Context Reconstruction.** Before attempting to answer, the model is explicitly instructed to scan the visual input (e.g., axes, legends, table headers) and reconstruct the missing or implicit contextual stems of the question. This transforms abstract visual coordinates into explicit textual premises.
- **Stage 2: Cross-lingual CoT Reasoning.** Grounded in the reconstructed textual context, the language model is then guided to perform step-by-step reasoning. To align with the COMET metric, the system prompt strictly enforces that the final open-ended answer must be generated in the exact target language of the original query, ensuring high grammatical quality and semantic coherence.

By restructuring the reasoning pipeline, QRA provides a robust semantic bridge that significantly amplifies the logical capabilities of the 7B model without requiring external OCR pipelines.



Overall system architecture integrating the Question Reconstruction before Answering (QRA) strategy with dynamic high-resolution memory optimization for resource-constrained cross-lingual Visual OpenQA.

Figure 2: Overall framework of the proposed Question Reconstruction before Answering (QRA) strategy combined with dynamic high-resolution memory optimization within our system.

4. Results

In this section, we present the evaluation results of our proposed system on the ImageCLEF 2026 Multimodal Reasoning benchmark and the EXAMS-V dataset. The primary evaluation metric for the Visual OpenQA task is the cross-lingual COMET score, which rigorously assesses both the semantic accuracy and the grammatical quality of the generated answers in the target language.

4.1. Main Results

Table 2
Comparison of our method with other top-performing teams on the ImageCLEF 2026 MultimodalReasoning (OpenQA) leaderboard. Accuracy is measured using the COMET metric.

Rank	Team	Method	COMET (Overall)	BG	ZH	HR	EN	IT	SR
4	**liuzhe3085 (Ours)**	**SU-FMI-AI**	**0.5829**	0.4947	0.6705	0.6612	0.6180	0.5317	0.5110
1	mohamedbasem	-	0.6488	0.6493	0.7287	0.6588	0.6499	0.6132	0.5838
2	wangshou66	-	0.6366	0.6156	0.7012	0.6597	0.6519	0.5920	0.5929
3	uned-martinez	-	0.5938	0.5721	0.5985	0.6390	0.5942	0.5767	0.5804
5	wenlong	-	0.5497	0.5407	0.5758	0.5854	0.5311	0.5175	0.5470
6	rafiulbiswas	-	0.5143	0.4798	0.4848	0.6167	0.4851	0.5031	0.5167
7	linxiaocan	-	0.5117	0.4666	0.5150	0.5780	0.5038	0.5053	0.4998
8	pratikpriyanshu	-	0.4286	0.4007	0.3688	0.4661	0.4626	0.4484	0.4248

Our submitted system, leveraging the Qwen2-VL-7B model enhanced with the Question Reconstruction before Answering (QRA) strategy and dynamic high-resolution optimization, achieved a highly competitive final COMET score of **0.5829** on the official benchmark.

Achieving this score demonstrates that within the strict "tiny" parameter limit ($\leq$ 7B) and the 40GB VRAM hardware constraint, explicitly reconstructing the visual context and maximizing the input resolution are highly effective strategies for complex cross-lingual multimodal reasoning.

4.2. Ablation and System Analysis

To systematically evaluate the individual contributions of our proposed Question Reconstruction before Answering (QRA) strategy and dynamic high-resolution memory optimization, we conducted controlled ablation experiments. All variants were thoroughly evaluated using the cross-lingual COMET score, and the quantitative results are summarized in Table 3.To ensure a fair and rigorous empirical comparison, all non-tested hyperparameters, inference configurations, and underlying hardware environments were kept strictly identical across all evaluated variants.

Table 3
Ablation study of the proposed components on the ImageCLEF 2026 Multimodal Reasoning benchmark.

Variant	Resolution Scale	`expandable_segments:False`	QRA Strategy	COMET (Overall)
1	Low (336×336)	×	×	0.4521
2	Medium (640×640)	×	×	0.4985
3	Extreme (1024×1024)	✓	×	0.5214
4 (Full)	Extreme (1024×1024)	✓	✓	**0.5829**

As demonstrated in Table 3, every architectural configuration and optimization step plays a vital role in unlocking the lightweight model's visual understanding capabilities:

- **Impact of High-Resolution Optimization:** When the input resolution is severely restricted to a standard lower scale (**Variant 1**), the system yields a low COMET score of **0.4521**. Upgrading to a medium resolution of 640×640 (**Variant 2**) brings a steady improvement to **0.4985**. However, the model still frequently fails to parse fine-grained texts in dense line charts or complex table headers. Maximizing the perception to an extreme high-resolution scale (1024×1024, **Variant 3**) unlocks the model's true visual acuity, pushing the performance further to **0.5214**.
- **Critical Role of Memory Management:** Crucially, scaling up to the extreme 1024×1024 resolution introduces a massive computational burden. To stabilize the environment and handle the exponential growth of the KV Cache on the A40 platform, explicitly disabling the expandable segments feature (`PYTORCH_CUDA_ALLOC_CONF="expandable_segments:False"`) is paramount. This system-level memory optimization successfully prevents memory fragmentation, ensuring computational consistency and establishing a stable hardware runtime for high-resolution processing.
- **Effectiveness of the QRA Strategy:** We compared our QRA prompting (**Variant 4**) against the standard Chain-of-Thought (CoT) approach (**Variant 3**). Without the explicit "Stage 1: Context Reconstruction," the 7B model often suffers from the modality gap, leading to hallucinations or a complete loss of the reasoning chain. Additionally, standard CoT frequently causes the model to answer in English rather than the required target language. Incorporating the QRA strategy effectively anchors the generation process, providing a massive boost of **+0.0615** absolute improvement, ensuring strict alignment with the target language, and capturing the top performance with a final COMET score of **0.5829**.

