Overview of the ImageCLEF 2026 Task on Multimodal Reasoning

Dimitar Dimitrov[1,*], Ming Shan Hee[2,*], Momina Ahsan[2,*], Sarfraz Ahmad[2], Georgi Pachov[1], Dimitrina Zlatkova[1], Zhuohan Xie[2], Preslav Nakov[2] and Ivan Koychev[1]

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

Abstract

This paper describes the ImageCLEF 2026 Multimodal Reasoning task, which focuses on evaluating vision–language systems on multilingual examination questions. The 2026 edition broadens the task formulation from option selection alone to a setting that includes both multiple-choice and direct answer generation. Participants were asked to solve four tracks: Visual MCQ, Textual MCQ, Visual OpenQA, and Textual OpenQA, making it possible to study the effects of both visual content and answer format. The benchmark contains 5,026 questions in English, Chinese, Bulgarian, Croatian, Italian, and Serbian, including 2,770 multiple-choice questions and 2,256 open-ended questions. The task had moderate participation with a total of 53 registered teams. Of these, 13 teams submitted results on the test set across all leaderboards, with 141 graded submissions overall. The shared task attracted submissions based on a diverse set of methods, including prompting, fine-tuning, LoRA-based adaptation, self-consistency, test-time scaling, answer extraction, language-specific prompting, and multi-stage reasoning. The strongest systems achieved 84.06% accuracy on Visual MCQ and 75.38% accuracy on Textual MCQ. In the open-ended setting, the best submissions reached COMET scores of 0.6488 on Visual OpenQA and 0.6563 on Textual OpenQA. The results indicate that current vision-language models are more reliable when selecting from answer options than when generating short answers directly. They also reveal substantial variation across languages and tracks, suggesting that multilingual visual reasoning, robust OCR, structured visual understanding, and answer normalization remain important challenges.

Keywords

multimodal reasoning, vision-language models, multilingual evaluation, visual question answering, free-form questions, open-ended questions, multiple-choice questions, low-resource languages

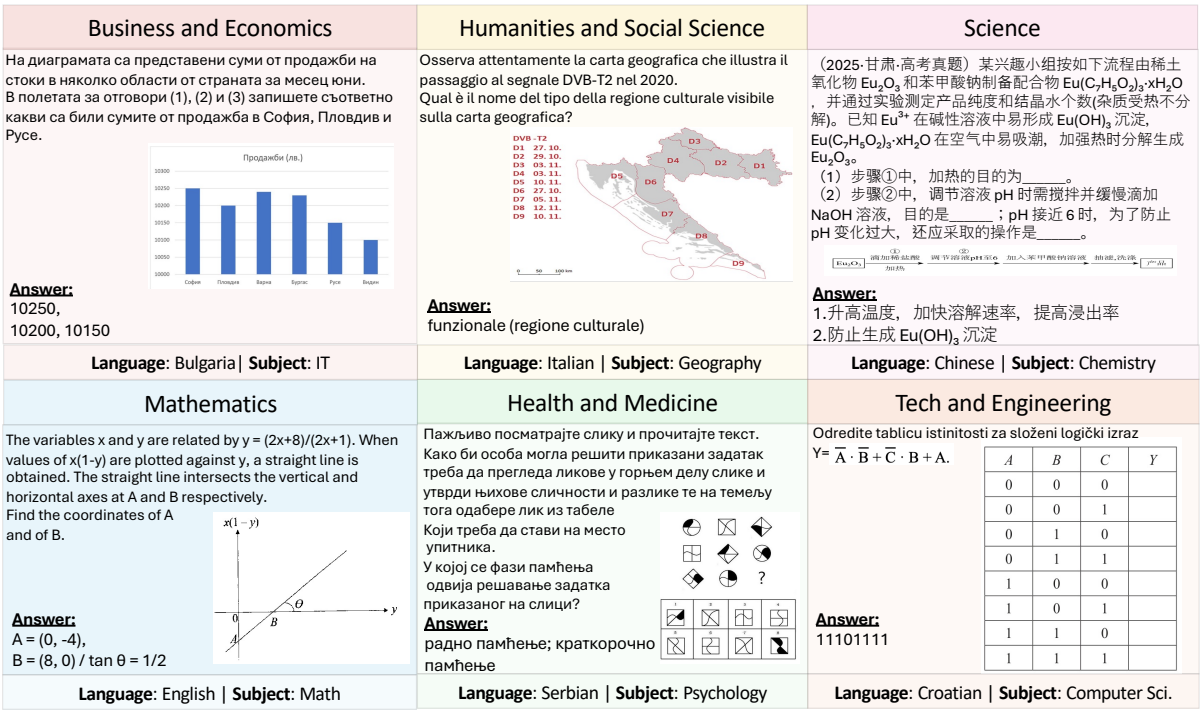
1. Introduction

Multimodal reasoning has become an increasingly important capability for vision–language models, particularly as these systems are applied to tasks that require the joint interpretation of images, text, symbols, and structured visual information. Foundational benchmarks such as VQA [1] and CLEVR [2] established visual question answering as a central testbed for multimodal understanding, while subsequent work on image captioning and vision–language pre-training showed how visual and textual representations can be learned jointly [3, 4, 5, 6, 7]. More recent multimodal large language models have further expanded these capabilities, enabling systems to process complex visual inputs and answer natural-language questions in increasingly flexible ways.

However, exam-style reasoning remains difficult for current models. Educational questions are often visually dense and linguistically precise: a single item may contain a diagram, a graph, a table, a formula, or a figure, together with instructions and answer constraints that must be interpreted correctly. Solving such questions requires a system to identify relevant information, combine it across modalities, and

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany
*Corresponding author.
ilijanovd@fmi.uni-sofia.bg (D. Dimitrov); mingshan.hee@mbzuai.ac.ae (M. S. Hee); momina.ahsan@mbzuai.ac.ae (M. Ahsan); sarfraz.ahmad@mbzuai.ac.ae (S. Ahmad); georgi.patchov@gmail.com (G. Pachov); didi.zlatkova93@gmail.com (D. Zlatkova); zhuohan.xie@mbzuai.ac.ae (Z. Xie); preslav.nakov@mbzuai.ac.ae (P. Nakov); koychev@fmi.uni-sofia.bg (I. Koychev)
0000-0003-1308-180X (D. Dimitrov); 0000-0002-6328-5889 (M. S. Hee); 0009-0004-5988-9673 (M. Ahsan); 0000-0003-1039-3787 (S. Ahmad); 0009-0008-1546-3805 (G. Pachov); 0009-0006-6766-4559 (D. Zlatkova); 0009-0008-2650-2857 (Z. Xie); 0000-0002-3600-1510 (P. Nakov); 0000-0003-3919-030X (I. Koychev)

Figure 1: Examples from the ImageCLEF 2026 Multimodal Reasoning dataset across languages, subjects, and visual formats. The examples illustrate the diversity of exam-style reasoning in the task, including charts, maps, scientific procedures, mathematical plots, visual pattern reasoning, and truth tables.



apply subject-specific reasoning. These requirements are further complicated in multilingual settings, where models must operate across different scripts, terminology, educational formats, and language-resource conditions. Figure 1 illustrates this diversity with examples from different languages and subjects, including questions involving charts, maps, chemistry procedures, mathematical plots, visual pattern reasoning, and truth tables. Therefore, evaluating multimodal models on realistic examination material provides a useful setting for measuring both reasoning ability and robustness.

