A Multi-Stage Extract-Verify-Answer Pipeline for Multilingual Exam Question Answering at ImageCLEF 2026

Xiaocan Lin[1], Xueyang Qin[1] and Zhongyuan Han[1,*]

[1]Foshan University, Foshan, China

Abstract

This paper presents our experimental investigation into multimodal reasoning for multilingual exam question answering, conducted in the context of the ImageCLEF 2026 MultimodalReasoning competition [1], which is part of the broader ImageCLEF 2026 campaign [2]. We systematically evaluate a range of approaches on the EXAMS-V dataset [3], including zero-shot prompting, QLoRA fine-tuning [4], reinforcement learning (REINFORCE), agent-based pipelines, and a task-oriented multi-stage extract-verify-answer (EVA) pipeline. Our key finding is that structured multi-stage prompting — which decomposes the MCQ task into extraction, verification, and answering phases — yields a significant accuracy improvement from 49.40% to 56.47% on the test2025 benchmark, outperforming both zero-shot baselines and supervised QLoRA fine-tuning. For open-ended question answering (OpenQA), QLoRA fine-tuning proves effective, improving COMET [5] from 0.6718 to 0.7042 on textual OpenQA and achieving METEOR [6] of 0.8056 on visual OpenQA. We conduct detailed error analysis identifying four failure categories, and perform prompt ablation studies showing that subject-specific role prompts provide marginal gains while chain-of-thought reasoning degrades performance in the current pipeline configuration.

Keywords

MultimodalReasoning, VLMs, exam QA, QLoRA, ImageCLEF, EXAMS-V

1. Introduction

The ImageCLEF 2026 MultimodalReasoning task evaluates whether a vision-language model can answer multilingual exam questions from either images or text [1]. The benchmark covers four subtasks: visual multiple-choice question answering (visual MCQ), textual MCQ, visual open question answering (visual OpenQA), and textual OpenQA. Questions span more than ten languages and multiple school subjects, which makes the task a realistic testbed for multilingual perception, reasoning, and answer generation within the broader ImageCLEF 2026 campaign [2].

A central difficulty in this task is that direct end-to-end prompting often entangles visual extraction errors with downstream reasoning. When the model misses formulas, hallucinates text, or accidentally changes option order during inference, the final answer can be wrong even if the underlying reasoning ability is strong. This problem is particularly severe for multilingual exam images that mix dense text, diagrams, tables, and scientific notation.

To reduce this error propagation, we study a multi-stage extract-verify-answer pipeline that explicitly separates information extraction, text verification, and final answer prediction. The same structured backbone is used for multilingual MCQ and adapted to OpenQA analysis, while additional experiments compare zero-shot prompting, OCR-assisted prompting, QLoRA fine-tuning, reinforcement learning, and an agent-based variant. Our results show that the proposed decomposition is the most effective strategy for visual MCQ, raising accuracy from 49.40% to 56.47% on the test2025 benchmark.

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

*Corresponding author.

linxx6513@gmail.com (X. Lin); qinxueyang@snu.edu.cn (X. Qin); hanzhongyuan@gmail.com (Z. Han)

0000-0001-8960-9872 (Z. Han)

2. Related Work

2.1. Vision-Language Models for Document Understanding

Recent advances in VLMs, particularly the Qwen-VL family (Qwen2.5-VL, Qwen3-VL), have demonstrated strong capabilities in document and image understanding [7]. These models process interleaved image-text inputs and can perform visual question answering, OCR, and reasoning over charts and diagrams. Our work builds on Qwen3-VL-8B-Instruct-FP8 as the primary backbone.

2.2. Parameter-Efficient Fine-Tuning

QLoRA (Quantized Low-Rank Adaptation) enables fine-tuning large models by quantizing the base model to 4-bit precision and training only low-rank adapter parameters [4]. Prior work has shown QLoRA to be effective for domain adaptation with minimal compute. We apply QLoRA with rank-16 adapters targeting all attention and MLP projection layers, training approximately 43.6M parameters out of 8.8B total (~0.5%).

2.3. Multi-Stage Reasoning Pipelines

Decomposing complex tasks into sequential sub-tasks has proven effective in NLP (e.g., chain-of-thought prompting [8], self-consistency [9]). For visual QA and document QA, extract-then-reason pipelines first convert images or layouts into textual or structured evidence before the final reasoning step. This design is attractive for exam QA because it exposes intermediate evidence, but it also risks propagating extraction errors into the final answer. Recent ImageCLEF-style systems have investigated ensemble VLMs, multi-prompt aggregation, and other practical strategies for multilingual multimodal reasoning [10, 11, 12]. Our EVA pipeline follows this general extract-then-reason family and evaluates whether an explicit verification pass can make the intermediate evidence more reliable in the ImageCLEF setting.

