Overview of ImageCLEF 2025 – Multimodal Reasoning

Dimitar Dimitrov[1,*], Ming Shan Hee[2], Zhuohan Xie[2], Rocktim Jyoti Das[2], Momina Ahsan[2], Sarfraz Ahmad[2], Nikolay Paev[1], Ivan Koychev[1] and Preslav Nakov[2]

[1]*Faculty of Mathematics and Informatics, Sofia University "St. Kliment Ohridski", Bulgaria*
[2]*Mohamed bin Zayed University of Artificial Intelligence, UAE*

Abstract

We present an overview of the first edition of the ImageCLEF Multimodal Reasoning Lab at the 2025 iteration of the Conference and Labs of the Evaluation Forum (CLEF). The goal of the task is to evaluate how well vision-language models can reason over complex visual and textual examination material. The test dataset consists of 3,565 questions in 13 different languages. Participants received an image of a question, which included answer choices and metadata outlining the nature of the visual content within the image. Their objective was to choose one correct answer from a group of three to five options. The task had moderate participation with a total of 51 registered teams. Of these, 11 teams submitted results on the test set across all 13 languages and the multilingual leaderboard, with 129 graded submissions overall. The teams mainly used zero-shot approaches, while some chose few-shot methods or fine-tuning. Qwen-VL was the most commonly used model, followed by Gemini. Participants focused on prompt engineering, mostly using variations of instruction prompts that guided the models through processing steps to reach a final answer. Some teams approached the task from an optimization perspective, showing that well-optimized models can achieve competitive performance with fewer parameters and faster inference times. This task contributes to the broader effort of expanding resources for vision-language reasoning evaluation, particularly in low-resource languages. The dataset has been publicly released, along with the gold labels for the test set. We hope this resource will support future research on multilingual and multimodal understanding and foster the development of better and more efficient vision-language models.

1. Introduction

Understanding and reasoning over both images and text has long been recognised as a core challenge for artificial intelligence. Early datasets such as VQA [1] and CLEVR [2] revealed how people naturally combine language and vision when answering exam-style questions, interpreting charts, or solving textbook problems. Pioneering captioning models like Show-and-Tell [3] and Show, Attend and Tell [4] demonstrated the feasibility of bridging the two modalities end-to-end, while vision–language pre-training frameworks including ViLBERT [5], LXMERT [6] and UNITER [7] established the joint representations that underpin many modern systems. Scaling these ideas to the large language model regime has produced today's Multimodal Large Language Models (MLLMs), for example, CLIP [8], Flamingo [9], and PaLM-E [10]. Over the past year, such MLLMs have achieved strong results on visual question answering, open-ended captioning, multimodal dialogue, and even step-by-step math reasoning. Nonetheless, rigorous audits show that their compositional reasoning abilities remain limited, especially when complex textual cues must be tightly integrated with visual evidence [11, 12, 13].

Over the past decade, a diverse suite of text-only reasoning benchmarks has driven significant progress in the development of language models capable of more structured and transparent problem solving. These benchmarks span a range of reasoning paradigms, including numerical reasoning [14], multi-hop commonsense inference [15, 16], deductive logic [17, 18], mathematical reasoning [19, 20], spatial planning [21], and domain-specific tasks in finance [22, 23], collectively establishing increasingly

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain
*Corresponding author.
ilijanovd@fmi.uni-sofia.bg (D. Dimitrov); mingshan.hee@mbzuai.ac.ae (M. S. Hee); zhuohan.xie@mbzuai.ac.ae (Z. Xie); zrocktim.jyotidas@mbzuai.ac.ae (R. J. Das); momina.ahsan@mbzuai.ac.ae (M. Ahsan); sarfraz.ahmad@mbzuai.ac.ae (S. Ahmad); paev@uni-sofia.bg (N. Paev); koychev@fmi.uni-sofia.bg (I. Koychev); preslav.nakov@mbzuai.ac.ae (P. Nakov)
0000-0003-1308-180X (D. Dimitrov); 0000-0002-6328-5889 (M. S. Hee); 0009-0008-2650-2857 (Z. Xie); 0009-0004-5988-9673 (M. Ahsan); 0000-0003-1039-3787 (S. Ahmad); 0009-0006-6125-2684 (N. Paev); 0000-0003-3919-030X (I. Koychev); 0000-0002-3600-1510 (P. Nakov)

rigorous benchmarks for evaluating reasoning capabilities in large language models. However, language-only evaluation captures only part of human problem-solving. Tasks such as interpreting geometric proofs, analyzing circuit schematics, or optimizing supply chains frequently require not only linguistic understanding but also the ability to extract and reason over structured visual information. This realization has led to a parallel surge of multimodal reasoning benchmarks: MATH-V adds diagrams to word problems [12]; NP-HardEval4V provides graph instances for knapsack and shortest-path tasks to isolate algorithmic reasoning [11]; MMMU spans 30 academic subjects to expose gaps between perception and domain expertise [24]; MLLM-COMPBench stresses pairwise relational comparisons [13]; and ScienceQA mixes visuals with natural-language questions, albeit only in English [25]. Taken together, the field is progressing from text-centric reasoning to integrative multimodal evaluation; yet, current suites still underrepresent the linguistic and structural diversity of real educational assessments, highlighting ample room for broader, more authentic benchmarks.

Recent benchmarks have taken steps to evaluate the reasoning abilities of multimodal models more precisely. Specifically, MATH-V tests how models handle math questions that include diagrams [12], while NPHardEval4V focuses on algorithmic tasks such as shortest paths and knapsack problems. It separates reasoning from recognition and instruction-following to test models in isolation [11]. MMMU introduces a broad set of problems across academic disciplines and highlights the gap between visual understanding and subject-specific reasoning [24]. MLLM-COMPBench, on the other hand, evaluates how well models handle comparisons between image pairs across multiple types of relativity, such as emotion, spatiality, and quantity, but does not consider structured exam-style questions or integrate text alongside image comparisons [13]. ScienceQA includes some visual elements and subject variation, but it is limited to English and lacks consistent metadata or support for multilingual evaluation [25]. As a result, current benchmarks either focus too narrowly on specific visual tasks or lack the language and structural variety found in real educational assessments.

