DS@GT at ImageCLEF 2026 MultimodalReasoning: Visual Multiple Choice Question Reasoning with Vision-Language Models

MultimodalReasoning at ImageCLEF 2026

Yue Li[1,*,†], Zhanxu Liu[1,†] and Chengxi Zhang[1,†]

[1]Georgia Institute of Technology, North Ave NW, Atlanta, GA 30332

Abstract

The ImageCLEF 2026 MultimodalReasoning Visual MCQ task asks a model to pick the right answer to multilingual exam questions given as images. We attempt two model tracks. On a normal-scale model (Qwen3.5-9B), we run GRPO fine-tuning with a reward that rewards correct answers, penalizes over-long responses, and gives a small bonus for output format. On a tiny model (Gemma 4 E2B - 5B), we fine-tune with LoRA on reasoning traces distilled from a larger teacher (Gemma 4 31B). GRPO raises test accuracy from 0.73 to 0.78, cuts the invalid rate, and shortens responses, and the gains hold across many subjects and languages. But a question-level check shows the model mostly improves its output form, not its reasoning. For the tiny model, SFT is necessary (+12 points on validation), while OCR effects are split- and prompt-dependent rather than reliably beneficial.

Keywords

Multimodal Reasoning, Vision-Language Models, Optical Character Recognition, Multilingual, ImageCLEF

1. Introduction

Recent years have brought major progress in Vision-Language Models (VLM), especially in image captioning, visual question answering, and visual dialogue. Open-source VLMs have advanced quickly in the past year, with better benchmark performance, longer context windows, multilingual support, and even video understanding in some systems [1]. Despite the progress, challenges remain, especially around compositional reasoning and linguistic bias, where models can still make mistakes in how they interpret object relationships in images [2].

The ImageCLEF 2026 MultimodalReasoning competition [3, 4] focuses on evaluating and advancing the next generation of frontier multimodal and VLM on challenging, reasoning-intensive academic exam problems. The core tasks of the competition are split into two primary paradigms: Multiple-Choice Question Answering (MCQ) and Open Question Answering (OpenQA). Our work was focused only on the Visual MCQ task.

Top performers from 2025 like ContextDrift [5] explored the boundaries of proprietary APIs by altering reasoning token budgets and testing zero-shot or few-shot prompts. Meanwhile, systems like MSA [6] deployed multi-stage ensemble pipelines integrating Optical Character Recognition (OCR) to safeguard prediction formatting on multilingual exam material [7]. The 2026 rules mandate a transition of these insights into the open-source domain.

Inspired by these recent trends and the success of reinforcement learning (RL) in scaling reasoning behaviors, this paper explores the training-time adaptation of open-source VLMs for multilingual exam reasoning. Specifically, we draw upon the architectural insights of DeepSeek-R1, which demonstrates that reasoning capabilities can be substantially incentivized via Group Relative Policy Optimization (GRPO) and a cold-start Supervised Fine-Tuning (SFT) phase. Our work evaluates a dual training

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

*Corresponding author.

†These authors contributed equally.

yli3720@gatech.edu (Y. Li); zliu943@gatech.edu (Z. Liu); czhang808@gatech.edu (C. Zhang)

strategy: applying parameter-efficient SFT via LoRA on a tiny-scale VLM, alongside executing GRPO updates on a normal-scale VLM.

2. Related Work

Recent years have seen increasing interest in multimodal reasoning tasks, particularly in multilingual exam-style question answering settings. ImageCLEF Multimodal Reasoning 2026 [4, 8] represents the second edition of the competition and further extends the task setting introduced in 2025. Building on the 2025 ImageCLEF Multimodal Reasoning task [7], the 2026 edition introduces two key changes: (1) the addition of an open-ended question answering (OpenQA) track alongside the existing MCQ track, and (2) a restriction to open-source models, prohibiting proprietary API-based systems. These changes reshape the design space for participating systems and motivate us to revisit the lessons of the 2025 top performers from an open-source perspective.

Among the top-performing systems in ImageCLEF 2025, ContextDrift [5] investigated the impact of prompt design, OCR augmentation, and reasoning token budget on multimodal reasoning performance. Their system utilized Gemini 2.5 Flash combined with OCR-extracted textual information and explored multiple prompting strategies, including zero-shot, few-shot, and structured prompting. One of their key findings was that increasing the thinking budget does not always improve performance. The authors suggested that multimodal reasoning tasks may exhibit an "optimal reasoning length," highlighting the importance of reasoning token control in multimodal systems.