2.4. Verification and Self-Correction

Verification and self-correction methods ask a model to revisit an intermediate or final output before accepting it. Self-consistency samples multiple reasoning paths [9], while Self-Refine uses iterative feedback and refinement without additional training data [13]. Our verification stage is related to this test-time correction paradigm, but it is narrower: it checks extracted question evidence against the original image and tries to preserve option labels and ordering. We therefore treat verification as a consistency-control component rather than as an independently large methodological contribution.

2.5. Reinforcement Learning for Language Models

REINFORCE and PPO-based approaches have been applied to align language models with reward signals. In the context of exam QA, accuracy on labeled questions provides a natural reward signal. We explore REINFORCE training starting from supervised fine-tuned adapters.

2.6. Agent-Based Reasoning

Tool-augmented agents that can interact with external tools (e.g., image zoom, OCR) have shown promise for complex reasoning tasks. ReAct interleaves reasoning traces with external actions [14], and Toolformer studies how language models can learn to use external tools [15]. We implement an ExamAgent using the Qwen-Agent framework with specialized tools for information extraction and answer verification, but we report it as an exploratory participant-system variant rather than as the main submission method.

3. Method

Figure 1 presents the overall design of our method. Instead of asking the model to directly answer from a noisy exam image, we first convert the image into structured evidence, then verify that evidence against the source image, and finally predict the answer from both modalities. This design is intended to reduce error accumulation between perception and reasoning stages.

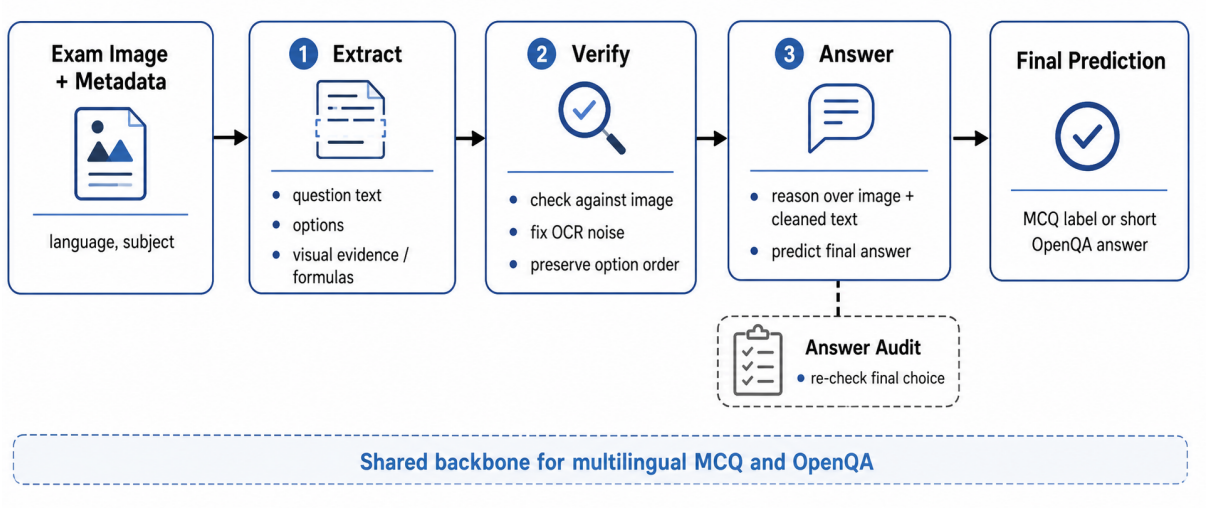

Figure 1: Overview of the proposed multi-stage extract-verify-answer pipeline. The same backbone is used for multilingual MCQ and adapted to OpenQA analysis, while an optional answer-audit stage re-checks the final decision.

3.1. Task Formulation

Let $x = (I, m)$ denote an input instance, where I is an exam image or a text-rendered exam item and m contains metadata such as language and subject. For MCQ, the model predicts an answer label $y \in \{A, B, C, D, E\}$, where the valid set depends on the question. For OpenQA, the model predicts a short free-form answer sequence $y = (w_1, \ldots, w_T)$. The objective is to learn or prompt a system that maps multilingual visual-textual evidence to the correct label or answer while remaining robust to OCR noise, dense layouts, and scientific notation.

3.2. Multi-Stage Extract-Verify-Answer Pipeline

Our main method decomposes the task into three core stages plus an optional audit stage.

