MSA at ImageCLEF 2025 Multimodal Reasoning: Multilingual Multimodal Reasoning With Ensemble Vision-Language Models*

Seif Ahmed[1,*,†], Mohamed Younes[1,†], Abdelrahman Moustafa[1,†], Abdulrahman Allam[1,†] and Hamza Moustafa[1,†]

[1]October University for Modern Sciences and Arts (MSA), Giza, Egypt

Abstract

We present a robust ensemble-based system for multilingual multimodal reasoning, designed for the ImageCLEF 2025 EXAMS-V challenge. Our approach integrates Gemini 2.5 Flash for visual description, Gemini 1.5 Pro for caption refinement and consistency checks, and Gemini 2.5 Pro as a reasoner which handles final answer selection, all coordinated through carefully engineered few-shot and zero-shot prompts. We conducted an extensive ablation study, training several large language models (Gemini 2.5 Flash, Phi-4, Gemma-3, Mistral) on an English dataset and its multilingual augmented version. Additionally, we evaluated Gemini 2.5 Flash in a zero-shot setting for comparison and found it to substantially outperform the trained models. Prompt design also proved critical: enforcing concise, language-normalized formats and prohibiting explanatory text boosted model accuracy on the English validation set from 55.9% to 61.7%. On the official leaderboard, our system (Team MSA) achieved first place overall in the multilingual track with 81.4% accuracy, and led 11 out of 13 individual language tracks, with top results such as 95.07% for Croatian and 92.12% for Italian. These findings highlight that lightweight OCR–VLM ensembles, when paired with precise prompt strategies and cross-lingual augmentation, can outperform heavier end-to-end models in high-stakes, multilingual educational settings.

Keywords

Multimodal Reasoning, Vision-Language Models, Large Language Models, Multilingual QA, ImageCLEF 2025, EXAMS-V 2025 Challenge

1. Introduction

Vision-Language Models (VLMs) have rapidly advanced in recent years, demonstrating remarkable capabilities in diverse multimodal tasks such as image captioning, visual question answering (VQA), and visual dialogue [1, 2]. Despite these successes, contemporary VLMs often encounter significant challenges in tasks demanding deep logical reasoning or inferencing [3, 4]. Limitations in the current generation of models are frequently revealed by complex queries involving intricate dependencies or hypothetical scenarios. Thus, it remains crucial to rigorously assess how well modern language and vision models can reason across complex multimodal inputs, especially in multilingual contexts [5, 6, 7]. For a detailed description of the shared task and competition, we refer the reader to the official overview papers [8, 9].

To address these challenges, three distinct tasks have emerged to evaluate VLM performance across various reasoning scenarios. Task 1, Visual Question Answering (VQA), requires the generation of accurate textual answers from image-question pairs, demanding precise interpretation and description of image content [3, 1]. Task 2, Visual Question Generation (VQG), involves generating contextually relevant questions from given images and associated answers, testing models' ability to deeply understand visual contexts and formulate meaningful textual queries [4]. Task 3, Visual Location Question Answering (VLQA), further extends these challenges by requiring spatial localization through

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

*Corresponding author.

†These authors contributed equally.

seifeldein.ahmed@msa.edu.eg (S. Ahmed); mohamed.tarek61@msa.edu.eg (M. Younes); abdelrahman.moustafa5@msa.edu.eg (A. Moustafa); abdulrahman.atif5@msa.edu.eg (A. Allam); hamza.moustafa@msa.edu.eg (H. Moustafa)

segmentation masks, necessitating accurate identification and delineation of regions of interest based on textual prompts. Our task was focused only on the Visual Question Answering (VQA) task.

Motivated by the complexities and novel demands of recent multimodal reasoning benchmarks [7, 5, 6], our approach leverages a strategic ensemble of advanced transformer-based models, specifically integrating Gemini 2.5 Flash for enhanced visual understanding and Gemini 1.5 Pro coupled with Gemini 2.5 Pro for sophisticated reasoning and answer aggregation. This hybrid approach exploits the complementary strengths of each model, achieving robust performance across multilingual datasets.

Our contributions in this paper are threefold: First, we provide a detailed examination of our system's architecture and the rationale behind model selection and combination. Second, we thoroughly analyze the performance of our system on multilingual multimodal reasoning tasks, emphasizing insights gained from multilingual diversity and complexity. Finally, we reflect on lessons learned from the evaluation and suggest pathways for future enhancements to strengthen multimodal reasoning capabilities.