Another strong-performing team, MSA [6], proposed an OCR–VLM ensemble pipeline that combined multi-stage caption generation, caption refinement, and answer verification to improve multilingual multimodal reasoning performance. Their work emphasized the importance of prompt engineering for stable answer prediction. Specifically, they observed that allowing models to generate unrestricted explanations often led to formatting errors and unnecessarily verbose outputs. In contrast, using a strict "letter-only" prompting strategy, where the model outputs only the final answer choice (A/B/C/D/E), significantly improved final accuracy. Their work further demonstrated that OCR augmentation and staged reasoning pipelines can improve robustness in multilingual exam-style reasoning tasks.

The 2026 edition's open-source requirement makes inference-time techniques on proprietary APIs (the dominant strategy in 2025) inapplicable, motivating a shift toward training-time adaptation of open-source VLMs. We draw inspiration from DeepSeek-R1 [9], which demonstrates that reasoning ability in language models can be substantially incentivized through reinforcement learning. DeepSeek-R1 introduces two key techniques: Group Relative Policy Optimization (GRPO), a simplified variant of PPO that replaces the critic network with a group-relative baseline computed from multiple sampled outputs per prompt, dramatically reducing RL training cost; and a cold-start SFT phase that fine-tunes the base model on a small set of curated chain-of-thought examples before RL, stabilizing subsequent reinforcement learning. To our knowledge, however, this paradigm has not yet been applied to the multilingual multimodal exam reasoning setting represented by EXAMS-V, and its behavior across different model scales is not yet understood. This gap motivates our two-pronged approach: SFT+LoRA on a tiny VLM, and GRPO on a normal-scale VLM.

3. Data and Tasks

The task of Multiple-Choice Question Answering (MCQ) can be interpreted as a classification task: Given an image of a multi-disciplinary exam question paired with text context and three to five multiple-choice options, the system must reason through the multimodal evidence to accurately select the single correct answer [4]. The problems are derived from high-quality academic datasets (MBZUAI/EXAMS-V) [10], featuring reasoning-heavy scientific and technical content heavily reliant on diagrams, charts, technical schematics, graphs, and tables.

The training dataset used in this study consists of 24,497 exam question samples. Each sample includes an answer key variable indicating the correct answer choice ("A" to "E" in various languages).

Table 1
Overall accuracy on the visual partition of EXAMS-V validation.

Model	Token budget	n	Acc (%)	Invalid (%)	Median len.	Mean len.
Qwen3.5-9B	6,144	1,195	70.13	21.17	2,190	4,324
Qwen3.5-27B	6,144	1,195	75.15	18.91	2,039	4,083

Note: A response is counted as *invalid* if no answer letter can be extracted after the closing `</think>` tag; invalid responses count as incorrect in the accuracy denominator.

The dataset is heterogeneous and comprises multimodal exam questions (combining visual elements and text) across various dimensions, including 14 different languages (imbalanced distribution), grade levels ranging from 4 to 12 (imbalanced, mainly 12th grade), 44 subjects, and question types with 76% text and 24% image+text. The dataset was partitioned into predetermined training set: 16,281 samples (66.46%), validation set: 4,651 samples (18.99%), and test set: 3,565 samples (14.55%). During the assessment exploration of the dataset, a data cleaning step was executed on most of the attributes to ensure consistency for downstream tasks and modeling.

The test dataset used in this study consists of 1117 exam question samples. Unlike the training set, this dataset does not contain an answer key variable, as the ground truth labels remain hidden for evaluation purposes. The dataset is heterogeneous and comprises multimodal exam questions across various dimensions, including 6 different languages with an imbalanced distribution (English: 44.76%, Chinese: 30.62%, Bulgarian: 9.85%, Croatian: 5.10%, Italian: 4.83%, and Serbian: 4.83%). After cross-lingual normalization to remove duplicate naming conventions across languages and splitting mixed categories into unified disciplines, the dataset spans multiple canonical academic subjects—with Mathematics, Physics, Chemistry, and Biology being the most prominent. Additionally, the dataset features a heavy integration of rich visual elements across its samples, consisting of 57.65% Diagrams, 26.68% Tables, 15.85% Graphs, and 11.55% Other categories, with many samples exhibiting multi-label co-occurrences of these visual features.

4. Experiments and Results

4.1. Baseline

To comply with the open-source requirements, initially we selected Qwen3.5 as our baseline for this task.