In this paper, we describe the ImageCLEF 2025 Multimodal Reasoning task under the ImageCLEF 2025 Lab [26]. The task is a shared benchmark designed to evaluate model performance on multiple-choice questions (MCQs) that may include both text and visual content. The dataset used is EXAMS-V [27], further enriched by a new test set. The combined corpus spans 14 languages and 44 academic and vocational subjects, covering science fields such as biology, physics, and chemistry, as well as disciplines like history, informatics, fine arts, and business. The languages represented include Arabic, Chinese, Urdu, Kazakh, and several European languages such as French, German, Bulgarian, and Polish. This wide linguistic coverage enables the benchmark to evaluate models in both high-resource and low-resource settings, making it well-suited for real-world multilingual education scenarios. Each example includes a question with three to five answer options with a single correct answer, and may also contain a visual element such as a graph, figure, table, or chemical diagram, depending on the question type as shown in Figure 1. The task reflects a realistic educational setting, especially for low-resource languages and disciplines, and supports model evaluation using a straightforward accuracy metric.

One of the primary objectives of this task is to evaluate systems in low-resource settings. This includes languages such as Arabic and Urdu, as well as scientific content that requires diagrams or tables to answer the question. Each sample in the dataset contains metadata, including subject, grade level, and the type of visual elements. This enables a deeper analysis of system performance and allows for comparisons of results across languages, grades, and question types.

2. Related Work

2.1. Multimodal Reasoning.

Multimodal reasoning has emerged as a critical research area at the intersection of vision and language, enabling models to integrate and interpret heterogeneous information sources. This capability is essential for real-world applications such as educational assessments [27, 28, 29, 30], online content moderation [31, 32, 33], and scientific analysis [25, 34], where information often includes structured tables, data charts, and annotated figures. In the education domain, students frequently engage with

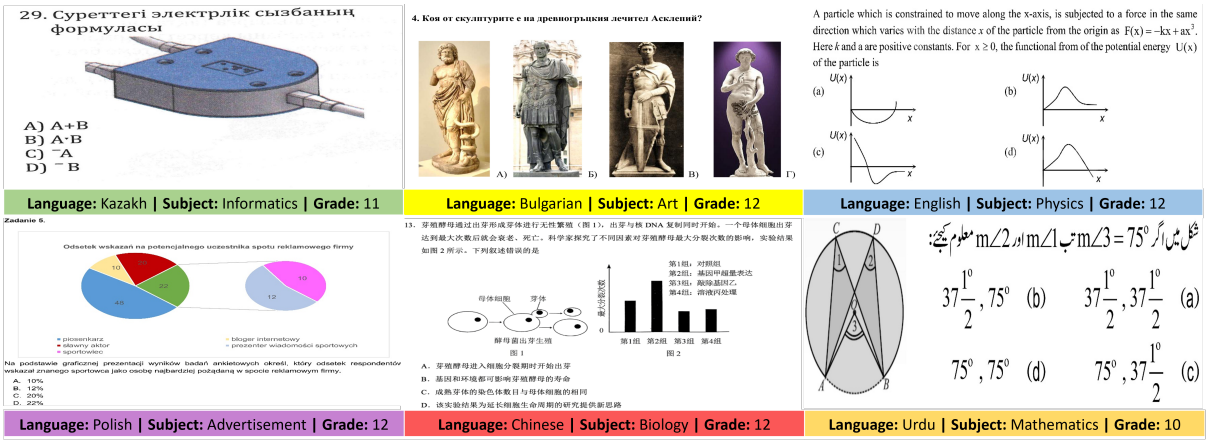

Figure 1: Sampled questions from the new ImageCLEF Multimodal Reasoning test set from different languages.

classroom materials that combine visual and textual information, ranging from annotated diagrams to chemical symbols and structures. As students increasingly use AI tools to tackle difficult questions, it becomes essential to evaluate multimodal models of structured educational content that reflect real-world curricula, ensuring that their responses are accurate and reliable.

Several evaluation benchmarks have been introduced across different languages and disciplines to assess the performance of models on examination questions. EXAMS-V [27] presents a multidisciplinary, multimodal, and multilingual benchmark consisting of 20,932 multiple-choice questions. These questions may include tables, figures, diagrams, maps, scientific symbols, and equations, and span more than 20 academic disciplines in 11 languages. Kaleidoscope [28] provides 20,911 exam questions in 18 languages and 14 subjects, ensuring linguistic and cultural authenticity through contributions from a diverse group of researchers around the world. More recently, MDK12-Bench [30] introduced more than 140,000 examples drawn from K-12 exams in six disciplines, enriched with difficulty labels and explanation rationales to support fine-grained reasoning evaluation. These benchmarks collectively emphasize the growing need to evaluate multimodal models through realistic, diverse, and linguistically rich educational scenarios. This *ImageCLEF 2025 - Multimodal Reasoning* task further contributes to this line of research by inviting participants to tackle real-world exam question challenges, with a test set comprising 3,565 questions to evaluate the models' capabilities.

2.2. Low-Resource Multilingual Languages.

Although current multimodal and language models have demonstrated remarkable capabilities in high-resource languages (i.e., languages with abundant data), many widely spoken languages in Southeast Asia, Central Asia, and Africa remain underrepresented due to limited data availability (i.e., low-resource languages). These low-resource languages pose a significant challenge for existing AI models, as they differ substantially in structure, grammar, and cultural context [35, 36]. Recent efforts have sought to bridge this gap by curating multilingual and culturally representative datasets. For example, KazMMLU [37] adapts the MMLU benchmark to Kazakh to assess general knowledge reasoning within Central Asian contexts. SGHateCheck [38] focuses on detecting hate language in Singaporean English, Malay, and Mandarin, with an emphasis on linguistic code-switching and localized harms. Similarly, IndoNLP [39] uses crowdsourcing to build benchmarks for several under-resourced Indonesian languages. Collectively, these datasets highlight the linguistic diversity and cultural specificity that are often overlooked in the mainstream benchmarks.

The *ImageCLEF 2025 Multimodal Reasoning Task* aims to address this gap by introducing a curated multimodal benchmark on K-12 examination questions that covers a diverse range of languages, scripts, and question types. The task releases a test set that extends the 11 languages in EXAMS-V with three new languages: Kazakh, Urdu, and Spanish, exam-style questions evaluating the robustness

and generalizability of multimodal models in multilingual contexts, thereby promoting research that inclusively supports a broader spectrum of linguistic communities.