2. Related Work

Recent advancements in multimodal and multilingual reasoning have underscored the complexity and richness of these domains. Benchmarks such as M4U [5, 10], M3Exam [6, 11], and PM4Bench [7] have emerged as pivotal platforms for evaluating large multimodal models across diverse languages and complex reasoning tasks. These benchmarks facilitate rigorous assessment of model capabilities in multilingual understanding, multimodal reasoning, and multi-level inference, encompassing various modalities such as text, images, and video.

The reasoning ability of language models, especially via chain-of-thought prompting, has also been extensively explored and shown to be particularly effective in multilingual contexts [12, 13]. This research emphasizes the necessity of developing robust models capable of handling multilingual data and highlights the benefits of incorporating explicit reasoning steps within model architectures. Recent large models like GPT-4 and Gemini have demonstrated significant progress in multilingual reasoning, maintaining logical coherence across diverse linguistic settings.

Multimodal reasoning tasks such as Visual Question Answering (VQA), Visual Question Generation (VQG), and Visual Location Question Answering (VLQA) have notably benefited from transformer-based architectures and vision-language model innovations [1, 4]. Techniques including Vision Transformers (ViT), SegFormer, and VisualBERT have shown promising results in interpreting visual information and generating relevant textual content. These transformer-based models leverage self-attention mechanisms to integrate visual and textual features, facilitating a nuanced understanding of multimodal inputs [5, 6].

Recent research also highlights the role of evaluation methodologies and metrics in accurately capturing model performance [3, 2]. Evaluations commonly include metrics such as accuracy, precision, recall, Intersection-over-Union (IoU), and Dice coefficients especially for tasks involving segmentation masks. The increasing complexity of multimodal tasks necessitates advanced evaluation strategies, as discussed in recent benchmarks, which systematically categorize challenges in visual question answering and generation, and underscore the importance of precise metrics to evaluate nuanced performances [8, 9].

Collectively, these studies underscore the ongoing need for sophisticated models capable of intricate multimodal reasoning, highlighting both the progress made and the challenges remaining in the field. Continued research and development are essential to addressing existing limitations and unlocking further advancements in multimodal and multilingual reasoning capabilities.

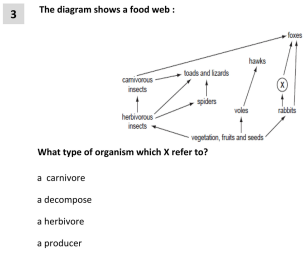
3. Dataset and Task Description

It is shown through Table 2, that the multilingual dataset under study consists of over 20,000 questions distributed across 13 languages including English, Chinese, German, Spanish, Arabic, Italian, Bulgarian, Croatian, Serbian, Urdu, Polish, and Kazakh. Each question is associated with metadata such as

sample_id, subject (e.g., biology, chemistry, physics), type (text or image_text), grade (ranging from 4 to 12), answer_key (A, B, C, D, or E), and language [as shown in Table 2]. The questions span a variety of educational domains and cognitive skills, presenting a comprehensive challenge for multimodal reasoning systems.

(a) Example of answer options entirely in **Arabic** although the metadata tag says "English".

(b) Example of answer options labeled in **Bulgarian letters** which the OCR fails to map to {A,B,C,D,E}.



(c) Example of answer options completely **unlabeled**.

Figure 1: Illustrative OCR-related challenges encountered in the dataset.

The dataset includes both multiple-choice questions and visual reasoning problems. However, several challenges were observed:

OCR-specific Challenges: Some items were printed in a language different from their metadata tag, while others lacked standard option labels (A–E) or used a different script problems that confused OCR and downstream prompts (see Figure. 1).

VLM-specific Challenges: Visual-language models often missed important details or made severe misinterpretations. Some image-based questions referenced diagrams that were missing entirely, leading to hallucinated or irrelevant answers. Table 1 shows that the gemini-2.5-flash model has misinterpreted the image saying that the vessel X is from the right ventricle.

Reasoner-specific Challenges: Large Language Models sometimes responded with full sentences or explanations instead of returning a concise choice like "A" or "D," which was required by the evaluation format.

The dataset statistics highlight the diversity of the challenge as shown in Table 2. For instance, Hungarian and Croatian had over 3,800 and 3,900 questions respectively, with a high proportion of visual questions. In contrast, English had fewer overall questions but maintained a balance between visual and textual modalities. This linguistic and subject-area diversity posed unique challenges for cross-lingual and multimodal generalization.