On EXAMS-V's validation split. On the 1,195 visual samples of the EXAMS-V validation partition, Qwen3.5-27B reaches 75.15% overall against 70.13% for Qwen3.5-9B, with invalid rates of 18.91% and 21.17%, respectively (Table 1). The larger model's advantage is broad: across subjects (Table 2) it leads or ties the 9B model on every category, with the largest margins on Geography and Informatics; across languages (Table 3) it leads everywhere except English, Bulgarian, and German. On Chinese, the largest language subset (475 items), the 27B model scores 71.79% against 66.32%. English is by a wide margin the weakest language for both models (49.45% and 52.75%).

On the Visual-MCQ test set. On the 1,117 questions of the official Visual-MCQ test set, Qwen3.5-27B again leads Qwen3.5-9B, at 75.02% versus 73.05% (Table 4). The table also includes a budget-extended Qwen3.5-9B reference run (30,000-token budget) at 78.96%. Unlike on EXAMS-V, the same-budget 27B advantage here is broad but not uniform: across subjects (Table 5) it leads the 9B model on Chemistry, Biology, Geography, and History, ties on Informatics, Fine Arts, and Psychology, but trails on Physics and Mathematics. The budget-extended 9B run is the strongest configuration on five of the nine subjects (Chemistry, Physics, Geography, Mathematics, and Fine Arts), while the 27B model keeps the lead on Biology and History. Across languages (Table 6), the same-budget 27B leads the 9B model everywhere

Table 2
Baseline accuracy (%) by subject on the EXAMS-V validation visual partition.

Subject	Count	Qwen3.5-9B	Qwen3.5-27B
Physics	328	68.60	70.12
Biology	260	72.69	78.85
Chemistry	208	69.23	69.71
History	184	78.80	84.78
Geography	99	63.64	81.82
Mathematics	31	64.52	67.74
Informatics	27	55.56	70.37
Science	26	80.77	80.77
Professional	20	55.00	60.00
Social	7	14.29	57.14
Sociology	5	80.00	80.00

Table 3
Baseline accuracy (%) by language on the EXAMS-V validation visual partition.

Language	Count	Qwen3.5-9B	Qwen3.5-27B
Chinese	475	66.32	71.79
Croatian	127	75.59	82.68
Italian	126	74.60	81.75
Serbian	126	75.40	84.92
Arabic	91	72.53	78.02
English	91	52.75	49.45
Bulgarian	59	88.14	86.44
Hungarian	32	81.25	84.38
German	29	72.41	68.97
Polish	20	55.00	60.00
French	19	73.68	84.21

Table 4
Overall accuracy on the visual-MCQ test set.

Model	Token budget	n	Acc (%)	Invalid (%)	Median len.	Mean len.
Qwen3.5-9B	6,144	1,117	73.05	22.83	2,150	4,558
Qwen3.5-27B	6,144	1,117	75.02	22.65	3,432	5,926
Qwen3.5-9B	30,000	1,117	**78.96**	15.40	2,229	12,191

except Croatian, and the budget-extended 9B leads or ties on every language except Italian. Chinese is by a wide margin the weakest language for all three configurations — the counterpart of English on EXAMS-V — and is also where the budget-extended run improves most over its 6k counterpart.

4.2. GRPO fine-tuning with Qwen3.5-9B

4.2.1. Setup

In a EXAMS-V validation result, it is obvious that for Physics, Chemistry and Mathematics, incorrectly answered questions had longer responses than other subjects (Figure 1), and accuracy on these subjects was comparatively low. We suspect that this is because these subjects involve more formulas and reasoning steps, leading the model to over-think and run past its output budget. Building on this observation, we designed three GRPO experiments to test whether this method can improve the model on its relatively weaker subjects: two *single-subject* runs trained only on Chinese Physics (with and without a length penalty in their reward functions), and a *tri-subject* run trained on Chinese Physics,

Table 5
Baseline accuracy (%) by subject on the Visual-MCQ test set.

Subject	n	Qwen3.5		
		9B (6k)	27B (6k)	9B (30k)
Chemistry	344	71.22	74.42	**80.23**
Biology	305	77.05	**81.64**	80.98
Physics	299	67.89	66.56	**73.58**
Geography	55	80.00	83.64	**85.45**
Mathematics	47	76.60	70.21	**80.85**
Fine Arts	32	75.00	75.00	**78.12**
History	18	72.22	**83.33**	77.78
Informatics	14	**92.86**	**92.86**	85.71
Psychology	3	**100.00**	**100.00**	**100.00**

Table 6
Baseline accuracy (%) by language on the Visual-MCQ test set.