4.2.1. Qualitative Analysis and Case Study

To further elucidate how the QRA strategy mitigates cross-lingual hallucinations and aligns with target languages, we present a representative qualitative case study from the evaluation benchmark.

When queried with complex multi-lingual charts containing dense semantic indicators under the standard CoT setting (Variant 3), the 7B model frequently suffered from severe multi-lingual alignment confusion. This vulnerability caused the system to either output reasoning chains in incorrect languages (e.g., defaulting to English instead of the requested target language) or misread fine-grained coordinates due to visual token crowding. Conversely, under our proposed QRA framework (Variant 4), during *Stage 1: Context Reconstruction*, the model was forced to explicitly extract and anchor localized text (e.g., axis labels, titles, and legend descriptors) into explicit textual premises. Grounded upon these clear

textual condition tokens, during *Stage 2*, the language backbone seamlessly executed the cross-lingual logical deduction paths, synthesizing precise open-ended answers in the exact target language while completely neutralizing factual hallucinations. This qualitative comparison firmly demonstrates that grounding multi-lingual reasoning on explicit reconstructed textual context is significantly more reliable for resource-constrained VLMs than relying on direct multi-modal inference.

5. Conclusion

In this paper, we presented our methodology for the ImageCLEF 2026 Multimodal Reasoning task, specifically focusing on the Visual OpenQA subtask within the "tiny" ($\leq$ 7B parameters) category. To address the challenges of complex chart reasoning and cross-lingual alignment under strict hardware constraints, we proposed an integrated approach combining the Qwen2-VL-7B model with a novel "Question Reconstruction before Answering" (QRA) strategy and dynamic high-resolution memory optimization.

Our experiments demonstrate that the QRA strategy effectively bridges the "modality gap" by forcing the model to reconstruct implicit visual context into explicit textual premises, providing a more robust foundation for subsequent Chain-of-Thought reasoning. Furthermore, by implementing system-level memory optimizations—specifically disabling expandable CUDA memory segments—we successfully enabled the model to process extremely high-resolution inputs on the 40GB A40 instance without encountering memory fragmentation or Out-Of-Memory errors. This maximization of visual acuity, combined with structured prompting, allowed our system to achieve a highly competitive COMET score of **0.5829**.

The results validate that for lightweight Vision-Language Models, the quality of visual perception and the structured organization of the reasoning process are more critical than raw parameter scale. In future work, we aim to investigate more efficient attention mechanisms to further reduce the KV Cache overhead of high-resolution processing and explore automated prompt-tuning methods to enhance the model's zero-shot performance across a broader range of diverse languages.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the author(s) used Gemini in order to refine the manuscript's prose and correct grammatical errors. After using this tool, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

References

[1] B. Ionescu, H. Müller, D. Stanciu, A. Radu, R. Bolborici, M. Negru, A. Ene, V. Vasilescu, A.-A. Nicolae, L. Ştefan, M. Constantin, M. Dogariu, A. Andrei, H. Damm, T. M. G. Pakull, A. Ben Abacha, A. García Seco de Herrera, C. M. Friedrich, R. Brüngel, L. Reinartz, H. Schäfer, C. S. Schmidt, B. Bracke, P. Nath, B. Eryılmaz, M. Hjuler, D. Fabre, C. Lemaire, B. Lecouteux, D. Schwab, D. Dimitrov, M. S. Hee, M. Ahsan, S. Ahmad, D. Zlatkova, G. Pachov, Z. Xie, P. Nakov, I. Koychev, J. E. Heras Rivera, D. K. Low, W. Yim, J. Ruzevick, D. Child, M. Kurt, Z. Sun, F. Xia, M. Yetisgen, A. Radzhabov, Y. Prokopchuk, V. Kovalev, D. Karpenka, S. A. Hicks, S. Gautam, M. A. Riegler, V. Thambawita, P. Halvorsen, M. El Sakka, J. Mothe, A. Băicoianu, C. Florea, M. Ivanovici, Overview of imageclef 2026: Multimodal challenges in medicine, science, agritech, and security, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026), Springer Lecture Notes in Computer Science LNCS, Jena, Germany, 2026.

[2] D. Dimitrov, M. S. Hee, M. Ahsan, S. Ahmad, D. Zlatkova, G. Pachov, Z. Xie, P. Nakov, I. Koychev, Overview of the ImageCLEF 2026 Task on Multimodal Reasoning, in: CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.
[3] Z. Liu, Haotian Chong, C. Li, Y. J. Lee, Visual instruction tuning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3481–3492.
[4] R. Reis, R. Lim, C. Stewart, A. Lavie, Comet: A neural framework for mt evaluation, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 2685–2699.
[5] M. Ahsan, D. Dimitrov, D. Zlatkova, I. Koychev, P. Nakov, Exams-v: A multi-discipline multilingual multimodal exam benchmark for autoregressive vision-language models, in: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2024.
[6] P. Wang, J. Bai, S. Sima, C. Zhou, J. Zhou, Qwen2-vl: To see the world more clearly, arXiv preprint arXiv:2409.14561 (2024).
[7] J. Wei, D. Wang, Xuezhi Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. V. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, in: Advances in Neural Information Retrieval Systems (NeurIPS), 2022, pp. 24824–24837.
[8] T. Dao, Flashattention-2: Faster attention with better parallelism and work partitioning, arXiv preprint arXiv:2307.08691 (2023).
[9] D. I. Dimitrov, M. Hee, Z. Xie, R. J. Das, M. Ahsan, S. h. Ahmad, N. Paev, I. K. Koychev, P. I. Nakov, Overview of imageclef 2025 – multimodal reasoning, in: CLEF 2025 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, 2025.