Existing benchmarks have made substantial progress toward evaluating multimodal reasoning. Some focus on general visual question answering, while others target specific skills such as mathematical reasoning with diagrams, scientific question answering, or multidisciplinary academic understanding [8, 9, 10]. Multilingual educational benchmarks such as EXAMS-V [11] further highlight the importance of evaluating vision–language models beyond English-only settings. Nevertheless, many benchmarks rely primarily on multiple-choice evaluation or focus on a single answer format. Real examination settings, in contrast, include both option selection and free-form answer generation, and they vary in how much visual information is needed to solve each question.

To address this gap, we organized the ImageCLEF 2026 Multimodal Reasoning task as part of the ImageCLEF 2026 Lab. The task evaluates multilingual exam-style question answering from question images and is divided into four tracks: Visual MCQ, Textual MCQ, Visual OpenQA, and Textual OpenQA. This design separates the effect of answer format from the effect of visual content. The MCQ tracks assess whether systems can select the correct option from a set of candidates, whereas the OpenQA tracks require systems to generate the answer directly. The visual tracks contain questions with diagrams, graphs, tables, or other figures, while the textual tracks contain questions without explicit visual elements but still presented as image-rendered exam items.

The dataset contains 5,026 questions in six languages: English, Chinese, Bulgarian, Croatian, Italian, and Serbian. It consists of 2,770 multiple-choice questions and 2,256 open-ended questions. The questions were collected from publicly available examination materials and processed to extract question images, gold answers, and metadata describing the presence and type of visual elements. The resulting data

was deduplicated and manually checked to ensure that the question crops, answers, and metadata were suitable for evaluation. For the open-ended training split, we additionally repurposed multiple-choice questions by removing answer options and using the corresponding answer text as the target.

The shared task was conducted in development and test phases. During development, participants could submit to open-ended leaderboards and receive feedback. During the final phase, participants submitted predictions to all four tracks under blind evaluation conditions. The submitted systems covered a range of approaches, including zero-shot and few-shot prompting, fine-tuning, LoRA adaptation, self-consistency, answer extraction, test-time scaling, and language-specific prompting. The results show that current systems can perform well on some multiple-choice settings, but open-ended generation and multilingual visual reasoning remain challenging, with clear variation across languages and tracks.

This paper provides an overview of the ImageCLEF 2026 Multimodal Reasoning task, including the dataset construction process, task organization, evaluation measures, submitted systems, and final results. By providing a multilingual benchmark with both multiple-choice and open-ended formats, the task aims to support more realistic evaluation of vision–language models and to encourage further research on robust multimodal reasoning over educational content.

2. Related Work

2.1. Multimodal Reasoning Benchmarks

Multimodal reasoning requires models to combine information from language, images, symbols, and structured visual elements. This setting is central to many real-world tasks, including educational assessment, chart and document understanding, scientific reasoning, and content analysis. Early visual question answering benchmarks such as VQA [1] and CLEVR [2] established image-based question answering as a standard evaluation paradigm, while later benchmarks introduced more complex visual structures, readable text, and domain-specific reasoning requirements. Recent work has further expanded multimodal evaluation to include charts, diagrams, tables, scientific figures, and document-like layouts, reflecting the fact that many practical reasoning tasks cannot be solved from text or images alone.

A number of recent benchmarks focus specifically on reasoning over complex multimodal inputs. ScienceQA evaluates science question answering with textual and visual contexts [8], while MathVista and MATH-V target visual mathematical reasoning over diagrams, tables, plots, and other structured images [12, 9]. MMMU evaluates expert-level multimodal reasoning across multiple academic disciplines and image types [13], and MLLM-COMPBench studies comparative reasoning over visual inputs [14]. These benchmarks show that current multimodal models still struggle with tasks that require precise visual grounding, symbolic interpretation, and multi-step reasoning. However, many such datasets are either primarily English-focused, domain-specific, or centered on a single evaluation format, which limits their ability to capture the diversity of real educational assessments.

2.2. Educational and Multilingual Exam Benchmarks

Educational benchmarks provide a natural testbed for evaluating multimodal reasoning because exam questions often combine text, visual information, subject knowledge, and answer constraints. M3Exam introduced a multilingual and multilevel examination benchmark based on real human exams, covering both text-only and multimodal questions [15]. EXAMS-V [11] extended this direction by introducing a multilingual, multidisciplinary, multimodal benchmark of school-exam questions in image form, covering multiple languages, subjects, and visual element types. Other recent resources further broaden the evaluation landscape: Kaleidoscope focuses on in-language exam questions across many languages and subjects [16], MDK12-Bench collects large-scale K–12 multimodal questions with difficulty and explanation annotations [17], and several domain-specific benchmarks evaluate multimodal reasoning in mathematics, science, and other educational settings.

These datasets highlight the importance of evaluating models beyond English and beyond generic image understanding. Multilingual exam data introduces additional challenges, including script variation, language-specific terminology, different educational conventions, and uneven model performance across high- and low-resource languages. This is particularly important for shared evaluations, where the goal is not only to measure overall performance, but also to identify which languages, subjects, and visual formats remain difficult for current systems. The ImageCLEF 2025 Multimodal Reasoning task followed this motivation by evaluating systems on multilingual multiple-choice exam questions, using EXAMS-V together with an additional test set [18]. The ImageCLEF 2026 task builds on this line of work with a broader formulation: it keeps the multilingual educational setting, but extends the evaluation beyond multiple-choice answering by adding open-ended answer generation and by separating visual and textual questions into distinct tracks.