Language	n	Qwen3.5		
		9B (6k)	27B (6k)	9B (30k)
English	500	79.40	80.60	**81.00**
Chinese	342	55.56	58.48	**69.88**
Bulgarian	110	80.91	83.64	**85.45**
Croatian	57	**85.96**	84.21	**85.96**
Italian	54	83.33	**88.89**	87.04
Serbian	54	85.19	87.04	**88.89**

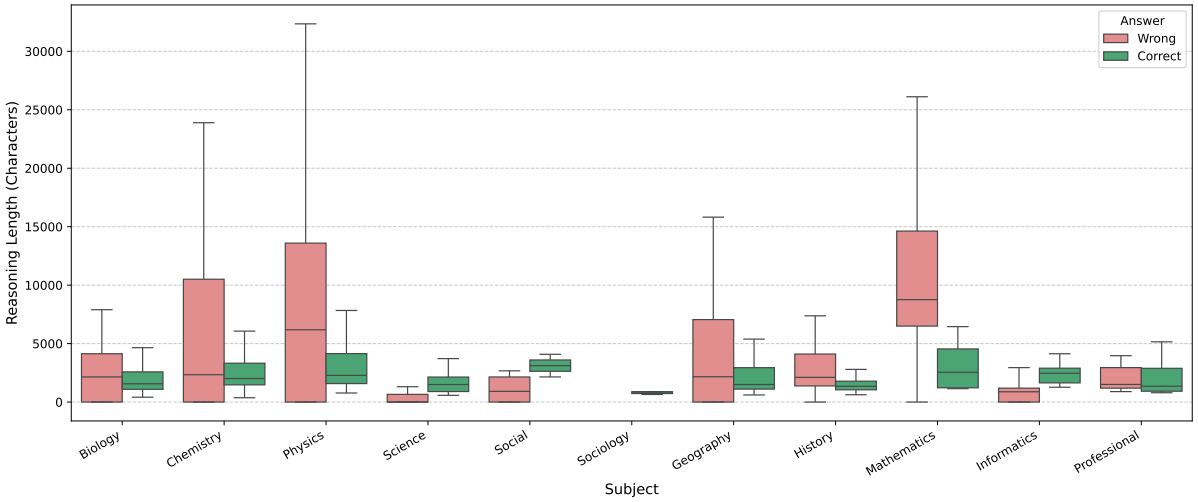

Figure 1: Reasoning character length vs. correctness by subject on the EXAMS-V validation set (Qwen3.5-9B with 20000 token budget).

Chemistry and Biology (see Table 7).

The training data are filtered from EXAMS-V; the train-time val data and test data are from the visual-MCQ test set. Train-time val data is veRL's periodic in-training validation set, on which the current policy is evaluated every 5 steps to track progress during GRPO; it is a monitoring signal only and is never used for reward computation or gradient updates.

Table 7
Experiment configuration.

	single-subject	tri-subject	single-subject (no-len)
Training data	Chinese Physics (n=371)	Chinese Phys./Chem./Bio. (n=927; Phys 371 / Chem 328 / Bio 228)	Chinese Physics (n=371)
Train-time val[1]	Chinese Physics (n=87)	Chinese Phys./Chem./Bio. (n=316; Phys 87 / Chem 175 / Bio 54)	Chinese Physics (n=87)
Reward[2]	$R = c - 0.3\,(1-c)\,\min(\ell/L,\,1) + 0.05\,f$		$R = c + 0.05\,f$
Hyperparameters	base model Qwen3.5-9B; LoRA rank=α=64 (all-linear); GRPO; lr=5×10^{-6}; KL coef =0.01 (low-var); train batch =16; PPO mini-batch =16; rollouts per prompt =4, T=1.0; max prompt =1024; max response L=6144; training steps =138.		

[1] "Train-time val" is veRL's periodic in-training validation set, on which the current policy is evaluated every 5 steps to track progress during GRPO; it is a monitoring signal only and is never used for reward computation or gradient updates.
[2] In the reward definitions, $c, f \in \{0, 1\}$ indicate a correct answer and a closed `</think>`, respectively, and ℓ is the response length.

Table 8
Overall metrics.

Run	n	Accuracy (%)	Invalid (%)	Median len.	Mean len.
baseline	1117	73.1	22.8	2150	4558
single-subject	1117	78.5	13.7	1723	3496
tri-subject	1117	79.2	14.0	1734	3489
single-subject (no-len)	1117	**79.5**	**12.1**	**1712**	**3355**

Table 9
Accuracy and invalid rate by subject.