Stage 1: Extract. Given the original exam image and task metadata, the model first produces a structured textual representation with four fields: `QUESTION`, `OPTIONS`, `VISUAL_EVIDENCE`, and `NUMBERS_AND_FORMULAS`. This stage externalizes the information that is usually buried inside a single end-to-end prompt and preserves the original question language.

Stage 2: Verify. The extracted text is then compared against the original image. The model is instructed to treat the image as the source of truth, remove hallucinated content, recover missing details, and preserve the original option labels and order. This stage is critical because many MCQ failures come from noisy extraction or destructive edits introduced during reasoning.

Stage 3: Answer. The final prediction stage consumes both the original image and the verified text. For MCQ, the model outputs a single answer option. For OpenQA-oriented analysis, the same backbone can be used to produce a short answer after the evidence has been cleaned and organized.

Optional Stage 4: Answer Audit. For difficult MCQ cases, we optionally add a final audit pass that re-checks the proposed answer against the source image and the verified context. This stage offers only marginal gains, but it is useful for understanding whether the remaining errors come from evidence extraction or from final answer instability.

4. Experiments

4.1. Experimental Setup

All experiments use Qwen3-VL-8B-Instruct-FP8 as the primary backbone and are conducted on a multi-GPU NVIDIA environment. We compare prompt-based, fine-tuning-based, reinforcement learning, and agent-based variants under a shared evaluation protocol.

4.1.1. Experimental Data

We use the MBZUAI/EXAMS-V dataset [3], a multilingual collection of exam questions designed for visual reasoning evaluation. The dataset covers English, Chinese, Arabic, Bulgarian, Croatian, Hungarian, Italian, Kazakh, Polish, Serbian, Spanish, Urdu, and additional languages, with subjects including Physics, Mathematics, Chemistry, Biology, Informatics/Computer Science, Art, and related domains.

The competition defines four subtasks, summarized in Table 2. Due to hidden official dev/test labels, OpenQA local experiments use a fixed train-holdout split (seed=42, 20% validation, language-stratified). MCQ experiments use the 2025 EXAMS-V MCQ train/test2025 protocol for local comparison. Table 1 therefore separates the data source and purpose of each track.

Table 1
Data sources and local evaluation protocols used in our experiments.

Track	Data source/protocol	Train	Holdout/Test	Use
visual_openqa	ImageCLEF 2026 OpenQA official train, local 80/20 split	423	105	Strict local validation
textual_openqa	ImageCLEF 2026 OpenQA official train, local 80/20 split	240	60	Strict local validation
MCQ	EXAMS-V 2025 train/test2025 protocol	16,281	3,565	Local MCQ comparison

The 240/60 textual OpenQA and 423/105 visual OpenQA splits are internal 80/20 splits of the official ImageCLEF 2026 training data, not official hidden dev/test splits. The MCQ row refers to the 2025 EXAMS-V MCQ train/test2025 setting used for local controlled comparison. Official ImageCLEF 2026 leaderboard results are reported separately in Section 4.6.

Table 2
Tasks and primary metrics.

Task	Type	Output	Primary Metric
visual_mcq	Multiple-choice	A/B/C/D/E	Accuracy
textual_mcq	Multiple-choice	A/B/C/D/E	Accuracy
visual_openqa	Open-ended	Text answer	BLEU-Avg, ROUGE-L, METEOR, COMET
textual_openqa	Open-ended	Text answer	BLEU-Avg, ROUGE-L, METEOR, COMET

4.1.2. Data Processing

We developed a deduplication tool to remove exact duplicates, cropped sub-images, and perceptually similar images using MD5 hashing, template matching with SSIM, and rotation-aware perceptual hashing (pHash). For OCR-related ablations, OCR outputs are treated as auxiliary evidence rather than a mandatory preprocessing step, since preliminary probes show that OCR can increase runtime substantially and is not consistently helpful for MCQ.

4.1.3. Evaluation Metrics

MCQ performance is measured by exact-match accuracy between the predicted answer label and the gold answer key. For OpenQA, we report BLEU-Avg, ROUGE-L, METEOR, and COMET [16, 17, 6, 5]. BLEU-Avg averages sentence-level BLEU-1 to BLEU-4, ROUGE-L measures longest-common-subsequence overlap, METEOR captures lexical alignment with stemming and synonymy, and COMET provides a semantic similarity estimate using the Unbabel/wmt22-comet-da model.

4.1.4. Baseline Methods

We compare five families of methods: (1) direct zero-shot prompting, (2) OCR-assisted prompting, (3) QLoRA fine-tuning [4], (4) reinforcement learning starting from supervised adapters, and (5) an agent-based pipeline with external tools. The proposed extract-verify-answer pipeline is evaluated against all of these baselines.