2.3. Multiple-Choice and Open-Ended Multimodal Question Answering

Most educational multimodal benchmarks rely heavily on multiple-choice evaluation. This format has practical advantages: it supports straightforward accuracy-based scoring, reduces ambiguity in evaluation, and allows direct comparison across systems. However, multiple-choice questions can also make the task easier by allowing models to exploit answer candidates, eliminate unlikely options, or rely on surface cues in the choices. In contrast, open-ended question answering requires models to generate the answer directly, without access to predefined options. This setting better reflects many real educational tasks, but it is also harder to evaluate because correct answers may vary in form, wording, units, or notation.

Open-ended visual and document-based question answering has therefore received increasing attention. TextVQA and OCR-VQA evaluate questions requiring text recognition from images [19, 20], while InfographicVQA, ChartQA, and FlowVQA focus on visually structured documents, charts, and flowcharts [21, 22, 23]. These datasets emphasize the importance of reading, grounding, and reasoning over visual layouts. At the same time, open-ended evaluation remains challenging, especially when answers are short, multilingual, mathematical, or semantically equivalent despite surface differences. For this reason, open-ended evaluation often requires semantic metrics in addition to lexical overlap metrics.

The ImageCLEF 2026 Multimodal Reasoning task combines both evaluation paradigms in a single shared benchmark. The MCQ tracks evaluate option selection using accuracy, while the OpenQA tracks evaluate free-form answer generation using text-generation metrics, with COMET as the official ranking metric. This design makes it possible to compare how systems behave when answer options are available versus when answers must be generated directly. It also enables separate analysis of visual and textual questions, allowing the evaluation to distinguish between failures caused by visual reasoning, language understanding, and answer generation.

3. Dataset

3.1. Data Sources

We collected the dataset from publicly available exam materials, covering levels from primary school to high school. Most of the questions are aimed at high-school students. For Bulgarian, we used high-school exam materials and added primary and middle-school Olympiad questions. The Chinese data comes from Gaokao exams, while the English data was taken from Singaporean primary and secondary school test papers. The Croatian, Italian, and Serbian data all come from Croatian national high-school exams. The Italian and Serbian sets are parallel versions of the Croatian data, meaning they contain the same exam questions translated into Italian and Serbian.

3.2. Data Extraction

Question extraction We used Gemini 3 Pro to identify the bounding box of each question on every page of the input PDF. Once the bounding boxes were detected, the questions were cropped and saved as image files using the OpenCV Python library.

Answer extraction We used Gemini 3 Flash to locate and extract the answer corresponding to each question identified during the question-extraction step. We ensured that each PDF included answers for the exam questions. In rare cases where an answer could not be found, the model was instructed to return "NOT_FOUND". All answers were extracted in LaTeX format and included only the final correct answer, without step-by-step reasoning.

Metadata extraction We also used Gemini 3 Flash to extract question-level metadata. In particular, the model identified whether a question contained visual elements and assigned the relevant category: Diagram, Graph, Table, or Other/Figure.

Deduplication Finally, we deduplicated the collected data. We first detected exact duplicates using MD5 hashes [24]. We then used the structural similarity index measure (SSIM) [25] to identify visually similar images, followed by perceptual hashing (pHash) [26] to detect near-duplicates. When duplicates were found, we kept the largest file as the representative image.

Verification We manually verified the extracted questions, answers, and metadata. First, we checked the question crops and adjusted them when necessary so that each crop fully contained the corresponding question. During this step, we also removed questions whose answer format differed from standard multiple-choice or open-ended questions, such as drawing tasks, long explanations, and proofs. Next, we reviewed the extracted answers. Most issues were related to LaTeX formatting, especially special characters that were either not escaped or escaped incorrectly. Finally, we verified the extracted metadata. This step was nearly perfect, with only two incorrect cases, both caused by overlaps between neighbouring questions.

Repurposing MCQs In addition to the open-ended questions we collected, we also repurposed MCQ-style questions from EXAMS-V [11] as open-ended questions. To do this, we removed the answer choices from the question images and replaced the correct answer key with the corresponding answer text. This repurposing was done to provide participants with additional training data.

3.3. Data Statistics

The dataset for this task contains a total of 5 026 questions[1], comprising 2 770 multiple-choice questions (MCQs) and 2 256 open-ended questions.

MCQ Data Table 1 presents the per-language distribution and key statistics of the MCQ test set. The **# Textual** column reports the 1,653 MCQs without visual elements, which are used for the Textual MCQ leaderboard, while the **# Visual** column reports the 1,117 MCQs with visual elements, which are used for the Visual MCQ leaderboard. Table 2 shows the per-language visual element distribution. Diagrams are the most frequent visual element in the MCQ test set, accounting for 644 out of 1,248 visual elements, followed by tables, graphs, and other visual elements.

[1] https://huggingface.co/collections/SU-FMI-AI/imageclef-2026-multimodal-reasoning

Table 1
MCQ test data statistics. Here, # **Visual** refers to questions with visual elements, and # **Textual** refers to questions without visual elements.

Language	ISO	Family	Grade	#Subj.	# Textual	# Visual	# Total Q.
English	en	Germanic	8-12	5	240	500	740
Chinese	zh	Sino-Tibetan	12	4	267	342	609
Bulgarian	bg	Slavic	4-12	11	451	110	561
Croatian	hr	Slavic	12	13	335	57	392
Serbian	sr	Slavic	12	11	187	54	241
Italian	it	Romance	12	10	173	54	227

Table 2
Visual element distribution of the **MCQ** test set. Note that a single question can contain multiple types of visual elements.

Language	Diagram	Graph	Table	Other	Total
English	308	72	203	10	593
Chinese	249	55	51	17	372
Bulgarian	35	5	18	57	115
Croatian	18	15	10	15	58
Italian	17	15	8	15	55
Serbian	17	15	8	15	55

OpenQA Data Table 3 presents the per-language distribution of the open-ended questions across the train, dev, and test splits. The OpenQA subset contains 2,256 questions in total and is split into two parts corresponding to the different leaderboards: 1,049 textual questions and 1,207 questions with visual elements. For the training data, we released repurposed MCQs, as described in Section 3.2, while the dev and test sets were constructed from the collected real open-ended questions.. Table 4 shows the distribution of visual elements across the OpenQA splits. Diagrams are the most frequent visual element, appearing in 625 questions, corresponding to 51.78% of the visual OpenQA data, followed by graphs, other figures, and tables. Since a single question may contain multiple visual elements, the reported percentages do not sum to 100%.

Table 3
Number of **open-ended** questions with textual and with visual element data per language.