Subject	n	Accuracy (%)				Invalid (%)			
		base	single	tri	single (no-len)	base	single	tri	single (no-len)
Chemistry	344	71.2	79.4	78.8	**79.9**	26.7	16.3	16.0	**15.4**
Biology	305	77.0	81.0	81.3	**82.0**	18.7	9.2	10.2	**8.2**
Physics	299	67.9	71.9	**75.9**	75.6	26.8	18.1	17.1	**13.4**
Geography	55	80.0	**83.6**	**83.6**	80.0	10.9	**5.5**	**5.5**	**5.5**
Mathematics	47	76.6	**85.1**	83.0	80.9	21.3	**10.6**	14.9	14.9
Fine Arts	32	75.0	**81.2**	75.0	**81.2**	21.9	**12.5**	18.8	**12.5**
History	18	72.2	77.8	**83.3**	72.2	16.7	16.7	**11.1**	16.7
Informatics	14	**92.9**	**92.9**	85.7	**92.9**	**0.0**	**0.0**	7.1	**0.0**

4.2.2. Results

Overall, all experiments improve over the visual-MCQ test set baseline (Table 8): accuracy rises, the invalid rate drops, and response length shrinks.

By subject and language. The improvement holds across many subjects and languages, outside the training distribution (Tables 9 and 10). Although the single-subject experiments only used Physics data, they achieve higher accuracy not in Physics, but other subjects (e.g. Chemistry and Biology). The median response length shrinks for almost all languages and subjects (Table 11), most obviously for Chinese, Physics, Chemistry, Mathematics, and Biology.

Table 10
Accuracy and invalid rate by language.

Language	n	Accuracy (%)				Invalid (%)			
		base	single	tri	single (no-len)	base	single	tri	single (no-len)
English	500	79.4	83.4	**85.0**	83.4	15.4	6.8	7.8	**6.4**
Chinese	342	55.6	64.6	64.6	**68.4**	42.4	28.9	28.1	**24.3**
Bulgarian	110	80.9	**85.5**	83.6	84.5	13.6	9.1	10.9	**8.2**
Croatian	57	86.0	**87.7**	**87.7**	82.5	8.8	7.0	**5.3**	8.8
Serbian	54	85.2	85.2	**90.7**	88.9	11.1	9.3	**3.7**	7.4
Italian	54	83.3	**90.7**	88.9	**90.7**	13.0	**1.9**	7.4	3.7

Table 11
Median completion length (characters).

Language	base	single	tri	single (no-len)
Chinese	6232	3573	3270	3146
Croatian	2100	1501	1496	1532
Bulgarian	2036	1763	1731	1689
Italian	1880	1506	1579	1593
English	1790	1467	1468	1449
Serbian	1699	1585	1588	1718

Subject	base	single	tri	single (no-len)
Mathematics	3031	2870	2497	2453
Chemistry	2774	1864	1970	1854
Physics	2271	1823	1889	1781
Informatics	2076	1823	1684	2008
Biology	2075	1551	1580	1553
History	1791	1423	1361	1385
Fine Arts	1542	1407	1449	1365
Geography	1428	1210	1220	1271

Table 12
Response length by baseline trace length.

Baseline length	$n_{\text{all-correct}}$	base (chars)	single	tri	no-len
$<$1.5 k	317	1,171	$0.97\times$	$0.97\times$	$0.99\times$
1.5–3 k	289	1,997	$0.89\times$	$0.90\times$	$0.85\times$
3–5 k	82	3,753	$0.68\times$	$0.67\times$	$0.73\times$
5–10 k	55	6,889	$0.45\times$	$0.43\times$	$0.43\times$

Where length compresses. Table 11 shows that GRPO shortens responses overall, but not uniformly. To localize this effect, we bucket questions by the baseline response length and measure lengths only on items that all four systems answer correctly (Table 12). On short baseline traces, the GRPO runs stay close to the baseline length; on 5–10k-character baseline traces, all three compress to about 0.43–$0.45\times$ baseline.

Training dynamics. Despite slight differences in metrics such as clip ratio and KL loss, all three experiments follow highly similar trajectories (e.g. Figure 2): training reward, validation reward, and validation accuracy rise, while the invalid rate falls.

Question analysis. At the question level, all experiments make some new corrections and mistakes (see Table 13). We manually inspected some samples:

- *Wrong→right.* The model hesitates less at intermediate steps, runs through the whole answering process and commits to a final answer.
- *Right→wrong.* The model 1) fails to map some language's answer options (e.g. Bulgarian) onto the A–E letters, 2) misreads the visual elements, 3) gets the right knowledge and context, but stuck at the wrong reasoning, 4) starts overthinking.