4.1.5. Reproducibility Details

All local split construction uses seed 42 and language-stratified sampling where labels are available. Prompt-based MCQ runs constrain the final output to a single option label from the valid answer set, and the answer parser normalizes the first valid capital-letter option. The exact extraction, verification, answer, and answer-audit prompts are included in Appendix A. For official submissions, the MCQ runs use the EVA pipeline with text verification, while the OpenQA runs use the full-data QLoRA fine-tuned model described below.

Table 3
QLoRA training configuration.

Parameter	Value
Quantization	4-bit NF4
LoRA rank	16
LoRA alpha	32
LoRA dropout	0.05
Target modules	`q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, `down_proj`
Learning rate	2e-4
Warmup ratio	0.03
Seed	42
Trainable parameters	~43.6M / 8.8B (~0.5%)

4.2. MCQ Experiments

4.2.1. Phase 1: Zero-Shot Baseline Probes

Initial experiments on EXAMS-V validation (50 samples) established baseline performance:

Full validation (4,651 samples) confirmed: baseline = 0.5063, OCR = 0.5063. **OCR provides no benefit for MCQ** and increases runtime by 5x.

Table 4
Zero-shot baseline probes on 50-sample validation.

Method	Prompt	OCR	Accuracy
baseline	baseline_v1	No	0.50
vote	mcq_vote_v1	No	0.44
ocr	baseline_v1	Yes	0.50
voteocr	mcq_vote_v1	Yes	0.46

4.2.2. Phase 2: QLoRA Fine-Tuning

Supervised fine-tuning on 2025 EXAMS-V with full training data (16,281 samples, 10 epochs) achieved `train_loss` = 0.0750 but only marginal test2025 accuracy improvement:

Table 5
MCQ accuracy with supervised fine-tuning.

Method	Accuracy
baseline (zero-shot)	0.4940
QLoRA fine-tuned (best checkpoint)	0.4948

Key finding: Standard SFT does not improve MCQ performance. The model overfits to training patterns without learning transferable reasoning.

4.2.3. Phase 3: Multi-Stage EVA Pipeline

The major breakthrough came from decomposing MCQ into three stages:

1. **Extract:** Extract structured text (QUESTION, OPTIONS, VISUAL_EVIDENCE, NUMBERS_AND_FORMULAS) from the exam image.
2. **Verify:** Cross-check extraction against the original image, fix OCR errors, preserve option order.
3. **Answer:** Use both the original image and cleaned text to select the answer.

Results on test2025 (3,565 questions):

Table 6
MCQ results on test2025.

Method	Accuracy	Correct	Delta vs Baseline
baseline (zero-shot)	0.4940	1,761	—
QLoRA fine-tune	0.4948	1,764	+3
EVA pipeline	0.5638	2,010	+249
EVA + text verification	0.5647	2,013	+252
EVA + QLoRA	0.5023	—	-148 (regression)

The EVA pipeline improves accuracy by **7.07 percentage points** (+252 questions). The gain comes entirely from the structured pipeline, not from fine-tuning.

The comparison between EVA without text verification (56.38%) and EVA with text verification (56.47%) should be read as a final-answer-level ablation of the verification stage. The isolated net gain is small (+0.09 percentage points, +3 examples), so we do not claim that verification alone is the source of the main improvement. Instead, we interpret it as a lightweight consistency-control stage inside the broader structured decomposition. As discussed in Section 4.4, verification can also introduce new errors, especially destructive edits that overwrite or reorder options.

4.2.4. Phase 4: Answer Audit

We also explored a fourth stage where the model re-checks its proposed answer against the image. Because the answer-audit run was recorded only as an estimated marginal gain and was not used for the official submission, we exclude it from the main quantitative comparison. We keep the prompt template in the appendix to document the explored system variant.

4.2.5. Phase 5: Prompt Ablation Study

Four prompt variants were tested on a focused "hard set" of 32 English/Physics questions:

Table 7
Prompt ablation on 32 English/Physics questions.

Prompt Variant	Accuracy	Notes
EVA baseline	0.0938	Standard extract-verify-answer
EVA + role prompt	0.1250	Subject-specific role prompt
EVA + CoT	0.0312	Chain-of-thought reasoning
EVA + role prompt + CoT	0.0625	Role + CoT combined

Findings:

- Role prompts provide marginal gains for Physics (+3.12 pp).
- Chain-of-thought reasoning **hurts** performance significantly (-6.26 pp).
- The combination of role + CoT is worse than role alone.
- No prompt-only path achieves 70% accuracy.

Extended to 96 samples across multiple languages, role prompts showed gains for English but **degradation** for Kazakh and Urdu, indicating language-specific sensitivities.