The task evaluated over this dataset is:

- **Task 1 – Visual Question Answering (VQA):** Assessing the ability to answer questions based on both images and accompanying text.

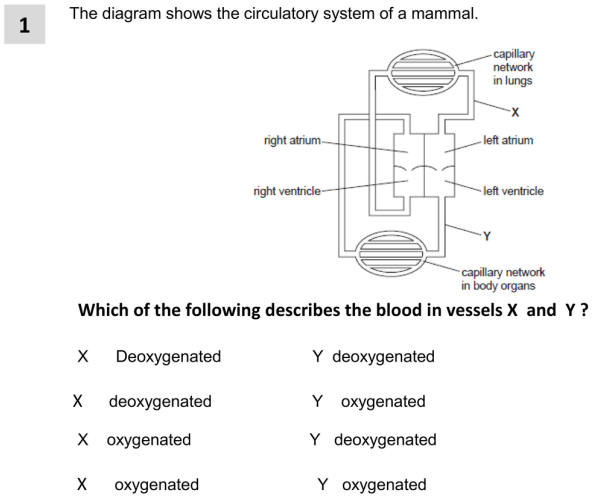
This task investigates different aspects of multimodal and multilingual reasoning and exposes the weaknesses and strengths of current VLM and LLM systems in handling such richly varied content.

4. Methodology

4.1. Overall Workflow

As shown in Figure 2, our system is a two–stage ensemble pipeline, inspired by recent advances in vision-language and large language models [1, 5, 7]. First, an **OCR–VLM stage** extracts rich textual

Table 1
VLM mis-interpretation vs. our ensemble fix.



Prompt	Extract the *Question* and *all answer options*, then provide a detailed, step-by-step description of every key visual element. Do not answer the question.
Gemini 2.5 Flash (VLM)	*Description (truncated):* Four-chamber heart; vessel X *from right ventricle to lungs*, vessel Y from left ventricle to body organs. *Predicted answer:* **B** (X deoxygenated, Y oxygenated) — *(incorrect)*.
Our Ensemble (VLM)	*Description (truncated):* Heart with labelled chambers; vessel X returns *oxygenated* blood from the lung capillary network to the left atrium, vessel Y carries *oxygenated* blood from the left ventricle to body organs *Predicted answer:* **D** (X oxygenated, Y oxygenated) — *(correct)*.
Ground Truth	*Answer:* **D**

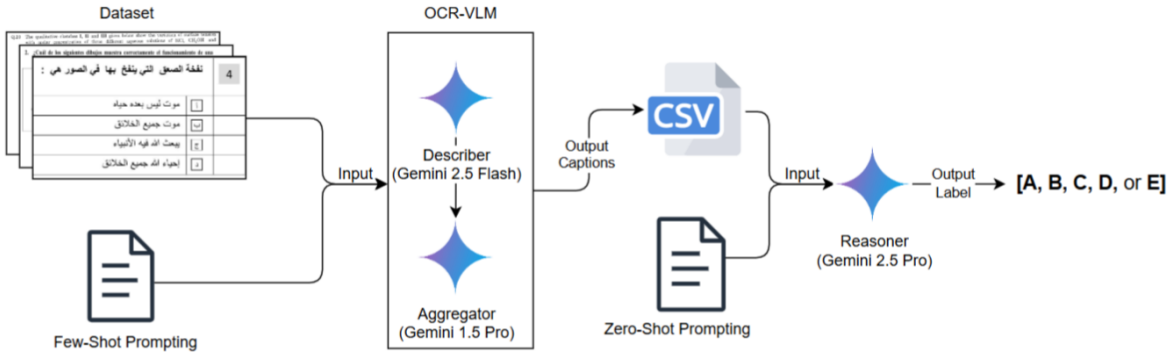


Figure 2: System pipeline: OCR–VLM ensemble (Gemini 2.5 Flash + Gemini 1.5 Pro) produces text for LLM answer selection (Gemini 2.5 Pro).

descriptions from each question image; second, a **Reasoner stage** maps the cleaned text to a final multiple–choice answer.