3. Dataset

3.1. Data Collection

The newly developed test set contains questions from the languages present in the EXAMS-V dataset, as well as questions in three additional languages: Kazakh, Urdu, and Spanish. The questions are sourced from PDFs available online, mostly from annual school exams. For some of the languages in the original dataset, we used the same sources, but we took more recent versions of the yearly exams. For Urdu and Spanish, we considered PDFs available online, while for Kazakh, questions were scanned from textbooks. After the PDFs are collected, we run a processing step that converts the PDF pages into a series of images using the pdf2image[1] Python package.

3.2. Data Annotations

The question extraction and annotation follow the steps used to create the original EXAMS-V dataset. We use open-source software[2] to annotate the bounding boxes of the questions that align with the annotation guidelines of the original EXAMS-V dataset - only multiple-choice questions with 3 to 5 options and exactly one correct answer are considered. The bounding box of each question encompasses all possible answers, as well as the images, tables, and other necessary information provided. In some languages, examination formats included questions that overflowed into multiple pages, which we discarded to simplify our annotation process. After annotating the bounding boxes, an automatic script extracts each question as a separate image. The next step of the pipeline is the creation of a metadata file that contains a unique ID, the path to the cropped image, the label for the correct answer, and annotations for the type of visual elements included. The source PDFs provide the correct gold label, reducing the risk of errors in annotation; therefore, a single annotator per record is sufficient.

3.3. Data Stats

The dataset for the Multimodal Reasoning task is divided into three subsets: train, validation, and test. Table 1 reports the statistics for the training and validation portions of the EXAMS-V dataset [27]. This dataset comprises 20,932 questions across 20 subjects, covering grades 4 to 12 in 11 languages, providing a solid foundation for training and evaluation. The newly developed test set includes a total of 3,565 new questions from recent public high school examinations. Table 2 presents detailed statistics on the new data. The test set includes all languages from EXAMS-V, except for French, and introduces three additional languages: Urdu, Kazakh, and Spanish. As with EXAMS-V, the new data preserves the same diversity in languages and subjects, as well as question complexity, and includes 203 parallel questions in three languages: Croatian, Serbian, and Italian. In terms of visual representation, the test data provides a higher proportion of questions with visual features in most languages as seen in Tables 2 and 3. We also observe that Figures and Graphs are the most common visual features across all languages.

4. Evaluation Framework

4.1. Task organization

The task was conducted in two phases: an exploration phase, during which participants familiarized themselves with the publicly available training and validation data [40], followed by a test phase. In the

[1]https://pypi.org/project/pdf2image/
[2]https://opencv.org/

Table 1
Original Exams-V train and validation data statistics. Here, # visual Q. refers to questions with multimodal context and # text Q. refers to text only questions.

Language	ISO	Family	Grade	#Subj.	# Total Q.	# 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	11	1,645	292	1,353
Arabic	ar	Semitic	4-12	6	823	117	706
Polish	pl	Slavic	12	1	2,511	422	2,089
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	13	3,969	700	3,269
Serbian	sr	Slavic	12	11	1,434	259	1,175

Table 2
New test data statistics. Here, # **visual Q.** refers to questions with multimodal context, and # **text Q.** refers to text-only questions. **Urdu, Kazakh, and Spanish* are new languages, with no training/validation data from Exams-V.

Language	ISO	Family	Grade	#Subj.	# Total Q.	# visual Q.	# text Q.
English	en	Germanic	12	1	512	62	450
Chinese	zh	Sino-Tibetan	12	4	407	195	212
German	de	Germanic	12	6	258	67	191
Arabic	ar	Semitic	10-12	4	222	168	54
Polish	pl	Slavic	12	7	259	104	155
Hungarian	hu	Finno-Ugric	12	6	247	30	217
Bulgarian	bg	Slavic	12	6	200	68	132
Croatian	hr	Slavic	12	5	203	58	145
Serbian	sr	Slavic	12	5	203	58	145
Italian	it	Romance	12	5	203	58	145
Urdu*	ur	Indo-Aryan	9-10	5	269	12	257
Kazakh*	kk	Turkic	11	4	243	243	0
Spanish*	es	Romance	12	10	339	209	130

test phase, images of the questions from the test dataset were released along with metadata describing the visual components within the questions. Participants submitted results to 14 different leaderboards: one multilingual leaderboard and 13 individual leaderboards, one for each language. Participants were allowed to make multiple submissions during this phase, but no feedback was provided. Final rankings were determined based on each participant's last submission at the end of the test phase.

4.2. Evaluation Measure

We use accuracy as the evaluation metric for the task. Accuracy is calculated as the percentage of questions where the model's selected answer matches the correct option. Since each question has a single correct answer, accuracy provides a simple and reliable way to compare model performance across different languages and subject areas.

4.3. Baselines

For our baseline experiments, we used the Instruct variants of four models: SmolLM-Instruct, SmolVLM-Instruct, OLMO-Instruct, and MOLMO-Instruct. These models span a range of modalities and sizes, providing a strong starting point for building an initial understanding of the task.

Table 3
Vision feature distribution of the test set. Note that a single question can contain multiple types of visual elements.

Language	Table	Figure	Graph	Chem. Struct.	Total
Arabic	9	126	37	6	181
Bulgarian	4	55	2	7	68
Chinese	28	126	41	23	218
Croatian	5	40	12	2	61
English	0	47	16	0	63
German	2	50	15	0	67
Hungarian	0	22	8	0	30
Italian	5	40	12	2	61
Kazakh	4	47	192	0	243
Polish	23	78	3	0	104
Serbian	5	40	12	2	61
Spanish	65	84	44	35	228
Urdu	0	12	0	0	12

We grouped the models based on their modality. SmolLM-Instruct and OLMO-Instruct are text-only models, while SmolVLM-Instruct and MOLMO-Instruct are multimodal models capable of processing both images and text. All models were evaluated in a zero-shot setting. For the language-only models (SmolLM and OLMO), we provided image captions as input instead of raw images. For the vision-language models (SmolVLM and MOLMO), the original image was used as input.

Each model group was tested with two types of prompts, i.e., **Prompt 1**: A short, direct instruction asking the model to select the correct answer based on the input (caption or image). **Prompt 2:** A more detailed, step-by-step reasoning prompt encouraging the model to extract and analyze all relevant elements, such as multilingual content, tables, or diagrams, before selecting an answer. For example, Prompt 1 for VLMs instructed the model to analyze the image and reply with just the letter of the correct option, while Prompt 2 guided the model to extract the question, options, and visual cues before answering. A similar prompt pair was used for LLMs but applied to the caption text.