Language	Textual				Visual			
	Train	Dev	Test	Total	Train	Dev	Test	Total
Bulgarian	50	75	75	**200**	84	50	75	**209**
Chinese	50	50	49	**149**	139	50	75	**264**
Croatian	50	75	75	**200**	96	50	75	**221**
English	50	75	75	**200**	66	50	75	**191**
Italian	50	25	75	**150**	87	20	74	**181**
Serbian	50	25	75	**150**	56	20	65	**141**
Total	**300**	**325**	**424**	**1,049**	**528**	**240**	**439**	**1,207**

Table 4
Visual element distribution of the **open-ended** train, dev, and test sets. Note that a single question can contain multiple types of visual elements, therefore, percentages do not sum to 100%.

Visual Element	Train	Dev	Test	Total	Percentage
Diagrams	241	149	235	625	51.78%
Graphs	107	44	98	249	20.63%
Tables	92	35	71	198	16.40%
Other Figures	88	35	96	219	18.14%

4. Evaluation Framework

4.1. Task Organization

The task was organized in two phases. During the development phase, participants worked on open-ended questions and were provided with repurposed MCQs as training data, along with a development leaderboard that offered instant feedback on real open-ended questions. Submissions were made to two separate leaderboards: Visual OpenQA, for questions containing visual elements, and Textual OpenQA, for questions without visual elements.

In the test phase, we released new MCQ data and additional real open-ended questions as the test dataset, together with metadata describing the visual components present in the questions. Participants submitted their results to four leaderboards: Visual OpenQA, Textual OpenQA, Visual MCQ, and Textual MCQ. Multiple submissions were allowed during this phase, but no feedback was provided. The final rankings were determined based on each participant's last submission at the end of the test phase.

This two-phase setup allowed participants to develop and tune their systems using feedback during the development phase, while ensuring that final evaluation was performed under blind test conditions.

4.2. Evaluation Measures

The ImageCLEF 2026 Multimodal Reasoning task consists of two evaluation settings: multiple-choice question answering (MCQ) and open question answering (OpenQA). The MCQ tracks are evaluated using classification accuracy, while the OpenQA tracks are evaluated using text generation metrics. This setup reflects the different objectives of the two tasks, where MCQ requires selecting a single correct option and OpenQA requires generating a complete answer.

For the MCQ tracks, we use accuracy as the official evaluation metric. Accuracy is computed as the percentage of questions for which the predicted answer exactly matches the gold answer. Since each question has a single correct option, accuracy provides a direct and interpretable measure of system performance across languages and subjects.

For the OpenQA tracks, systems generate free-form answers that are compared against the reference answers using automatic text generation metrics. The official ranking metric is COMET [27], which evaluates the semantic similarity between the generated and reference answers. In addition, we report BLEU [28], ROUGE-L [29], and METEOR [30] to provide complementary lexical and sequence-level evaluations. The leaderboards are ranked using the overall COMET score, while language-specific COMET scores are also reported together with the additional metrics.

4.3. Baselines

As both the Visual MCQ and OpenQA tasks provide examination questions in image format, we first extract the question text using the open-weight vision-language model `Qwen/Qwen3.5-35B-A3B`. The extracted text is used as auxiliary OCR guidance rather than as a replacement for the original image. To make training and inference feasible while preserving visual detail, images are resized only when they exceed the preprocessing limits: a maximum total area of 4,194,304 pixels or a maximum long side of 16,777,216 pixels, while preserving the aspect ratio.

Both the MCQ and OpenQA pipelines fine-tune `Qwen/Qwen3.5-9B`[2] using supervised instruction tuning with LoRA. After running several empirical experiments, the final configuration applies LoRA only to the attention modules, with a rank of 32, an alpha of 32, a dropout rate of 0.0, a learning rate of 2×10^{-5}, a maximum gradient norm of 0.5, a maximum sequence length of 512, gradient accumulation of 1, and two training epochs. The vision encoder is frozen in both pipelines, while the vision-to-language projection varies depending on the pipeline. In the OpenQA pipeline, the vision-to-language projection parameters are frozen, whereas in the MCQ pipeline, these projection layers remain trainable.

The MCQ model is trained on EXAMS-V[3], where each training example consists of the original image containing the question and answer options, followed by a multiple-choice instruction prompt. The OpenQA model is trained on ImageCLEF-MR2026-OpenQA-Visual[4], where each example consists of the original image followed by an instruction prompt containing the extracted question text. The target response is formatted as `FINAL ANSWER: <LETTER>` for MCQ and `FINAL ANSWER: <ANSWER>` for OpenQA. During supervised fine-tuning, the prompt tokens are masked so that the loss is computed only over the assistant's response.

5. Overview of the Systems and Results

5.1. Competition Results

We report only the test leaderboards because participation during the development phase was very limited. Tables 5 and 6 present the final results for the four competition tracks. The Visual MCQ track attracted the largest participation, with nine teams submitting systems. The Textual MCQ track received three submissions, while the Visual and Textual OpenQA tracks attracted eight and two participating teams, respectively.

The Visual MCQ track achieved the highest overall performance among the four tracks. Team **spirosbax** ranked first with an overall accuracy of 84.06%, followed by **DS@GT** with 79.86%. Both systems consistently achieved strong results across all six evaluation languages, with the highest scores generally observed for Italian, Croatian, and Bulgarian. In the Textual MCQ track, **pratikpriyanshu** obtained the highest overall accuracy of 75.38%, followed by **linxiaocan** with 70.84%.

The OpenQA tracks were more challenging than the MCQ tracks, as systems were required to generate complete answers instead of selecting from predefined options. In the Visual OpenQA track, **FAU** achieved the highest overall COMET score of 0.6488, narrowly outperforming **wangshou66** (0.6366). For the Textual OpenQA track, **linxiaocan** ranked first with a COMET score of 0.6563, ahead of **pratikpriyanshu** (0.5285).

Across both MCQ and OpenQA tracks, performance varied across languages. Italian and English generally achieved the highest scores in the MCQ tracks, whereas Chinese consistently obtained the highest COMET scores in the OpenQA tracks. Bulgarian and Serbian remained comparatively more challenging across most submissions, indicating that multilingual reasoning and generation continue to present difficulties for current models.

5.2. Detailed System Descriptions

Table 7 presents an overview of the approaches used by the participants. In the rest of this section, we present brief descriptions of the participant systems, the leaderboards in which they participated, and relevant keywords. Teams are ordered alphabetically

[2] https://huggingface.co/Qwen/Qwen3.5-9B
[3] https://huggingface.co/datasets/MBZUAI/EXAMS-V
[4] https://huggingface.co/datasets/SU-FMI-AI/ImageCLEF-MR2026-OpenQA-Visual

Table 5
Results for the ImageCLEF 2026 MCQ tracks. Rankings are based on overall accuracy.