Table 2
Dataset statistics by language, showing number of subjects, questions, and visual/textual distribution. [14]

Language	ISO	Family	Grade	# Subjects	# Questions	# Visual Q. / Text Q.
English	en	Germanic	11, 12	4	724	181 / 543
Chinese	zh	Sino-Tibetan	8–12	6	2,635	1,991 / 644
French	fr	Romance	12	3	439	50 / 389
German	de	Germanic	12	5	819	144 / 675
Italian	it	Romance	12	6	1,645	292 / 1,353
Arabic	ar	Semitic	4–12	6	823	117 / 706
Polish	pl	Slavic	12	6	2,158	72 / 2,086
Hungarian	hu	Finno-Ugric	12	6	3,801	495 / 3,306
Bulgarian	bg	Slavic	4, 12	4	2,132	435 / 1,697
Croatian	hr	Slavic	12	6	3,969	700 / 3,269
Serbian	sr	Slavic	12	11	1,434	259 / 1,175

4.2. Stage 1: OCR–VLM Ensemble

Gemini 2.5 Flash (describer). We employ Gemini 2.5 Flash to generate a detailed natural-language caption of the input image. A *few–shot* prompt (1 example) is prepended to encourage the model to:

- Preserve mathematical symbols and subscripts,
- Normalise answer-option markers ("(A)", "A. ", "①", *etc.*),
- Output in the language inferred from document metadata.

Few-shot prompting and multilingual captioning have proven effective in recent VLM research [15, 12].

Gemini 1.5 Pro (aggregator). The caption is passed together with the original image to Gemini 1.5 Pro, which acts as a verifier. It is prompted to correct label mismatches, flag missing diagrams ("diagram above" errors), and translate stray text into the declared language.

4.3. Stage 2: Reasoner

Gemini 2.5 Pro receives the caption from each row in the CSV plus a *zero-shot* prompt, following best practices in multilingual reasoning evaluation [5, 6, 3]. We chose Gemini 2.5 Pro over Gemini 2.5 Flash for the final reasoning stage due to its superior performance in complex reasoning tasks and better adherence to strict output formatting requirements [16, 17]. While Flash excels in vision-language understanding, Pro demonstrates enhanced logical reasoning capabilities and more reliable response formatting, which are critical for the multiple-choice answer selection task. Gemini 2.5 Pro was selected due to its state-of-the-art performance in Global MMLU (Massive Multitask Language Understanding) with a score of 89.8%, making it a very reliable choice for this task [18].

Zero-Shot Reasoner Prompt

You are given a multiple-choice question extracted from an exam.
The question description is: {caption}

Perform the following analysis:
1. Carefully read and interpret the full question description provided in the caption.
2. Identify the main question being asked.
3. Extract all available answer options presented in the description.
4. Pay close attention to any data mentioned (tables, diagrams, charts, graphs, chemical structures, etc.).
5. Analyze all information in context.
6. Select the correct answer based solely on your analysis of the provided description.

Your final response MUST be ONLY the single letter of the correct answer option ["A", "B", "C", "D", or "E"] in English.
Absolutely NO other text, explanation, reasoning, or formatting is allowed in your response. Just the letter.

5. Experiments and Results

5.1. Experimental Setup

All submissions were evaluated on the public leaderboard for *MultimodalReasoning* [8]. Accuracy is computed as the fraction of questions for which the system returned the correct letter (A–E), following the competition's official evaluation protocol [9]. Our system runs the two–stage pipeline described in Section 4: Gemini 2.5 Flash → Gemini 1.5 Pro for OCR + VLM, followed by Gemini 2.5 Pro for reasoning and answer selection. Unless otherwise stated, ensemble inference uses `temperature=1.5 (2.5 Flash)`, `1.5 (1.5 Pro)`, and `0.2 (2.5 Pro)`.

5.2. Performance

To assess the effectiveness of our approach, we compared our system's accuracy against the organiser-supplied baseline across all supported languages. Table 3 summarises the official results, showing the substantial performance gains achieved by our ensemble pipeline. Notably, our system ranked first on the multilingual leaderboard and achieved top ranks in nearly all individual language tracks.

5.3. Ablation Study: Model Architecture and Prompt Engineering

We conducted a comprehensive ablation study to evaluate the impact of (1) model architecture and scale, (2) multilingual data augmentation, and (3) prompt engineering on multilingual multimodal reasoning performance.

Model architecture and multilingual data augmentation. The original English dataset consisted of 377 training and 347 validation questions. To enrich training data with cross-lingual reasoning patterns, we expanded this dataset to 6,841 training and 2,990 validation items by translating questions from 12 other languages into English using Gemini 1.5 Pro. We then fine-tuned three large language models—**Phi-4** (14B parameters), **Gemma-3** (12B parameters), and **Mistral** (7B parameters)—on both the original and expanded datasets. Additionally, Gemini 2.5 Flash was evaluated in a zero-shot setting via API to justify its selection as the vision-language component in our system.