We used the same two prompts across all subjects and languages, and made no further tuning or task-specific engineering. This setup allowed us to evaluate how well these models could generalize to the ImageCLEF MCQ format under consistent conditions.

5. Overview of the Systems and Results

5.1. Competition Results

Table 4 shows participant results on the test set on all 14 leaderboards. The most popular leaderboards were English, Multilingual, and Chinese, with 10, 9, and 7 teams participating, respectively. Some teams participated in multiple leaderboards, with two submitting to all 14 and another two submitting to 13. All teams significantly outperformed the baseline, except for *elenat* in the Multilingual and Bulgarian leaderboards. The task proved to be moderately difficult, with some teams achieving over 90% accuracy using the most recent commercial VLMs. Team **MSA** performed exceptionally well across the board, securing first place in 11 of the 13 leaderboards they entered.

Participating teams utilized a combination of proprietary and open-source large VLMs, including Qwen2.5-VL, Gemini, SmolLM, and Deepseek. The majority of approaches employed zero-shot or few-shot techniques, leveraging metainformation about visual elements. The most common and successful prompt strategies consisted of few-shot prompts that leverage image descriptions generated by different VLMs. Most notably, **MSA** used Gemini 2.5 Flash to generate captions of the input image, which is then further validated and refined before passing it in a zero-shot prompt using Gemini 2.5 Pro. **ContextDrift** employed a different prompting strategy, which relied on a sophisticated pipeline for

1-shot inference using Gemini 2.5 Flash, complemented by OCR-extracted textual content. Team **ymgclef** applied a multi-prompt ensemble by combining a base prompt, Chain-of-Thought prompt, and a Role-Playing prompt and achieved competitive results with this approach, despite using Qwen-VL, which underperformed compared to the Gemini models.

A few teams experimented with model fine-tuning; however, they achieved lower rankings, primarily due to the use of smaller models constrained by limited computational resources. **lekshmiscopevit** opted for parameter-efficient fine-tuning of Qwen2.5-VL with LoRa, achieving competitive results in the Multilingual leaderboard while reducing memory requirements by 75%. Team **plutohbj** also chose a fine-tuning approach, using LoRa for efficient tuning. They added cross-modal attention to enhance multimodal performance and applied stable optimization methods.

Overall, the top-performing systems, utilized by the leading two to three teams in most leaderboards, were from the Gemini family—specifically, Gemini 1.5 Pro and Gemini 2.5 Flash. Following in rankings was Qwen2.5-VL, which was the only model used for fine-tuning. Many participants combined different models or prompt strategies in ensembles to further boost their scores. An interesting observation is that participant models show comparable performance on parallel questions in Italian and Croatian, likely due to their shared use of the Latin script. In contrast, performance on Serbian is significantly lower (by approximately 20%), suggesting that models face greater difficulty with languages written in Cyrillic script. This performance gap is likely attributable to the lower representation of such languages in the training data.

5.2. Detailed System Descriptions

Table 5 presents an overview of the approaches used by the participants. In the rest of the section, we present short participant system descriptions. Teams are ordered alphabetically.

ayeshaamjad [41]

(Keywords: *Gemini-1.5, DeepSeek-R1-Distill-LLaMA, Structured Parsing, Modular Pipeline, Zero-shot*)

The authors propose a modular two-stage pipeline combining structured visual parsing with language-based reasoning. First, Gemini-1.5 Flash is used to decompose images into structured JSON outputs containing fields like question text, options, diagrams, labels, and tables. This is achieved through a carefully crafted zero-shot prompt that suppresses reasoning and ensures accurate multilingual layout parsing. The structured outputs are then passed to DeepSeek-R1-Distill-LLaMA, which uses a strict answer-only prompt to select the correct option. The entire pipeline is zero-shot, language-agnostic, and operates without fine-tuning, relying on prompt design for robustness and consistency.

bingezzzleep [42]

(Keywords: *Prompt-Enhancement, Feature Alignment, Qwen-VL*)

The authors propose a three-part system that constructs textual and visual features from the input image and fuses them into the prompt of the LLM. The prompt consists of three parts. The first part is a prompt-tuned standardized prompt, explaining the task. The second part is an encoding of the input image via Vision Transformer that is compressed into a smaller feature set by the cross-attention module. The third part is an LLM-processed description of the image, done by VLM. The three-part prompt is given to Qwen-VL-Max for question answering. They apply a prompt-tuning strategy for optimization. The results show that this prompt-enhancing strategy performs better than the direct model use.

ContextDrift [49]

(Keywords: *Gemini-2.5, LRM, LMM, Thinking Budget, Few-Shot, Zero-Shot*)

The authors propose a pipeline that performs few-shot inference using Gemini-2.5-Flash (Thinking) with extracted textual content. Textual content is extracted using OCRSpace, a cloud-based OCR service,

Table 4
Results for the ImageCLEF 2025 Multimodal Reasoning task on all 14 leaderboards. Baseline system submitted by the organizers. In the case of equal scores, participants are assigned the same rank and ordered alphabetically. †Participants submitted as different teams, but wrote a single working notes paper as co-authors.