4.2.6. Phase 6: Reinforcement Learning (REINFORCE)

We trained REINFORCE agents starting from SFT adapters with entropy regularization. 28 RL experiment runs were conducted with batch size sweeps (16–64) and learning rate exploration.

RL training showed:

- Reward signal from accuracy on labeled questions.
- Entropy regularization to prevent mode collapse.
- Training plots available showing reward/loss/entropy dynamics.

The RL approach did not surpass the EVA pipeline in our experiments, suggesting that the bottleneck is in prompt-based reasoning structure rather than answer selection policy.

4.2.7. Phase 7: Agent-Based Pipeline

An ExamAgent using Qwen-Agent framework with tool-calling capabilities was implemented with:

- ExtractExamInfoTool for structured information extraction.
- VerifyAnswerTool for cross-checking against images.
- SubmitAnswerTool for answer submission.
- ImageZoomInToolQwen3VL for region-focused analysis.

The agent approach provides a framework for interactive reasoning but did not outperform the offline EVA pipeline.

4.3. OpenQA Experiments

Because the official dev/test labels are hidden, OpenQA results should be interpreted under two different training protocols. In the *partial-train → holdout* protocol, we first carve out a fixed holdout split from the official training set, train on the remaining subset, and evaluate on the holdout set. This is our strict local generalization protocol. In the *full-train → holdout* protocol, we train on the full official training set and then reuse the same holdout split only as a diagnostic reference. Since the holdout examples are included in the full training data, these diagnostic scores should not be interpreted as strict generalization results or compared directly with official leaderboard scores. The fixed holdout split uses random seed 42, validation ratio 0.2, and language-stratified sampling, yielding 240/60 samples for textual OpenQA and 423/105 samples for visual OpenQA.

4.3.1. Textual OpenQA

We retain both *partial-train → holdout* and *full-train → holdout* results for textual OpenQA.

Table 8
Textual OpenQA results under partial-train → holdout.

Method	BLEU-Avg	ROUGE-L	METEOR	COMET
baseline	0.1380	0.1810	0.1545	0.6718
OCR	0.1219	0.1936	0.1773	0.6822
QLoRA ckpt-30	0.2187	0.2438	0.1903	0.7042

A. partial-train (240) → holdout (60). Per-language COMET for the best partial-train model (ckpt-30):

Table 9
Per-language COMET for textual OpenQA (partial-train ckpt-30).

Language	COMET
English	0.7929
Italian	0.7765
Bulgarian	0.7407
Croatian	0.7217
Chinese	0.6308
Serbian	0.5624

Table 10
Diagnostic textual OpenQA results under full-train → holdout; not a strict generalization setting.

Method	BLEU-Avg	ROUGE-L	METEOR	COMET
QLoRA ckpt-133	0.9833	0.6167	0.8047	0.9833

B. diagnostic full-train (official train) → holdout (60). Key findings:

- Under the strict *partial-train → holdout* protocol, QLoRA fine-tuning consistently improves all textual OpenQA metrics, unlike the weaker gains observed for MCQ.
- OCR improves semantic metrics (ROUGE-L, METEOR, COMET) under partial-train evaluation, but reduces surface-form matching (BLEU).

- The much stronger full-train ckpt-133 results indicate that additional labeled samples substantially improve answer fitting, although these diagnostic scores should not be interpreted as strict holdout-protocol results.
- COMET and METEOR remain more reliable indicators for multilingual short-answer evaluation than BLEU alone.
- Since fine-tuning is clearly effective for OpenQA and the strongest model is obtained with more labeled training data, our final official OpenQA submission uses the full-train fine-tuned model.

4.3.2. Visual OpenQA

For visual OpenQA, the currently verifiable local records cover both the strict *partial-train → holdout* setting and the stronger diagnostic *full-train → holdout* reference setting, so we report them separately.

Table 11
Visual OpenQA results under partial-train → holdout.

Method	BLEU-Avg	ROUGE-L	METEOR	COMET
baseline	0.1222	0.1837	0.1454	0.6367
QLoRA ckpt-106 (task-specific visual)	0.1489	0.1916	0.1713	0.6678

Table 12
Diagnostic visual OpenQA results under full-train → holdout; not a strict generalization setting.

Method	BLEU-Avg	ROUGE-L	METEOR	COMET
QLoRA ckpt-85	0.9359	0.6529	0.8056	0.9680

Under the strict *partial-train → holdout* protocol, the best confirmed task-specific visual checkpoint is ckpt-106, which improves over the baseline on ROUGE-L, METEOR, and COMET. The diagnostic full-train ckpt-85 model improves much more strongly, especially on COMET, showing that visual OpenQA also benefits from fine-tuning and from access to more labeled data. These full-train numbers document the model selected for submission but should not be read as strict local generalization. For the final submission, we therefore prioritized the fully trained fine-tuned model rather than the reduced-data protocol.