Table 4 summarizes the results. The findings reveal that multilingual augmentation significantly improves performance for larger models: Phi-4 and Gemma-3 gained +19.63 and +19.96 percentage points, respectively. However, Mistral (7B) showed only minimal benefit (+0.74 pp), suggesting insufficient

Table 3
"Baseline" is the organizer-supplied reference system. Δ denotes the absolute accuracy gain.

Language	Baseline	MSA	Δ	Rank
Multilingual	27.01%	81.40%	+54.39%	1st
Arabic	27.03%	67.57%	+40.54%	1st
Chinese	26.78%	83.05%	+56.27%	1st
German	31.01%	89.15%	+58.14%	1st
Italian	24.14%	92.12%	+67.98%	1st
Spanish	31.56%	71.98%	+40.42%	1st
Urdu	30.11%	80.67%	+50.56%	1st
Serbian	23.65%	71.43%	+47.78%	1st
Croatian	27.09%	95.07%	+67.98%	1st
Polish	29.34%	82.24%	+52.90%	1st
Kazakh	27.38%	81.48%	+54.10%	1st
English	24.80%	86.52%	+61.72%	2nd
Bulgarian	24.50%	75.00%	+50.50%	3rd

capacity for complex cross-lingual reasoning. Gemini 2.5 Flash achieved a substantial gain of +12.79 pp, from 66.86% on the unexpanded dataset to 79.65% on the expanded dataset, outperforming all other models and validating its role in our system.

Table 4
Model ablation results on unexpanded and expanded datasets. Gemini 2.5 Flash was evaluated zero-shot via API (not fine-tuned).

Model	Parameters	Accuracy (%)	
	(B)	Unexpanded Dataset	Expanded Dataset
Gemini 2.5 Flash*	–	66.86	**79.65**
Phi-4	14	36.02	55.65
Gemma-3	12	23.92	43.88
Mistral	7	27.09	27.83

Prompt engineering. We further analyzed the role of prompt design by testing different prompting strategies on the English validation set. Switching from a verbose descriptive prompt to a strict "answer-letter-only" instruction boosted Gemini Flash accuracy from 55.9% to 57.1%. Replacing Flash with Gemini 1.5 Pro under the same prompt further increased accuracy to 61.7%, suggesting that larger models can exploit strict prompts more effectively (Table 5).

Table 5
Prompt-ablation results on the English validation split for the Reasoner stage.

Model	Prompt Style	Shots	Accuracy (%)
2.5 Flash	long descriptive	few	55.91
2.5 Flash	strict letter-only	few	57.06
1.5 Pro	strict letter-only	few	**61.67**

These results emphasize the importance of both architectural choices and precise prompt design in building effective multilingual multimodal reasoning systems.

5.4. Discussion

Our experiments highlight several key insights:

First, the ablation study demonstrates that both model scale and multilingual data augmentation are critical for achieving high reasoning accuracy. Larger models such as Phi-4 and Gemma-3 benefited substantially from training on the expanded dataset, whereas Mistral (7B) showed minimal improvement, indicating limited capacity for complex cross-lingual reasoning. Gemini 2.5 Flash, even without fine-tuning, consistently outperformed these models, underscoring the value of large-scale pretraining and advanced multimodal capabilities.

Second, prompt engineering played a pivotal role in optimizing performance. Strict output constraints, which prohibited explanatory text and enforced concise letter-only answers, reduced failure cases caused by "overflow" responses. Gemini 1.5 Pro exploited this prompt design more effectively than Gemini 2.5 Flash, suggesting a synergy between prompt quality and model capacity.

Finally, our findings reinforce the design choices of our ensemble system. By combining lightweight OCR–VLM components for vision-language understanding with a reasoning-optimized LLM, we achieved state-of-the-art performance in multilingual educational QA tasks.

6. Conclusion

In this paper, we presented a robust ensemble-based approach for multilingual multimodal reasoning, integrating Gemini 2.5 Flash and Gemini 1.5 Pro for vision-language tasks with Gemini 2.5 Pro as the final reasoner. Through careful prompt engineering and strict output normalization, our system achieved state-of-the-art performance on the ImageCLEF 2025 Multimodal Reasoning leaderboard, ranking first overall and securing the top position in 11 out of 13 language-specific tracks. The ablation study highlighted the importance of model architecture, multilingual data augmentation, and precise prompt design, demonstrating significant accuracy gains and validating the choice of Gemini 2.5 Flash as the backbone for our system, especially in handling languages with complex scripts.