Rank	Team	Acc	Rank	Team	Acc	Rank	Team	Acc
Multilingual			**English**			**Bulgarian**		
1	MSA	0.8140	1	ContextDrift†	0.8965	1	ContextDrift†	0.9050
2	ymgclef	0.5994	2	MSA	0.8652	1	ContextDrift†	0.9050
3	lekshmiscopevit	0.5770	3	ayeshaamjad	0.8125	2	ymgclef	0.7750
4	bingezzzleep	0.5619	4	ContextDrift†	0.8086	3	bingezzzleep	0.7500
5	plutohbj	0.5226	5	ymgclef	0.5938	3	MSA	0.7500
6	deng113abc	0.5195	6	deng113abc	0.5371	4	plutohbj	0.7300
7	mhl2001	0.4418	7	bingezzzleep	0.5312	5	baseline	0.2450
8	yaozihang	0.4376	8	plutohbj	0.4922	6	elenat	0.2350
9	baseline	0.2701	9	mhl2001	0.4629	**German**		
10	elenat	0.2188	10	yaozihang	0.4570	1	MSA	0.8915
Kazakh			11	elenat	0.2520	2	ymgclef	0.7403
1	MSA	0.8148	12	baseline	0.2480	3	bingezzzleep	0.6860
2	ymgclef	0.5350	**Chinese**			4	plutohbj	0.6783
3	bingezzzleep	0.4938	1	MSA	0.8305	5	yaozihang	0.4961
4	plutohbj	0.4444	2	ayeshaamjad	0.6560	6	mhl2001	0.4922
5	baseline	0.2738	3	plutohbj	0.5921	7	baseline	0.3101
Polish			4	bingezzzleep	0.5799	**Urdu**		
1	MSA	0.8224	5	mhl2001	0.5553	1	MSA	0.8067
2	ymgclef	0.7181	6	ymgclef	0.5283	2	ymgclef	0.3941
3	bingezzzleep	0.5792	7	yaozihang	0.4791	3	bingezzzleep	0.3569
4	plutohbj	0.5251	8	baseline	0.2678	3	yaozihang	0.3569
5	baseline	0.2934	**Arabic**			4	baseline	0.3011
Italian			1	MSA	0.6757	**Croatian**		
1	MSA	0.9212	2	ayeshaamjad	0.4775	1	MSA	0.9507
2	bingezzzleep	0.6059	3	mhl2001	0.4730	2	bingezzzleep	0.6207
2	plutohbj	0.6059	4	ymgclef	0.4324	3	ymgclef	0.5764
3	ymgclef	0.6010	5	plutohbj	0.3514	4	plutohbj	0.5616
4	baseline	0.2414	6	bingezzzleep	0.3243	5	baseline	0.2709
Spanish			7	baseline	0.2703	**Serbian**		
1	MSA	0.7198	**Hungarian**			1	MSA	0.7143
2	ymgclef	0.6696	1	ymgclef	0.6518	2	bingezzzleep	0.6059
3	bingezzzleep	0.6608	2	bingezzzleep	0.5425	3	ymgclef	0.5468
4	plutohbj	0.5723	3	plutohbj	0.4696	4	plutohbj	0.5320
5	baseline	0.3156	4	mhl2001	0.3563	5	baseline	0.2365
			5	baseline	0.2348			

and is passed along with the image to the Gemini 2.5 Flash (Thinking) model for answer classification. Their extensive experiments on the validation set reveal two critical findings. First, the impact of performing an OCR augmentation is marginal and varies depending on the language. Although the augmentation marginally improves model performance in the English dataset, it slightly decreases performance on the Bulgarian dataset. Second, the number of thinking tokens improves the overall performance of the model. However, a higher number of thinking tokens can have adverse effects, particularly on Biology and Chemistry questions containing graphical elements.

deng113abc [43]

(Keywords: *Qwen-VL-2.5, Prompt engineering, Chain-of-thought prompting*)

The authors propose a two-step prompting strategy called "Question Reconstruction before An-

Table 5
Overview of the approaches used by the participating systems.

Team	Reasoning						Captioning				Prompting			Adaptation			Misc		
	Qwen-VL	Gemini 1.5	Gemini 2.5	GPT 4.1	DeepSeek-R1	SmolLM	Gemini 1.5	Gemini 2.5	Qwen-VL	BLIP	Instruction	Chain-of-thought	Role-Playing	Zero-shot	Few-shot	Fine-tuning	LoRa	Optimization	Ensemble
ayeshaamjad [41]					☑		☑				☑			☑					
bingezzzleep [42]	☑								☑		☑			☑					
deng113abc [43]	☑											☑		☑					
elenat [44]						☑				☑	☑			☑					
lekshmiscopevit [45]	☑										☑					☑	☑	☑	
mhl2001 [46]	☑												☑		☑				
MSA [47]			☑				☑	☑			☑			☑	☑				☑
plutohbj [48]	☑										☑					☑	☑		
ContextDrift [49]			☑									☑	☑	☑	☑				
yaozihang [50]	☑												☑	☑				☑	
ymgclef [51]				☑							☑	☑	☑	☑					☑

swering" (QRA) for multimodal question answering. In the first step, the model uses image features and metadata such as language and subject to complete missing parts of the question. This helps the model better understand the problem. In the second step, the completed question is passed through a Chain-of-Thought prompting format, guiding the model to reason step-by-step and give an answer. The method works in a zero-shot setting, without OCR or fine-tuning. The team achieved 6th place in both multilingual and English tracks, showing strong performance across languages and improving significantly over the official baseline.

elenat [44]

(Keywords: *BLIP, Prompt Ablation, SmolLM-360M, Zero-shot*)

The authors propose a zero-shot pipeline that integrates image captioning with compact language model reasoning. First, images are processed using BLIP (Base or Large) to generate captions under three different prompt conditions: no prompt, “A photo of”, and “Describe what you see”. These prompts are designed to influence the verbosity and descriptiveness of the generated captions. The resulting caption is then inserted into a fixed-format prompt and passed to SmolLM-360M, which is a 360M parameter transformer optimized for low-resource inference. It selects the correct answer from multiple choices. The model is used without fine-tuning. This setup enables efficient multimodal reasoning under minimal compute, and allows the authors to assess how prompt formulation affects downstream answer prediction accuracy.

lekshmiscopevit [45]

(Keywords: *Qwen2.5-VL, LoRa, Model optimization, Efficient inference, Quantization*)

The authors propose an approach that utilizes the Qwen2.5-VL-72B Instruct model. They apply parameter-efficient fine-tuning technique, such as LoRA or QLoRA, explicitly optimized for the EXAMS-V dataset. Initial experiments suggest that fine-tuning with as few as 0.1% of parameters could yield accuracy improvements of 5–8% while maintaining generalization capabilities. To address performance disparities across languages, the authors used language-specific adapter modules that are dynamically integrated with the base model, with a focus on enhancing performance for underrepresented languages.

Through 4-bit quantization, specialized prompting techniques, and robust answer extraction methods, the team achieved comparable performance across all languages and subjects while reducing memory requirements by up to 75%.

mhl2001 [46]

(Keywords: *Qwen2.5-VL, Few-shot, Prompt engineering*)

The authors present a prompt-tuned pipeline leveraging Qwen2.5-VL-Plus to address the ImageCLEF 2025 Multimodal Reasoning task. Their system upgrades the baseline SmolVLM-2.5-1.7B model with Qwen2.5-VL-Plus and introduces a hybrid prompt design composed of multilingual system instructions and exemplar-based few-shot samples. The final prompt includes both macro-level role definition and task-specific examples, aiming to improve reasoning and answer classification. Their pipeline processes image-question pairs and generates structured responses using regex-based extraction. Evaluation on the EXAMS-V benchmark reveals a 63% performance gain over the baseline (from 0.2701 to 0.4418) on the Multilingual dataset, with other notable improvements in Chinese and German. Their method demonstrates robust multilingual generalization without task-specific fine-tuning.