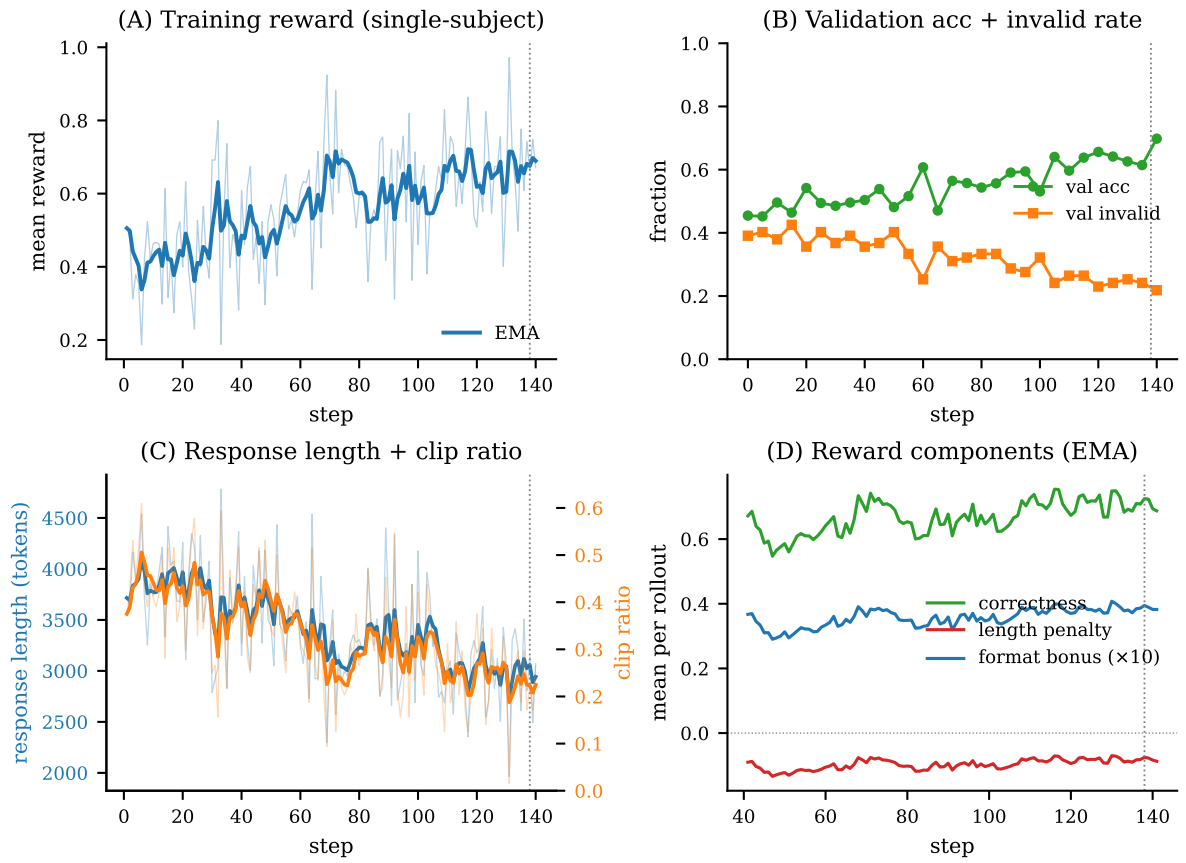

Figure 2: Single-subject experiment training curves.

Table 13
Paired question outcome counts vs. baseline.

Run	right→right	wrong→wrong	wrong→right	right→wrong	net
single-subject	783	207	94	33	+61
tri-subject	782	198	103	34	+69
single-subject (no-len)	785	198	103	31	+72

Table 14
Consistent mistakes.

	single (n=207)	tri (n=198)	single (no-len) (n=198)
invalid → invalid	129 (62.3%)	126 (63.6%)	114 (57.6%)
invalid → valid wrong	44 (21.3%)	35 (17.7%)	47 (23.7%)
valid wrong → invalid	6 (2.9%)	6 (3.0%)	3 (1.5%)
valid wrong → same wrong	23 (11.1%)	27 (13.6%)	32 (16.2%)
valid wrong → different wrong	5 (2.4%)	4 (2.0%)	2 (1.0%)

Among the consistent mistakes across baseline and experiments (Table 14), the invalid rates stay very high, and most of them, invalid → invalid group, are from Chemistry, Physics and Biology.

Summary. Compared with the baseline, GRPO results show a clear link between accuracy and model response length. However, the results do not distinguish whether GRPO primarily reduces overthinking, improves reasoning quality, or both.

Across the three GRPO variants, despite different data recipe or reward function, the quantitative metrics and qualitative patterns remain highly similar.