Visual MCQ								
Rank	Team	Overall	EN	BG	ZH	HR	IT	SR
1	spirosbax	**0.8406**	0.8560	0.8909	0.7807	0.8772	0.9074	0.8704
2	DS@GT	0.7986	0.8520	0.8636	0.6725	0.8596	0.8704	0.8333
3	FAU	0.7108	0.7480	0.6909	0.6345	0.7368	0.8148	0.7593
4	*baseline*	0.6437	0.7440	0.3000	0.5936	0.6316	0.6852	0.7037
5	zhaijinghe	0.6150	0.6760	0.6182	0.5351	0.6140	0.5926	0.5741
6	sjaini	0.5927	0.6920	0.5091	0.4415	0.5965	0.7222	0.6667
7	begyed	0.5774	0.6000	0.6636	0.4708	0.6667	0.7037	0.6481
8	linxiaocan	0.5560	0.6040	0.5455	0.5117	0.5088	0.5926	0.4259
9	pratikpriyanshu	0.5076	0.5720	0.4364	0.4181	0.5614	0.5556	0.5185
10	wether	0.4673	0.5460	0.5636	0.3187	0.4561	0.4815	0.4815
Textual MCQ								
Rank	Team	Overall	EN	BG	ZH	HR	IT	SR
1	pratikpriyanshu	**0.7538**	0.8375	0.6763	0.6479	0.8030	0.8497	0.8075
2	linxiaocan	0.7084	0.7917	0.6098	0.6891	0.7552	0.7630	0.7326
3	*baseline*	0.6177	0.8292	0.1929	0.7116	0.7821	0.7861	0.7861
4	rafiulbiswas	0.6152	0.7125	0.4346	0.6629	0.6836	0.7341	0.6257

Table 6
Results for the ImageCLEF 2026 OpenQA tracks. Rankings are based on the overall COMET score. BLEU, ROUGE-L, and METEOR are reported as additional evaluation metrics.

Visual OpenQA											
#	Team	COMET	COMET per language						Additional metrics		
			BG	ZH	HR	EN	IT	SR	BLEU	ROUGE-L	METEOR
1	FAU	**0.6488**	0.6493	0.7287	0.6588	0.6499	0.6132	0.5838	0.1391	0.2762	0.2383
2	wangshou66	0.6366	0.6156	0.7012	0.6597	0.6519	0.5920	0.5929	0.1308	0.2717	0.2388
3	uned-martinez	0.5938	0.5721	0.5985	0.6390	0.5942	0.5767	0.5804	0.0980	0.2452	0.1842
4	liuzhe3085	0.5829	0.4947	0.6705	0.6612	0.6180	0.5317	0.5110	0.1556	0.3351	0.2295
5	*baseline*	0.5613	0.5585	0.4873	0.6046	0.6228	0.5423	0.5507	0.1078	0.2272	0.1777
6	wenlong	0.5497	0.5407	0.5758	0.5854	0.5311	0.5175	0.5470	0.0680	0.1713	0.1185
7	rafiulbiswas	0.5143	0.4798	0.4848	0.6167	0.4851	0.5031	0.5167	0.0735	0.2013	0.1331
8	linxiaocan	0.5117	0.4666	0.5150	0.5780	0.5038	0.5053	0.4998	0.0541	0.1434	0.0936
9	pratikpriyanshu	0.4286	0.4007	0.3688	0.4661	0.4626	0.4484	0.4248	0.0375	0.0993	0.0712
Textual OpenQA											
#	Team	COMET	COMET per language						Additional metrics		
			BG	ZH	HR	EN	IT	SR	BLEU	ROUGE-L	METEOR
1	linxiaocan	**0.6563**	0.5634	0.7539	0.6559	0.6723	0.6703	0.6557	0.1871	0.3032	0.2139
2	*baseline*	0.6123	0.5837	0.5278	0.6694	0.6227	0.6363	0.6540	0.1764	0.3050	0.2102
3	pratikpriyanshu	0.5285	0.4309	0.4969	0.5553	0.5088	0.6097	0.5584	0.1142	0.2300	0.1591

DS@GT [32]

Tasks: *Visual MCQ*
Keywords: *Qwen3.5-9B, Gemma 4 E2B, GRPO, LoRA, Supervised Fine-Tuning, Distillation*
Description: The authors investigate training-time adaptation of open-source vision-language models for multilingual Visual MCQ reasoning. Their primary system applies Group Relative Policy Opti-

Table 7
Overview of the approaches used by the participating systems. "Size" indicates the competition track in which the submitted model belongs: *Tiny* ($\leq$7B parameters) or *Normal* ($\geq$8B parameters). "Models" lists the primary vision-language backbones used by each team. "Prompting" summarizes the main prompting and reasoning strategies. "Adaptation" indicates whether the system relies on zero-shot inference, model fine-tuning, or reinforcement learning. "Data" denotes the use of data augmentation. "Misc." includes additional techniques such as LoRA, OCR, model ensembling, and memory optimization. A square (☑) indicates that the corresponding technique was employed by the system.

Team	Size		Models					Prompting			Adaptation			Data	Misc.			
	Tiny	Normal	Qwen2-VL	Qwen2.5-VL	Qwen3-VL	Qwen3.5	Gemma	CoT	Self-Consistency	QRA	Zero-shot	Fine-tuning	RL	Augmentation	LoRA	OCR	Ensemble	Memory Opt.
MarsadLab [31]		☑		☑	☑			☑	☑			☑			☑			
DS@GT [32]	☑					☑	☑					☑	☑		☑	☑		
FAU [33]		☑		☑	☑						☑				☑	☑	☑	
linxiaocan [34]		☑			☑			☑			☑	☑	☑		☑	☑		
liuzhe3085 [35]	☑		☑					☑		☑	☑							☑
Nika [36]	☑			☑		☑		☑	☑		☑							
SRH-ReliableAI [37]		☑			☑			☑	☑			☑			☑			
UNED-Martinez [38]	☑			☑				☑			☑							
wenlong [39]	☑			☑								☑		☑	☑			
zhai [40]	☑				☑			☑	☑		☑					☑		

mization (GRPO) to Qwen3.5-9B, exploring several reward formulations and subject-specific training configurations that encourage correct predictions, valid answer formatting, and shorter reasoning traces. The resulting model achieves second place on the Visual MCQ leaderboard with an overall accuracy of 79.86%. In addition, they study a smaller Gemma 4 E2B model adapted with LoRA-based supervised fine-tuning on reasoning traces distilled from Gemma 4 31B, and perform ablation studies on prompt design, decoding budget, and OCR augmentation. Their experiments show that GRPO consistently improves answer validity and reduces excessive reasoning length, while supervised fine-tuning is essential for improving smaller models and OCR provides mixed benefits depending on the evaluation setting.