MSA [47]

(Keywords: *Gemini 2.5, Gemini 1.5, Ensemble, Zero-shot*)

The authors propose a two-stage ensemble pipeline fully based on the use of a proprietary family of models - Gemini. During the first stage, the authors use Gemini 2.5 Flash to generate detailed descriptions of the input image. They use a 1-shot prompt to encourage the model to preserve uniformity. The generated caption, together with the input image, is passed to Gemini 1.5 Pro, which verifies the format in the correct language, inferred from the input. Then, the generated caption, together with the input image, is passed to Gemini 1.5 Pro, which verifies the correctness of the labels and translates any stray text into the declared language. In the second stage, the authors employ Gemini 2.5 Pro with a zero-shot prompt to solve the question and output the correct answer. The system achieves very strong results, placing 1st in 11 of 13 tracks.

plutohbj [48]

(Keywords: *LoRa, Meta learning, Qwen2.5-VL*)

The authors propose the Meta-LoRa framework, which enhances Qwen2.5-VL through three training techniques: (1) they use dynamic parameter adaptation to improve performance and efficiency during training (LoRa) (2), multimodal feature fusion by computing cross-modal attention to enhance understanding between both visual and textual modalities (3), stable optimization using cosine annealing, gradient clipping, and KL regularization. The model is trained on the Exams-V dataset and achieves competitive results on the leaderboard.

yaozihang [50]

(Keywords: *Qwen-VL, Prompt Engineering, Zero-shot*)

The authors propose an approach that utilizes the Qwen2.5-VL model in a zero-shot setting. They craft a prompt that ensures consistent analysis of the input images, containing questions, and outputs a concise answer in the correct format required by the task. To ensure the validity and efficiency of the experimental evaluation, this paper designs and implements a pipeline, comprising modules such as data organization compression, API request construction, and exception handling.

ymgclef [51]

(Keywords: *Multi-Prompt Ensemble, Qwen-VL-Plus, GPT-4.1-2025-04-14*)

The authors proposed a zero-shot ensemble pipeline in which the final classification is obtained by performing model inference using three different types of prompt templates: Base Prompt, Chain-of-Thought Prompt, and Role-Playing Prompt. The multi-prompt ensemble strategy enables the model to focus on informational features at different levels of abstraction, which improves the model's multimodal reasoning capabilities. Their experiments on an open-source model, Qwen-VL-Plus, and a proprietary model, GPT-4.1-2025-04-14, demonstrate the effectiveness of the multi-prompt strategy.

6. Conclusion and Future Work

We presented an overview of the Multimodal Reasoning task, part of the ImageCLEF lab at CLEF 2025. This task aimed to assess the reasoning abilities of large vision-language models when applied to complex visual and textual examination content. Participants were provided with an image containing a question and 3 to 5 answer choices. The goal was to select the single correct answer from the given options.

Most submissions employed zero-shot large multilingual vision-language models, primarily Qwen and Gemini, often incorporating extensive prompt engineering. A smaller subset of submissions explored few-shot learning and fine-tuning approaches, emphasizing model optimization. These approaches demonstrated that models with fewer parameters can achieve performance levels comparable to their larger counterparts. The top-performing systems were based on the proprietary Gemini model family, achieving over 80% accuracy overall and surpassing 90% in some languages. Systems using the open-source Qwen models ranked closely behind, suggesting that open-source approaches have significantly reduced the performance gap with proprietary models, particularly in low-resource languages. However, performance differences remain in favor of commercial models for high-resource languages.

In the future, we plan to expand both the linguistic and visual complexity of the evaluation setting. We plan to introduce additional languages, incorporate more diverse and challenging examination materials, and include more complex visual elements. Furthermore, the task will be extended to cover assessments from university-level examinations, broadening the scope and difficulty of the reasoning challenges.

Acknowledgement

The work of Dimitar Dimitrov and Ivan Koychev is partially funded by the EU NextGenerationEU project, through the National Recovery and Resilience Plan of the Republic of Bulgaria, project SUMMIT, No BG-RRP-2.004-0008.

Declaration on Generative AI

The authors did not employ any Generative AI tools in the preparation of this manuscript.

References

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, D. Parikh, VQA: visual question answering, in: 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015, IEEE Computer Society, 2015, pp. 2425–2433. URL: https://doi.org/10.1109/ICCV.2015.279. doi:10.1109/ICCV.2015.279.

[2] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, R. B. Girshick, CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, IEEE Computer Society, 2017, pp. 1988–1997. URL: https://doi.org/10.1109/CVPR.2017.215. doi:10.1109/CVPR.2017.215.

[3] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, Show and tell: A neural image caption generator, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015, IEEE Computer Society, 2015, pp. 3156–3164. URL: https://doi.org/10.1109/CVPR.2015.7298935. doi:10.1109/CVPR.2015.7298935.

[4] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, in: F. R. Bach, D. M. Blei (Eds.), Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015, volume 37 of *JMLR Workshop and Conference Proceedings*, JMLR.org, 2015, pp. 2048–2057. URL: http://proceedings.mlr.press/v37/xuc15.html.

[5] J. Lu, D. Batra, D. Parikh, S. Lee, Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, in: H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, 2019, pp. 13–23. URL: https://proceedings.neurips.cc/paper/2019/hash/c74d97b01eae257e44aa9d5bade97baf-Abstract.html.

[6] H. Tan, M. Bansal, LXMERT: Learning cross-modality encoder representations from transformers, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 5100–5111. URL: https://aclanthology.org/D19-1514. doi:10.18653/v1/D19-1514.

[7] Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, J. Liu, Uniter: Universal image-text representation learning, in: European conference on computer vision, Springer, 2020, pp. 104–120.

[8] 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: M. Meila, T. Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event, volume 139 of *Proceedings of Machine Learning Research*, PMLR, 2021, pp. 8748–8763. URL: http://proceedings.mlr.press/v139/radford21a.html.

[9] J. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. L. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. Binkowski, R. Barreira, O. Vinyals, A. Zisserman, K. Simonyan, Flamingo: a visual language model for few-shot learning, in: S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh (Eds.), Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022, 2022. URL: http://papers.nips.cc/paper_files/paper/2022/hash/960a172bc7fbf0177ccccbb411a7d800-Abstract-Conference.html.