Our findings underscore that combining lightweight, well-calibrated OCR–VLM pipelines with targeted prompt strategies can outperform heavier end-to-end models, particularly in high-stakes educational scenarios requiring reliable automatic grading. Nonetheless, challenges remain, especially regarding the handling of ambiguous diagrams and enforcing strict output formats in low-resource languages. Future work will explore reinforcement learning for format adherence, enhanced diagram processing, and further augmentation for underrepresented languages.

Overall, our results confirm that prompt-centric system design and ensemble modeling represent a powerful paradigm for advancing multilingual and multimodal question answering [8, 9, 5, 6, 3].

Declaration on GenAI use

During the preparation of this work, the author(s) used ChatGPT in order to: Drafting content, 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] J. Zhang, J. Huang, S. Jin, S. Lu, Vision-language models for vision tasks: A survey, IEEE Transactions on Pattern Analysis and Machine Intelligence (2024).

[2] Y. Wang, W. Chen, X. Han, X. Lin, H. Zhao, Y. Liu, B. Zhai, J. Yuan, Q. You, H. Yang, Exploring the reasoning abilities of multimodal large language models (mllms): A comprehensive survey on emerging trends in multimodal reasoning, arXiv preprint arXiv:2401.06805 (2024).

[3] J. Bi, S. Liang, X. Zhou, P. Liu, J. Guo, Y. Tang, L. Song, C. Huang, G. Sun, J. He, et al., Why reasoning matters? a survey of advancements in multimodal reasoning (v1), arXiv preprint arXiv:2504.03151 (2025).

[4] P. Lu, S. Mishra, T. Xia, L. Qiu, K.-W. Chang, S.-C. Zhu, O. Tafjord, P. Clark, A. Kalyan, Learn to explain: Multimodal reasoning via thought chains for science question answering, Advances in Neural Information Processing Systems 35 (2022) 2507–2521.

[5] H. Li, et al., M4u: Evaluating multilingual understanding and reasoning for large multimodal models, arXiv preprint arXiv:2405.15638 (2024).

[6] Y. Huang, et al., M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models, in: NeurIPS Datasets and Benchmarks Track, 2023.

[7] J. Gao, J. Song, J. Wu, R. Zhu, G. Shen, S. Wang, X. Wei, H. Yang, S. Zhang, W. Li, B. Wang, D. Lin, L. Wu, C. He, Pm4bench: A parallel multilingual multi-modal multi-task benchmark for large vision language model, 2025. URL: https://arxiv.org/abs/2503.18484. arXiv:2503.18484.

[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] 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. Garc 'ia 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.

[10] H. Wang, J. Xu, S. Xie, R. Wang, J. Li, Z. Xie, B. Zhang, C. Xiong, X. Chen, M4u: Evaluating multilingual understanding and reasoning for large multimodal models, 2025. URL: https://arxiv.org/abs/2405.15638. arXiv:2405.15638.

[11] W. Zhang, M. Aljunied, C. Gao, Y. K. Chia, L. Bing, M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models, in: A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, S. Levine (Eds.), Advances in Neural Information Processing Systems, volume 36, Curran Associates, Inc., 2023, pp. 5484–5505. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/117c5c8622b0d539f74f6d1fb082a2e9-Paper-Datasets_and_Benchmarks.pdf.

[12] F. Shi, M. Suzgun, M. Freitag, X. Wang, S. Srivats, S. Vosoughi, H. W. Chung, Y. Tay, S. Ruder, D. Zhou, D. Das, J. Wei, Language models are multilingual chain-of-thought reasoners, 2022. URL: https://arxiv.org/abs/2210.03057. arXiv:2210.03057.

[13] X. Zhou, et al., Language models are multilingual chain-of-thought reasoners, arXiv preprint arXiv:2210.03057 (2022).

[14] 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.

[15] K. Zhou, J. Yang, C. C. Loy, Z. Liu, Learning to prompt for vision-language models, International Journal of Computer Vision 130 (2022) 2337–2348.

[16] Google DeepMind, Gemini 2.5 pro vs flash: Performance comparison and model selection, https://deepmind.google/technologies/gemini/pro/, 2025. Accessed: 2025-03-15.

[17] Google AI, Gemini 2.5 pro: Benchmark results and technical specifications, https://blog.google/technology/ai/google-gemini-ai-update-december-2024/, 2025. Accessed: 2025-01-10.

[18] Google DeepMind, Gemini 2.5 pro: Our latest advances in reasoning, coding, and multimodal understanding, https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/, 2025. Accessed: 2025-05-28.