FAU [33]

Tasks: *Visual MCQ, Visual OpenQA*
Keywords: *Candidate Label Scoring, Next-Token Logit Scoring, Weighted Score Fusion, Output Control, Concise OpenQA Prompting, Ensembles, OCR, LoRA*
Description: The authors propose an inference-engineering approach focused on controlling model outputs rather than fine-tuning a single model. For Visual MCQ, they score the candidate directly from next-token logits, producing reusable score vectors that can be merged across models through weighted score fusion and voting. For Visual OpenQA, they use enhanced images, concise no-reasoning prompts, deterministic decoding, and answer cleanup to remove reasoning traces, XML-like tags, answer prefixes, and formatting artifacts. The cleaned answers from several Qwen-family models are then combined through COMET-weighted answer-level ensembling, including position-weighted clustering and union-based variants. The paper also reports important negative findings: LoRA/QLoRA fine-tuning on limited heterogeneous data and raw OCR prompt injection both reduced performance.

linxiaocan [34]

Tasks: *Visual MCQ, Textual MCQ, Visual OpenQA, Textual OpenQA*
Keywords: *Qwen3-VL-8B, Extract-Verify-Answer, QLoRA, Text Verification, Reinforcement Learning, Agent-Based Pipeline*
Description: The authors investigate several approaches for multilingual exam question answering, including zero-shot prompting, OCR-assisted prompting, QLoRA fine-tuning, reinforcement learning, agent-based reasoning, and a multi-stage extract-verify-answer (EVA) pipeline. Their main system decomposes the MCQ task into three stages: extracting structured evidence from the exam image, verifying the extracted text against the original image to correct OCR errors and preserve option order, and answering using both the image and the cleaned context. This structured prompting strategy improves MCQ performance over the zero-shot baseline and is used for their official MCQ submissions, where the team ranks second on the Textual MCQ leaderboard with 0.7084 accuracy and seventh on the Visual MCQ leaderboard with 0.5560 accuracy. For OpenQA, the authors find QLoRA fine-tuning more effective than prompt-only methods, especially for short-answer generation, and submit a full-data fine-tuned model. This system ranks first on the Textual OpenQA leaderboard with a COMET score of 0.6563 and seventh on the Visual OpenQA leaderboard with a COMET score of 0.5117. Their analysis shows that fine-tuning helps OpenQA more than MCQ, while visual MCQ remains limited by diagram understanding, OCR noise, option-order errors, and answer instability.

liuzhe3085 [35]

Tasks: *Visual OpenQA*
Keywords: *Qwen2-VL-7B, Question Reconstruction before Answering (QRA), Cross-lingual Prompting, High-Resolution Processing, Memory Optimization*
Description: The authors propose a lightweight Visual OpenQA system based on Qwen2-VL-7B that combines a two-stage Question Reconstruction before Answering (QRA) prompting strategy with hardware-aware memory optimization. The QRA framework first reconstructs the missing textual context from the input image before performing cross-lingual chain-of-thought reasoning, reducing the modality gap between visual perception and language generation. To improve recognition of fine-grained visual content under the task's hardware constraints, the authors employ dynamic high-resolution image processing together with GPU memory optimization, enabling stable inference without external OCR or additional fine-tuning. Their ablation study shows that both high-resolution processing and the QRA prompting strategy contribute substantially to performance, resulting in a fourth-place ranking on the Visual OpenQA leaderboard with an overall COMET score of 0.5829.

MarsadLab [31]

Tasks: *Textual MCQ, Visual OpenQA*
Keywords: *Qwen2.5-VL-7B, Qwen3-VL-8B, LoRA, QLoRA, Letter-Logit Scoring, Native-Language Prompting, Self-Consistency*
Description: The authors participate in the Textual MCQ and Visual OpenQA subtasks. For Textual MCQ, they fine-tune Qwen2.5-VL-7B-Instruct with 4-bit NF4 quantization and LoRA adapters on a language-filtered subset of EXAMS-V containing text-only questions in the six target languages. To handle language imbalance, they cap the number of training examples per language and train the model to predict only the final answer letter. At inference time, they avoid free-form generation and instead use a single forward pass to score the logits of candidate answer letters A–D, aggregating over common token variants. This makes the system faster and prevents invalid answer formats, leading to a third-place ranking on the Textual MCQ leaderboard with 0.6152 overall accuracy. For Visual OpenQA, they use Qwen3-VL-8B-Instruct through vLLM with native-language prompts and an explicit tagged answer format. They compare zero-shot chain-of-thought prompting, five-sample self-consistency, and subject-specialized prompts, and submit the self-consistency setting for the final run. Their answer extraction combines answer-tag parsing, multilingual final-answer patterns, and a last-line fallback,

followed by normalization for voting. The system ranks sixth on the Visual OpenQA leaderboard with an overall COMET score of 0.5143, with Bulgarian reported as the weakest language in both subtasks.

Nika [36]

Tasks: *Visual MCQ*
Keywords: *Qwen3.5-4B, Qwen2.5-VL-7B, Test-Time Scaling, Self-Consistency, Process Reward Models, Beam Search*
Description: The authors investigate test-time scaling strategies for multilingual Visual MCQ reasoning using small open-source vision-language models. They compare self-consistency, describe-then-reason with PRM-guided beam search, and post-hoc answer selection using both generative critics and process reward models. Their experiments show that increasing the decoding budget and improving answer parseability contribute substantially more than increasing the number of sampled reasoning chains or using more sophisticated search and verification methods. The final system employs Qwen3.5-4B with self-consistency, an increased decoding budget, and guided answer repair, achieving first place on the Visual MCQ leaderboard with an overall accuracy of 84.06%. The authors further report that neither PRM-guided beam search nor trained reward models consistently outperform simple majority voting, highlighting the importance of strong base models and efficient inference over more complex test-time reasoning strategies.