[10] D. Driess, F. Xia, M. S. M. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, W. Huang, Y. Chebotar, P. Sermanet, D. Duckworth, S. Levine, V. Vanhoucke, K. Hausman, M. Toussaint, K. Greff, A. Zeng, I. Mordatch, P. Florence, Palm-e: An embodied multimodal language model, in: A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, J. Scarlett (Eds.), International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA, volume 202 of *Proceedings of Machine Learning Research*, PMLR, 2023, pp. 8469–8488. URL: https://proceedings.mlr.press/v202/driess23a.html.

[11] L. Fan, W. Hua, X. Li, K. Zhu, M. Jin, L. Li, H. Ling, J. Chi, J. Wang, X. Ma, et al., Nphardeval4v: A dynamic reasoning benchmark of multimodal large language models, ArXiv preprint abs/2403.01777 (2024). URL: https://arxiv.org/abs/2403.01777.

[12] K. Wang, J. Pan, W. Shi, Z. Lu, H. Ren, A. Zhou, M. Zhan, H. Li, Measuring multimodal mathematical reasoning with math-vision dataset, in: A. Globersons, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. M. Tomczak, C. Zhang (Eds.), Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024, 2024. URL: http://papers.nips.cc/paper_files/paper/2024/hash/

ad0edc7d5fa1a783f063646968b7315b-Abstract-Datasets_and_Benchmarks_Track.html.

[13] J. Kil, Z. Mai, J. Lee, A. Chowdhury, Z. Wang, K. Cheng, L. Wang, Y. Liu, W. Chao, Mllm-compbench: A comparative reasoning benchmark for multimodal llms, in: A. Globersons, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. M. Tomczak, C. Zhang (Eds.), Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024, 2024. URL: http://papers.nips.cc/paper_files/paper/2024/hash/32923dff09f75cf1974c145764a523e2-Abstract-Datasets_and_Benchmarks_Track.html.

[14] D. Dua, Y. Wang, P. Dasigi, G. Stanovsky, S. Singh, M. Gardner, DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 2368–2378. URL: https://aclanthology.org/N19-1246. doi:10.18653/v1/N19-1246.

[15] Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. Cohen, R. Salakhutdinov, C. D. Manning, HotpotQA: A dataset for diverse, explainable multi-hop question answering, in: E. Riloff, D. Chiang, J. Hockenmaier, J. Tsujii (Eds.), Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 2369–2380. URL: https://aclanthology.org/D18-1259. doi:10.18653/v1/D18-1259.

[16] M. Geva, D. Khashabi, E. Segal, T. Khot, D. Roth, J. Berant, Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies, Transactions of the Association for Computational Linguistics 9 (2021) 346–361. URL: https://aclanthology.org/2021.tacl-1.21. doi:10.1162/tacl_a_00370.

[17] W. Yu, Z. Jiang, Y. Dong, J. Feng, Reclor: A reading comprehension dataset requiring logical reasoning, in: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020, OpenReview.net, 2020. URL: https://openreview.net/forum?id=HJgJtT4tvB.

[18] J. Liu, L. Cui, H. Liu, D. Huang, Y. Wang, Y. Zhang, Logiqa: A challenge dataset for machine reading comprehension with logical reasoning, in: C. Bessiere (Ed.), Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020, ijcai.org, 2020, pp. 3622–3628. URL: https://doi.org/10.24963/ijcai.2020/501. doi:10.24963/ijcai.2020/501.

[19] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, et al., Training verifiers to solve math word problems, ArXiv preprint abs/2110.14168 (2021). URL: https://arxiv.org/abs/2110.14168.

[20] D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, J. Steinhardt, Measuring mathematical problem solving with the math dataset, ArXiv preprint abs/2103.03874 (2021). URL: https://arxiv.org/abs/2103.03874.

[21] O. Choukrani, I. Malek, D. Orel, Z. Xie, Z. Iklassov, M. Takáč, S. Lahlou, Llm-babybench: Understanding and evaluating grounded planning and reasoning in llms, ArXiv preprint abs/2505.12135 (2025). URL: https://arxiv.org/abs/2505.12135.

[22] Z. Chen, W. Chen, C. Smiley, S. Shah, I. Borova, D. Langdon, R. Moussa, M. Beane, T.-H. Huang, B. Routledge, W. Y. Wang, FinQA: A dataset of numerical reasoning over financial data, in: M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 3697–3711. URL: https://aclanthology.org/2021.emnlp-main.300. doi:10.18653/v1/2021.emnlp-main.300.

[23] Z. Xie, D. Sahnan, D. Banerjee, G. Georgiev, R. Thareja, H. Madmoun, J. Su, A. Singh, Y. Wang, R. Xing, et al., Finchain: A symbolic benchmark for verifiable chain-of-thought financial reasoning, ArXiv preprint abs/2506.02515 (2025). URL: https://arxiv.org/abs/2506.02515.

[24] X. Yue, Y. Ni, T. Zheng, K. Zhang, R. Liu, G. Zhang, S. Stevens, D. Jiang, W. Ren, Y. Sun, C. Wei, B. Yu, R. Yuan, R. Sun, M. Yin, B. Zheng, Z. Yang, Y. Liu, W. Huang, H. Sun, Y. Su, W. Chen, MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA,

June 16-22, 2024, IEEE, 2024, pp. 9556–9567. URL: https://doi.org/10.1109/CVPR52733.2024.00913. doi:10.1109/CVPR52733.2024.00913.

[25] P. Lu, S. Mishra, T. Xia, L. Qiu, K. Chang, S. Zhu, O. Tafjord, P. Clark, A. Kalyan, Learn to explain: Multimodal reasoning via thought chains for science question answering, in: S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh (Eds.), Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022, 2022. URL: http://papers.nips.cc/paper_files/paper/2022/hash/11332b6b6cf4485b84afadb1352d3a9a-Abstract-Conference.html.

[26] 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.

[27] 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.

[28] I. Salazar, M. F. Burda, S. B. Islam, A. S. Moakhar, S. Singh, F. Farestam, A. Romanou, D. Boiko, D. Khullar, M. Zhang, et al., Kaleidoscope: In-language exams for massively multilingual vision evaluation, ArXiv preprint abs/2504.07072 (2025). URL: https://arxiv.org/abs/2504.07072.

[29] S. Park, G. Kim, Evaluating multimodal generative ai with korean educational standards, ArXiv preprint abs/2502.15422 (2025). URL: https://arxiv.org/abs/2502.15422.