4.3. Gemma 4 E2B (5B)

4.3.1. Setup

We evaluate Gemma 4 E2B and a LoRA-fine-tuned variant trained on reasoning traces distilled from the larger Gemma 4 31B teacher; Table 15 summarizes the distillation and SFT configuration. Pre-submission ablations are run on a subset of the released EXAMS-V validation split matched to the per-language and per-subject distribution of the official test set using only public metadata. The official test labels were not used for training, hyperparameter tuning, prompt selection, or model selection. Student inference uses greedy decoding, 4-bit weight quantization, and `enable_thinking=True`. OCR is used only in the OCR-enabled inference ablations and OCR-aware test submission; it is not used during teacher-trace generation or LoRA SFT training. For OCR-enabled runs, PaddleOCR is applied to the full question image with metadata-based language routing. The original image is provided first, and the OCR text is appended as a noisy auxiliary context block rather than as a replacement for visual input.

Table 15
Distillation, SFT, and OCR configuration for Gemma 4 E2B.

Stage	Setting
Teacher distillation	Gemma 4 31B-IT, 4-bit NF4
Source split	EXAMS-V train, image_text only
Samples per question	3 (stochastic, T=0.7, top-p=0.95, top-k=64)
Max new tokens (teacher)	700
Quality filter	$\geq 2/3$ correct teacher runs; valid answer and reasoning trace
Final SFT samples	2,201 (zh 1072 / hr 483 / bg 336 / it 140 / sr 109 / en 61)
Top subjects	Physics 718, Geography 453, Biology 354, Chemistry 324
Student base model	Gemma 4 E2B (4-bit NF4)
LoRA rank / α / dropout	32 / 32 / 0
Adapted modules	q/k/v/o, gate/up/down, vision adapter
Optimizer / lr / schedule	AdamW (8-bit) / 5×10^{-6} / cosine, warmup 20
Batch size	2×4 grad-accum $= 8$
Max sequence length	2,048
Precision / Epochs	bf16 / 2
OCR engine	PaddleOCR with per-image language routing
OCR usage	Inference-time ablation only; not used for teacher distillation or SFT
OCR injection	Image-first prompt; OCR text appended, image remains primary evidence

4.3.2. Effect of Supervised Fine-Tuning (SFT)

Under matched prompting (`reasoning_commit`, 2,000-token budget), LoRA fine-tuning raises validation accuracy from 0.3300 to 0.4500, a gain of +12.00 pp (Table 16, rows 1–2). We discuss the corresponding test-set behavior in Section 4.3.5.

4.3.3. Prompt and Decoding-Budget Sensitivity

We iteratively refined the Gemma 4 E2B student prompt by inspecting sampled reasoning traces and constraining their output format. The filed Gemma submission used the `reasoning_commit` prompt, which asks the model to solve the question from the image, produce 3–5 short numbered reasoning steps in English, avoid explicit self-correction phrases such as "Wait", "Let me reconsider", "Actually", and "On second thought", and end with a fixed final answer line of the form `Answer: X`. Numeric, Cyrillic,

Arabic, or other non-Latin option labels are mapped to Latin A/B/C/D/E according to their order of appearance. This fixed final-answer format makes answer extraction deterministic. At evaluation time, we extract the final answer from the last `Answer: X` pattern, where $X \in \{A, B, C, D, E\}$; outputs without a valid final option are marked invalid and counted as incorrect.

Holding the LoRA model and decoding fixed, validation accuracy across prompt variants ranged from 0.30 to 0.45. Longer reasoning chains (5–8 steps) consistently under-performed, and the best configuration was `reasoning_commit` at a $2{,}000$-token decoding budget (Table 16, row 2), which we adopted for our final Gemma submission. We also evaluated two diagnostic prompt variants: `reasoning_direct_choice`, which imposes a stricter four-choice A–D answer format, and `reasoning_direct_choice_ocr`, which adds OCR text as auxiliary input while treating the original image as the source of truth. These A–D constrained prompts were used only for diagnostic ablations on the evaluated four-choice subset and were not the prompt used for the filed Gemma submission. The full Gemma prompt templates are provided in Appendix A.

4.3.4. OCR Ablation

Prior work on this task has suggested that OCR-extracted text can help small VLMs read dense multilingual question images [7]. We therefore evaluate OCR as an inference-time auxiliary signal for the Gemma student, using the setup in Table 15.

On the validation subset, OCR did not provide a consistent benefit. For the controlled `reasoning_commit` comparison, adding OCR text reduced accuracy from 0.4500 to 0.3700 while keeping the model, decoding budget, and final-answer format fixed (Table 16). The OCR-aware `reasoning_direct_choice_ocr` variant also underperformed its non-OCR counterpart on validation. These results show that OCR is not reliably helpful on this validation subset, possibly because the extracted text adds noise when the original image is already available.