SRH-ReliableAI [37]

Tasks: *Textual MCQ, Visual MCQ, Textual OpenQA, Visual OpenQA*
Keywords: *Qwen3-VL-8B-Thinking, QLoRA Fine-Tuning, Thinking Tokens, Subject-Aware Prompting, Language-Matched Instructions, Multi-Stage Answer Extraction, Self-Consistency Voting*)
Description: The authors propose a fine-tuning-based pipeline built around Qwen3-VL-8B-Thinking, a vision-language model with built-in chain-of-thought reasoning through explicit thinking tokens. The core system applies QLoRA fine-tuning on 16,300 EXAMS-V multiple-choice samples, using 4-bit quantization and LoRA adapters across the model's linear layers to adapt the model efficiently for multilingual exam question answering. At inference time, the authors combine this fine-tuned thinking model with subject-aware prompting, which injects domain context from the question metadata, and language-matched answer instructions, which encourage the model to answer consistently in the language of the input question. The paper also reports several useful design observations beyond the final system: direct, thinking, and decomposed prompts were compared; token budget was shown to be critical for thinking models.

UNED-Martinez [38]

Tasks: *OpenQA*
Keywords: *Qwen2.5-VL-7B, Vision-Language Models, Hybrid Per-Language Prompting, Zero-Shot Inference, Chain-of-Thought prompting*
Description: The authors propose a zero-shot multilingual Visual OpenQA pipeline based on Qwen2.5-VL-7B-Instruct. Instead of using a single prompt across all languages, they introduce a per-language prompting strategy, where English and Bulgarian use concise English prompts, while Chinese, Croatian, Italian, and Serbian use prompts written in the corresponding native language. A normalization and validation pipeline is applied to remove answer prefixes, reasoning traces, markdown artifacts, and invalid outputs before submission. Their ablation study shows that this asymmetric prompting strategy outperforms uniform English prompting and chain-of-thought variants, leading the system to rank third on the Visual OpenQA final-phase leaderboard.

wenlong [39]

Tasks: *OpenQA*
Keywords: *Qwen2.5-VL-7B, GPT-Assisted Multilingual Data Augmentation, Fine-Tuning, Visual Encoder Tuning, LoRA*
Description: The authors propose a data-centric adaptation pipeline for multilingual Chart-VQA using Qwen2.5-VL-7B. Their main idea is to address the scarcity of task-specific training data by generating multilingual question–answer pairs for each chart image using GPT-assisted augmentation, followed by human verification to preserve factual correctness and semantic fidelity. The augmented data is then used to fine-tune the model with a hybrid strategy: the vision encoder and projection layers are fully fine-tuned to better capture chart-specific elements such as axes, legends, labels, and data markers, while LoRA is applied to the language model component to preserve general linguistic ability with fewer trainable parameters. This combination of curated multilingual expansion and targeted visual adaptation leads to consistent improvements, ranking fifth on the Visual OpenQA leaderboard.

zhai [40]

Tasks: *MCQ*
Keywords: *Qwen3-VL-4B-Thinking, Structured CoT, Self-Consistency Voting, Zero-Finetuning, Multi-Stage Answer Extraction, OCR-then-Reason, Few-Shot*
Description: The authors explore how far a single small open-weight thinking VLM can be pushed on multilingual Visual MCQ reasoning under a without fine-tuning. Their final system uses Qwen3-VL-4B-Thinking as the only inference backbone. The authors apply five-sample self-consistency voting at low temperature, with deterministic tie-breaking based on a hash of the image path. They also design a multi-stage answer extractor that removes the thinking block, maps Cyrillic option labels to Latin A–E, applies a regex cascade, and falls back to a reverse scan when needed. Beyond the final system, the paper reports several useful negative results: language-aware prompting, higher image resolution, few-shot prompting, OCR-then-reason, GRPO, and LoRA SFT did not provide stable gains in their screening experiments. The submitted system uses Structured CoT, self-consistency with five samples, 1024-pixel image preprocessing, and deterministic answer extraction, reaching fourth place with 0.6150 accuracy on the official test set.

6. Conclusion and Future Work

We presented the ImageCLEF 2026 Multimodal Reasoning task, a shared benchmark for evaluating multilingual exam-style question answering with vision–language models. The task extends the previous ImageCLEF multimodal reasoning setting by covering both multiple-choice and open-ended question answering, with separate tracks for visual and textual questions. The dataset contains 5,026 questions across six languages and includes diagrams, graphs, tables, figures, and text-only exam items rendered as images. The results show that current systems perform better in multiple-choice settings than in open-ended generation. The strongest submissions reached 84.06% accuracy on Visual MCQ and 75.38% accuracy on Textual MCQ, while the best OpenQA systems achieved COMET scores of 0.6488 on Visual OpenQA and 0.6563 on Textual OpenQA. Participant systems used a range of strategies, including prompting, LoRA/QLoRA fine-tuning, self-consistency, test-time scaling, answer extraction, and multi-stage reasoning pipelines. Across tracks, performance varied substantially by language and answer format, indicating that multilingual visual reasoning and direct answer generation remain challenging.

Future work will focus on expanding the benchmark to more languages, subjects, and visual formats, especially for underrepresented scripts and educational contexts. We also plan to improve open-ended evaluation through stronger normalization and semantic matching, and to provide more detailed diagnostic analyses by language, subject, grade level, answer type, and visual element category.

Acknowledgments

We are thankful to all the people that took part in ensuring the quality of the constructed dataset and namely: Kristian Nikolov and Owais Aijaz.

The work of Dimitar Dimitrov, and Ivan Koychev is partially supported by the project UNITe BG16RFPR002-1.014-0004 funded by PRIDST.

The work of Dimitrina Zlatkova and Georgi Pachov is partially supported by 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

During the preparation of this work, the author(s) used Generative AI in order to: Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, D. Parikh, VQA: visual question answering, in: 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] 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. Mo-

hamed, 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.

[9] 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.

[10] 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.

[11] 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.

[12] P. Lu, H. Bansal, T. Xia, J. Liu, C. Li, H. Hajishirzi, H. Cheng, K.-W. Chang, M. Galley, J. Gao, Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts, in: The Twelfth International Conference on Learning Representations, 2024. URL: https://openreview.net/forum?id=KUNzEQMWU7.

[13] 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: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9556–9567. doi:10.1109/CVPR52733.2024.00913.

[14] 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.

[15] W. Zhang, S. M. Aljunied, C. Gao, Y. K. Chia, L. Bing, M3exam: a multilingual, multimodal, multilevel benchmark for examining large language models, in: Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23, Curran Associates Inc., Red Hook, NY, USA, 2023.