4.4. Error Analysis

Analysis of 1,552 errors in the best MCQ run (EVA + text verification) reveals four error categories:

Table 13
Error categories in MCQ.

Error Type	Description	Impact
Type A	Image semantics not structured (visual Physics/Math)	Hardest for graphs, diagrams, geometric figures
Type B	Multi-language OCR noise causing option mapping errors	Affects Urdu, Arabic, Kazakh, Bulgarian
Type C	Verify overwriting and reordering original options	Destructive edits during verification
Type D	Answer stage instability	Same extract/verify, different final answer

These two error-distribution tables further show that failures concentrate on science-heavy questions and languages with more challenging OCR characteristics, which is consistent with the preceding error taxonomy.

Table 14
Error distribution by subject.

Subject	Error Count	Percentage
Physics	526	33.89%
Mathematics	255	16.43%
Chemistry	170	10.95%
Biology	164	10.57%
Informatics	84	5.41%

Table 15
Error distribution by language.

Language	Error Count	Percentage
English	225	14.50%
Chinese	158	10.18%
Urdu	144	9.28%
Arabic	135	8.70%
Kazakh	130	8.38%

4.5. Cross-Experiment Comparison Summary

To distinguish this section from the preceding error analysis and the following official leaderboard report, we summarize the core metrics of the main experimental settings separately below. OpenQA columns summarize the locally reported model-selection results; strict partial-train and diagnostic full-train protocols are separated in Section 4.3.

Table 16
Cross-experiment comparison summary.

Approach	MCQ Accuracy	Textual OpenQA COMET	Visual OpenQA METEOR
Zero-shot baseline	0.4940	0.6718	0.1454
QLoRA fine-tune	0.4948	0.7042	0.8056
EVA pipeline	0.5638	—	—
EVA + text verification	0.5647	—	—
EVA + QLoRA	0.5023	—	—

4.6. Official Leaderboards

The following rankings are transcribed from the official ImageCLEF 2026 MultimodalReasoning leaderboard snapshot dated May 20, 2026. Our official MCQ submissions correspond to the EVA pipeline + text verification method, while our official OpenQA submissions correspond to the full-data fine-tuned model.

The official leaderboard scores are computed on hidden evaluation data and are therefore the primary external reference for competition performance. They are not directly comparable with the local diagnostic full-train-to-holdout scores reported above, which are included only to explain model-selection decisions before submission.

Across the four official tracks, our methods are more competitive on textual tasks than on visual ones. In particular, the EVA pipeline + text verification submission ranks near the top on textual MCQ while remaining mid-to-upper tier on visual MCQ, whereas the full-data fine-tuned OpenQA model reaches 1st place on textual OpenQA, suggesting that supervised fine-tuning is especially effective for generative short-answer settings.

These leaderboard outcomes are also consistent with the experimental findings discussed earlier: the main bottleneck on visual tracks remains image perception and structured evidence extraction, while textual tracks benefit more directly from explicit verification or supervised fine-tuning. We first report the two MCQ leaderboards, followed by the two OpenQA leaderboards.

Table 17
Official visual MCQ leaderboard (May 20, 2026).

Rank	Participant	Accuracy
1	spirosbax	0.8406
2	DS@GT	0.7986
3	mohamedbasem	0.7108
4	zhaijinghe	0.6150
5	sjaini	0.5927
6	begyed	0.5774
7	linxiaocan (Our)	0.5560
8	pratikpriyanshu	0.5076
9	wether	0.4673

Table 18
Official textual MCQ leaderboard (May 20, 2026).

Rank	Participant	Accuracy
1	pratikpriyanshu	0.7538
2	linxiaocan (Our)	0.7084
3	rafiulbiswas	0.6152

Table 19
Official visual OpenQA leaderboard (May 20, 2026).

Rank	Participant	COMET	BLEU	ROUGE-L	METEOR
1	mohamedbasem	0.6488	0.1391	0.2762	0.2383
2	wangshou66	0.6366	0.1308	0.2717	0.2388
3	uned-martinez	0.5938	0.0980	0.2452	0.1842
4	liuzhe3085	0.5829	0.1556	0.3351	0.2295
5	wenlong	0.5497	0.0680	0.1713	0.1185
6	rafiulbiswas	0.5143	0.0735	0.2013	0.1331
7	linxiaocan (Our)	0.5117	0.0541	0.1434	0.0936
8	pratikpriyanshu	0.4286	0.0375	0.0993	0.0712

Table 20
Official textual OpenQA leaderboard (May 20, 2026).