[30] P. Zhou, F. Zhang, X. Peng, Z. Xu, J. Ai, Y. Qiu, C. Li, Z. Li, M. Li, Y. Feng, et al., Mdk12-bench: A multi-discipline benchmark for evaluating reasoning in multimodal large language models, ArXiv preprint abs/2504.05782 (2025). URL: https://arxiv.org/abs/2504.05782.

[31] M. S. Hee, W. Chong, R. K. Lee, Decoding the underlying meaning of multimodal hateful memes, in: Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China, ijcai.org, 2023, pp. 5995–6003. URL: https://doi.org/10.24963/ijcai.2023/665. doi:10.24963/IJCAI.2023/665.

[32] M. S. Hee, S. Sharma, R. Cao, P. Nandi, P. Nakov, T. Chakraborty, R. Lee, Recent advances in online hate speech moderation: Multimodality and the role of large models, Findings of the Association for Computational Linguistics: EMNLP 2024 (2024) 4407–4419.

[33] D. Dimitrov, F. Alam, M. Hasanain, A. Hasnat, F. Silvestri, P. Nakov, G. Da San Martino, SemEval-2024 task 4: Multilingual detection of persuasion techniques in memes, in: A. K. Ojha, A. S. Doğruöz, H. Tayyar Madabushi, G. Da San Martino, S. Rosenthal, A. Rosá (Eds.), Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 2009–2026. URL: https://aclanthology.org/2024.semeval-1.275.

[34] C. Ma, J. Ding, J. Zhang, Z. Ma, H. Qing, B. Gao, L. Chen, M. Song, et al., Sci-reason: A dataset with chain-of-thought rationales for complex multimodal reasoning in academic areas, ArXiv preprint abs/2504.06637 (2025). URL: https://arxiv.org/abs/2504.06637.

[35] Y. Susanto, A. V. Hulagadri, J. R. Montalan, J. G. Ngui, X. B. Yong, W. Leong, H. Rengarajan, P. Limkonchotiwat, Y. Mai, W. C. Tjhi, Sea-helm: Southeast asian holistic evaluation of language models, ArXiv preprint abs/2502.14301 (2025). URL: https://arxiv.org/abs/2502.14301.

[36] R. Ng, T. N. Nguyen, Y. Huang, N. C. Tai, W. Y. Leong, W. Q. Leong, X. Yong, J. G. Ngui, Y. Susanto, N. Cheng, et al., Sea-lion: Southeast asian languages in one network, ArXiv preprint abs/2504.05747 (2025). URL: https://arxiv.org/abs/2504.05747.

[37] M. Togmanov, N. Mukhituly, D. Turmakhan, J. Mansurov, M. Goloburda, A. Sakip, Z. Xie, Y. Wang, B. Syzdykov, N. Laiyk, et al., Kazmmlu: Evaluating language models on kazakh, russian, and regional knowledge of kazakhstan, ArXiv preprint abs/2502.12829 (2025). URL: https://arxiv.org/abs/2502.12829.

[38] R. C. Ng, N. Prakash, M. S. Hee, K. T. W. Choo, R. K.-w. Lee, SGHateCheck: Functional tests for detecting hate speech in low-resource languages of Singapore, in: Y.-L. Chung, Z. Talat, D. Nozza, F. M. Plaza-del Arco, P. Röttger, A. Mostafazadeh Davani, A. Calabrese (Eds.), Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024), Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 312–327. URL: https://aclanthology.org/2024.woah-1.24.

[39] S. Cahyawijaya, H. Lovenia, J. R. A. Moniz, T. H. Wong, M. R. Farhansyah, T. T. Maung, F. Hudi, D. Anugraha, M. R. S. Habibi, M. R. Qorib, et al., Crowdsource, crawl, or generate? creating sea-vl, a multicultural vision-language dataset for southeast asia, ArXiv preprint abs/2503.07920 (2025). URL: https://arxiv.org/abs/2503.07920.

[40] 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.

[41] A. Amjad, F. Seemab, S. Kausar, S. Latif, M. Fatima, ayeshaamjad at ImageCLEF 2025 Multimodal Reasoning: visual question answering with structured data extraction and robust reasoning, in: [52], 2025.

[42] Q. Wu, L. Kong, J. Yan, J. Li, Team bingezzzleep at ImageCLEF 2025 Multimodal Reasoning: a multimodal feature alignment prompt-enhanced method for multimodal reasoning, in: [52], 2025.

[43] S. Deng, G. Niu, X. Yao, H. Mo, T. Li, S. Jiao, Bridging the modality gap through cot-enhanced multimodal reasoning, in: [52], 2025.

[44] E. Tosheva, D. Dimitrov, I. Koychev, P. Nakov, Elenat at Image CLEF 2025 Multimodal Reasoning: Zero-shot reasoning with blip and smollm, in: [52], 2025.

[45] T. Srikumar, S. Kesavan, A. M B, D. Samuel, K. Maneesh Ram, G. E, V. K. Singh, L. Kalinathan, Leveraging qwen2.5-vl-72b-instruct for visual question answering: A Study on the EXAMS-V Benchmark in ImageCLEF 2025, in: [52], 2025.

[46] H. Mo, G. Niu, S. Deng, X. Yao, T. Li, S. Jiao, Multimodal Reasoning in Multilingual Visual Question Answering: A prompt-tuned qwen2.5-vl-plus approach, in: [52], 2025.

[47] A. Seif, M. Younes, A. Moustafa, A. Allam, H. Moustafa, MSA at ImageCLEF 2025 Multimodal Reasoning: multilingual multimodal reasoning with ensemble vision-language models, in: [52], 2025.

[48] B. Huang, C. Zhong, K. Yan, Team plutohbj at ImageCLEF 2025 Multimodal Reasoning: meta-learning lora fine-tuning for multimodal reasoning, in: [52], 2025.

[49] V. Krazheva, D. Markova, D. Dimitrov, I. Koychev, P. Nakov, ContextDrift at ImageCLEF 2025 Multimodal Reasoning: Evaluating vlms' multimodal, multilingual and multidomain reasoning capabilities via thinking budget variations and textual augmentation, in: [52], 2025.

[50] X. Yao, G. Niu, T. Li, H. Mo, S. Deng, S. Jiao, Enhancing Multilingual VQA with structured prompts and vision-language alignment, in: [52], 2025.

[51] J. Yan, L. Kong, Q. Wu, J. Li, Multi-prompt ensemble reasoning for multimodalreasoning, in: [52], 2025.

[52] G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CLEF 2025, 2025.