[16] I. Salazar, M. F. Burda, S. B. Islam, A. S. Moakhar, S. Singh, F. Farestam, A. Romanou, D. Boiko, D. Khullar, M. Zhang, D. Krzemiński, J. Novikova, L. Shimabucoro, J. M. Imperial, R. Maheshwary, S. Duwal, A. Amayuelas, S. Rajwal, J. Purbey, A. Ruby, N. Popovič, M. Suppa, A. T. Wasi, R. M. R. Kadiyala, O. Tsymboi, M. Kostritsya, B. soltani moakhar, G. da Costa Merlin, O. F. Coletti, M. Jabbarishiviari, M. FARAHANIFARD, S. A. Fernandez, M. Grandury, D. Abulkhanov, D. Sharma, A. G. D. Mitri, L. B. Marchezi, S. Heydari, J. Obando-Ceron, N. Kohut, B. Ermis, D. Elliott, E. Ferrante, S. Hooker, M. Fadaee, Kaleidoscope: In-language exams for massively multilingual vision evaluation, in: The Fourteenth International Conference on Learning Representations, 2026. URL: https://openreview.net/forum?id=zCYXhSy9UH.

[17] 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.
[18] D. I. Dimitrov, M. Hee, Z. Xie, R. J. Das, M. Ahsan, S. Ahmad, N. Paev, I. K. Koychev, P. I. Nakov, Overview of imageclef 2025–multimodal reasoning (2025).
[19] A. Singh, V. Natarjan, M. Shah, Y. Jiang, X. Chen, D. Parikh, M. Rohrbach, Towards vqa models that can read, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 8317–8326.
[20] A. Mishra, S. Shekhar, A. K. Singh, A. Chakraborty, OCR-VQA: visual question answering by reading text in images, in: 2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, September 20-25, 2019, IEEE, 2019, pp. 947–952. URL: https://doi.org/10.1109/ICDAR.2019.00156. doi:10.1109/ICDAR.2019.00156.
[21] M. Mathew, V. Bagal, R. Tito, D. Karatzas, E. Valveny, C. V. Jawahar, Infographicvqa, in: IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, Waikoloa, HI, USA, January 3-8, 2022, IEEE, 2022, pp. 2582–2591. URL: https://doi.org/10.1109/WACV51458.2022.00264. doi:10.1109/WACV51458.2022.00264.
[22] A. Masry, D. X. Long, J. Q. Tan, S. R. Joty, E. Hoque, Chartqa: A benchmark for question answering about charts with visual and logical reasoning, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022, Findings of ACL, Association for Computational Linguistics, 2022, pp. 2263–2279. URL: https://doi.org/10.18653/v1/2022.findings-acl.177. doi:10.18653/V1/2022.FINDINGS-ACL.177.
[23] S. Singh, P. Chaurasia, Y. Varun, P. Pandya, V. Gupta, V. Gupta, D. Roth, Flowvqa: Mapping multimodal logic in visual question answering with flowcharts, in: L. Ku, A. Martins, V. Srikumar (Eds.), Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024, Findings of ACL, Association for Computational Linguistics, 2024, pp. 1330–1350. URL: https://doi.org/10.18653/v1/2024.findings-acl.78. doi:10.18653/V1/2024.FINDINGS-ACL.78.
[24] R. Rivest, The MD5 message-digest algorithm, Technical Report, 1992.
[25] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE transactions on image processing 13 (2004) 600–612.
[26] C. Zauner, Implementation and benchmarking of perceptual image hash functions (2010).
[27] R. Rei, C. Stewart, A. C. Farinha, A. Lavie, COMET: A neural framework for MT evaluation, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 2685–2702. URL: https://aclanthology.org/2020.emnlp-main.213/. doi:10.18653/v1/2020.emnlp-main.213.
[28] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: P. Isabelle, E. Charniak, D. Lin (Eds.), Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 2002, pp. 311–318. URL: https://aclanthology.org/P02-1040/. doi:10.3115/1073083.1073135.
[29] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013/.
[30] S. Banerjee, A. Lavie, METEOR: An automatic metric for MT evaluation with improved correlation with human judgments, in: J. Goldstein, A. Lavie, C.-Y. Lin, C. Voss (Eds.), Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Association for Computational Linguistics, Ann Arbor, Michigan, 2005, pp. 65–72. URL: https://aclanthology.org/W05-0909/.
[31] M. R. Biswas, W. Zaghouani, MarsadLab at ImageCLEF 2026 Multimodal Reasoning: LoRA Fine-Tuned Vision Language Models for Multilingual Textual MCQ and Visual Open Question Answering, in: [41], 2026.
[32] Z. Liu, C. Zhang, Y. Li, DS@GT at ImageCLEF 2026 MultimodalReasoning: Visual Multiple Choice Question Reasoning with Vision-Language Models, in: [41], 2026.

[33] M. Elmashtooly, V. Christlein, FAU at ImageCLEF 2026 Task on Multimodal Reasoning: Robust Candidate Scoring and Concise Multilingual Visual Answering, in: [41], 2026.
[34] X. Lin, X. Qin, Z. Han, A Multi-Stage Extract-Verify-Answer Pipeline for Multilingual Exam Question Answering at ImageCLEF 2026, in: [41], 2026.
[35] H. Liu, G. Niu, S. Wang, MarsadLab at ImageCLEF 2026 Multimodal Reasoning: LoRA Fine-Tuned Vision Language Models for Multilingual Textual MCQ and Visual Open Question Answering, in: [41], 2026.
[36] P.-J. Yang, S. Baxevanakis, Nika at ImageCLEF 2026 Task on Multimodal Reasoning: More Tokens, Fewer Trees — Test-Time Scaling for Small VLMs on Multilingual Visual MCQ, in: [41], 2026.
[37] P. Priyanshu, S. Chandna, SRH-ReliableAI at ImageCLEF 2026 Task on Multimodal Reasoning: Fine-Tuning Thinking Vision-Language Models for Multilingual Exam Question Answering, in: [41], 2026.
[38] F. Martínez Marco, UNED-Martinez at ImageCLEF 2026 Task on Multimodal Reasoning: Hybrid Per-Language Prompting with Qwen2.5-VL-7B for Multilingual Visual OpenQA, in: [41], 2026.
[39] W. Li, LoRA-based Fine-tuning of Qwen Model for Visual Question Answering, in: [41], 2026.
[40] J. Zhai, H. Chen, X. Qiu, zhai at ImageCLEF 2026 Task on Multimodal Reasoning: Prompt Design and Self-Consistency on a 4 B Open-Weight Thinking VLM, in: [41], 2026.
[41] E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), Working Notes of CLEF 2026 - Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS.org, 2026.