Rank	Participant	COMET	BLEU	ROUGE-L	METEOR
1	linxiaocan (Our)	0.6563	0.1871	0.3032	0.2139
2	pratikpriyanshu	0.5285	0.1142	0.2300	0.1591

Within the complete official rankings, our full-data OpenQA fine-tuning method places 1st in textual OpenQA, while our MCQ EVA pipeline + text verification method places 2nd in textual MCQ and 7th in both visual tracks.

5. Discussion

5.1. Why the EVA Pipeline Works

The EVA pipeline's effectiveness stems from decomposing a complex visual reasoning task into manageable sub-problems. The extraction stage converts visual information into structured text, the verification stage corrects OCR errors and hallucinations while preserving option integrity, and the answer stage leverages both visual and textual evidence. This is analogous to how human test-takers approach exam questions: first read the question carefully, then verify understanding, then reason about the answer.

5.2. Role and Limits of Verification

The verification stage is useful as a consistency-control step, but our existing ablation shows that its isolated final-answer gain is modest: EVA rises from 56.38% to 56.47% after adding text verification. This means the main improvement should be attributed to the overall structured decomposition rather than to verification alone. The error taxonomy also shows that verification can hurt some cases by overwriting or reordering options, so future systems should add stricter option-preservation constraints or post-hoc option canonicalization.

5.3. Why QLoRA Fails for MCQ but Succeeds for OpenQA

The divergence between MCQ and OpenQA fine-tuning results is noteworthy. For MCQ, the model must select from fixed options — a task that depends more on accurate visual understanding than on answer format learning. Overfitting to training patterns may actually hurt MCQ by biasing the model toward incorrect reasoning shortcuts. For OpenQA, fine-tuning helps because the task benefits from learning domain-specific answer phrasings and formatting conventions.

5.4. Limitations of Prompt Engineering

Our prompt ablation study shows diminishing returns from prompt engineering alone. While role prompts provide marginal gains for specific subjects, chain-of-thought reasoning degrades performance. This suggests the bottleneck is in the model's visual understanding capabilities, not in its reasoning instruction following.

5.5. Error Analysis Implications

The dominance of Physics (33.89%) and Mathematics (16.43%) in errors indicates that visual understanding of scientific diagrams, graphs, and equations remains the primary challenge. Multi-language OCR noise (Type B) affects non-Latin scripts disproportionately, suggesting the need for language-specific preprocessing.

6. Conclusion

We present a comprehensive experimental study on multimodal reasoning for multilingual exam question answering. Our key contributions are:

1. **Multi-stage EVA pipeline:** A task-oriented extract-verify-answer pipeline that improves MCQ accuracy by 7.07 percentage points (from 49.40% to 56.47%) over zero-shot baselines, demonstrating that structured decomposition of visual reasoning tasks is more effective than end-to-end approaches in this setting.
2. **QLoRA effectiveness spectrum:** QLoRA fine-tuning significantly improves OpenQA performance (COMET +3.24 pp, METEOR +66 pp for visual) but fails to improve MCQ accuracy, revealing task-dependent effectiveness of supervised fine-tuning.

3. **Prompt engineering limits:** Subject-specific role prompts provide marginal gains for MCQ, while chain-of-thought reasoning consistently degrades performance in the current pipeline configuration.
4. **Error taxonomy:** We identify four error categories (visual structure, OCR noise, option reordering, answer instability) and show that Physics and Mathematics questions remain the primary challenge.
5. **Official leaderboard reporting:** We include the complete four-track official leaderboards dated May 20, 2026. Our full-data OpenQA fine-tuning method ranks 1st in textual OpenQA, while our MCQ EVA pipeline + text verification method ranks 2nd in textual MCQ and 7th in both visual tracks.
6. **Practical insights:** OCR preprocessing is unnecessary for MCQ but beneficial for OpenQA semantic metrics; COMET and METEOR are more reliable than BLEU for multilingual short-answer evaluation.

Future work should focus on: (a) improving visual understanding of scientific diagrams through specialized training data, (b) developing language-specific OCR and preprocessing pipelines, (c) implementing option canonicalization to prevent reordering errors, and (d) exploring consistency-based answer selection strategies.

Acknowledgments

This work is supported by the National Social Science Foundation of China (24BYY080).

Declaration on Generative AI

During the preparation of this work, the authors used a generative AI tool for limited figure production in order to create the overview diagram in Figure 1. The authors also used language tools for limited grammar and spelling checks. After using these tools, the authors carefully reviewed and edited the content as necessary and assume full responsibility for the content of this publication.

References

[1] D. Dimitrov, M. S. Hee, M. Ahsan, S. Ahmad, D. Zlatkova, G. Pachov, Z. Xie, P. Nakov, I. Koychev, Overview of the ImageCLEF 2026 Task on Multimodal Reasoning, in: CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.

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

[3] 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: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024, pp. 7768–7791.

[4] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer, Qlora: Efficient finetuning of quantized llms, Advances in neural information processing systems 36 (2023) 10088–10115.
[5] R. Rei, J. G. De Souza, D. Alves, C. Zerva, A. C. Farinha, T. Glushkova, A. Lavie, L. Coheur, A. F. Martins, Comet-22: Unbabel-ist 2022 submission for the metrics shared task, in: Proceedings of the Seventh Conference on Machine Translation (WMT), 2022, pp. 578–585.
[6] S. Banerjee, A. Lavie, Meteor: An automatic metric for mt evaluation with improved correlation with human judgments, in: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, 2005, pp. 65–72.
[7] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, J. Zhou, Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. URL: https://arxiv.org/abs/2308.12966. arXiv:2308.12966.
[8] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al., Chain-of-thought prompting elicits reasoning in large language models, Advances in neural information processing systems 35 (2022) 24824–24837.
[9] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, D. Zhou, Self-consistency improves chain of thought reasoning in language models, arXiv preprint arXiv:2203.11171 (2022).
[10] 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).
[11] S. Ahmed, M. T. Younes, A. Moustafa, A. Allam, H. Moustafa, Msa at imageclef 2025 multimodal reasoning: Multilingual multimodal reasoning with ensemble vision language models, arXiv preprint arXiv:2507.11114 (2025).
[12] J. Yan, L. Kong, Q. Wu, J. Li, Multi-prompt ensemble reasoning for multimodalreasoning (2025).
[13] A. Madaan, N. Tandon, P. Gupta, S. Hallinan, L. Gao, S. Wiegreffe, U. Alon, N. Dziri, S. Prabhumoye, Y. Yang, S. Gupta, B. P. Majumder, K. Hermann, S. Welleck, A. Yazdanbakhsh, P. Clark, Self-refine: Iterative refinement with self-feedback, Advances in Neural Information Processing Systems 36 (2023).
[14] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, Y. Cao, React: Synergizing reasoning and acting in language models, in: International Conference on Learning Representations, 2023.
[15] T. Schick, J. Dwivedi-Yu, R. Dessi, R. Raileanu, M. Lomeli, L. Zettlemoyer, N. Cancedda, T. Scialom, Toolformer: Language models can teach themselves to use tools, in: Advances in Neural Information Processing Systems, volume 36, 2023.
[16] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.
[17] C.-Y. Lin, Rouge: A package for automatic evaluation of summaries, in: Text summarization branches out, 2004, pp. 74–81.

A. Prompt Templates

A.1. Extract Prompt

```
Extract the multiple-choice question from this exam image.
The question language is {language}. The subject is {subject}.
Do not solve it. Return plain text with the following sections in order:
QUESTION:
OPTIONS:
VISUAL_EVIDENCE:
NUMBERS_AND_FORMULAS:
Keep the original language. Preserve the option order exactly as shown in the image.
```

A.2. Verify Prompt

```
You are verifying a noisy text extraction against the original exam image.
The question language is {language}. The subject is {subject}.
The image is the source of truth. The input extraction is only a draft to be checked
    against the image.
Compare the draft with the image line by line. Fix OCR errors, remove hallucinated text,
    restore missing details visible in the image, and keep only information that is
    supported by the image.
Preserve the original option labels and option order exactly as they appear in the image.
     Never reorder, sort, renumber, or relabel the options.
Do not answer the question. Return plain text only using exactly these sections:
QUESTION:
OPTIONS:
VISUAL_EVIDENCE:
NUMBERS_AND_FORMULAS:
```

A.3. Answer Prompt

```
You are solving an exam multiple-choice question from image evidence and cleaned context.
The question language is {language}. The subject is {subject}.
Use both image and context. Choose the single best option.
Respond with ONLY one capital letter from {A,B,C,D,E}.

CLEANED CONTEXT:
{context_text}
```

A.4. Answer Audit Prompt

```
You are auditing a proposed answer to an image-based multiple-choice exam question.
The question language is {language}. The subject is {subject}.
Look at the original image again and compare it against the cleaned context and the
    proposed answer.
If the proposed answer is supported by the image and context, keep it. If it is not
    supported, change it.
Respond with ONLY one final capital letter from {A,B,C,D,E}.

PROPOSED_ANSWER: {proposed_answer}

CLEANED CONTEXT:
{context_text}
```