Because the OCR-enabled variants did not improve validation accuracy, our filed Gemma submission used the no-OCR `reasoning_commit` configuration. We treat the later test-set OCR result as diagnostic rather than as evidence that OCR alone is beneficial.

Table 16
Accuracy of pre-submission runs on a subset of the EXAMS-V validation split.

Model	Prompt	Max tokens	OCR	Accuracy
Gemma 4 E2B (base)	`reasoning_commit`	2,000	No	0.3300
Gemma 4 E2B + SFT	`reasoning_commit`	**2,000**	**No**	**0.4500**
Gemma 4 E2B + SFT	`reasoning_commit`	2,000	Yes	0.3700
Gemma 4 E2B + SFT	`reasoning_direct_choice`	2,000	No	0.4300
Gemma 4 E2B + SFT	`reasoning_direct_choice_ocr`	2,000	Yes	0.3700

4.3.5. Test-Set Results

We report final test-set numbers in Table 17 (overall) and Table 18 (per-subject and per-language for the base vs. SFT comparison). The held-out test accuracy of our filed submission is 0.4673 (Gemma 4 E2B + SFT, `reasoning_commit`, no OCR), a $+3.94$ pp improvement over the base model—smaller in absolute size than the $+12.00$ pp gain observed on the validation subset (Section 4.3.2), but consistent in sign.

Replacing `reasoning_commit` with `reasoning_direct_choice` adds a small +0.36 pp gain. The OCR-aware diagnostic prompt reaches 0.4942, the best Gemma test result among our submitted variants (Table 17). However, this run is not a pure OCR-only ablation: it combines OCR text with the stricter `reasoning_direct_choice` prompt and an explicit A–D answer-space constraint. Therefore, the test-set gain cannot be attributed to OCR content alone.

This result contrasts with the validation subset, where OCR reduced accuracy under both evaluated prompt styles (Section 4.3.4). We interpret this as split- and prompt-dependent behavior rather than a

Table 17
Official test-set outcomes for Gemma 4 E2B variants.

Model	Prompt	OCR	Accuracy
Gemma 4 E2B (base)	`reasoning_commit`	No	0.4279
Gemma 4 E2B + SFT	`reasoning_commit`	**No**	**0.4673**
Gemma 4 E2B + SFT	`reasoning_direct_choice`	No	0.4709
Gemma 4 E2B + SFT	`reasoning_direct_choice_ocr`	Yes	0.4942

general OCR benefit. In particular, our evidence supports only the narrower claim that the OCR-aware diagnostic system performed best on the official test subset; it does not establish that PaddleOCR augmentation is reliably beneficial for Gemma 4 E2B.

Table 18
Test-set accuracy by subject and by language: base Gemma 4 E2B vs. Gemma 4 E2B + SFT.

Subject	n	base	+SFT	Language	n	base	+SFT
Chemistry	344	0.3721	**0.4535**	English	500	0.4980	**0.5460**
Biology	305	0.4525	**0.4557**	Chinese	342	0.2778	**0.3187**
Physics	299	0.4147	**0.4749**	Bulgarian	110	0.5182	**0.5636**
Geography	55	0.4909	0.4909	Croatian	57	**0.4912**	0.4561
Mathematics	47	**0.5319**	0.4043	Italian	54	**0.5000**	0.4815
Fine Arts	32	0.4375	**0.5000**	Serbian	54	0.4074	**0.4815**
History	18	0.5556	**0.6667**				
Informatics	14	**0.7143**	0.6429				

At the breakdown level, the SFT gains concentrate on the heavy STEM subjects (Chemistry $+8.14$ pp, Physics $+6.02$ pp) and on the high-volume languages (English $+4.80$ pp, Chinese $+4.09$ pp, Bulgarian $+4.55$ pp, Serbian $+7.41$ pp). The Mathematics (-12.77 pp) and Informatics (-7.14 pp) regressions coincide with their under-representation in the distillation set (123 and 25 training examples respectively, vs. 718 for Physics and 354 for Biology in Table 15), suggesting that limited teacher coverage may be one factor limiting where the student improves.

5. Discussion and future work

We tried two models and two kinds of methods: training-time (GRPO, SFT) and inference-time (prompting, decoding budget, OCR). Both raised accuracy, but mainly by changing the form of the output (e.g. producing shorter answers), not by improving the model's reasoning, for which we found no clear evidence.

Two possible directions to follow. First, some questions are simply beyond the model's reach—those where every sampled answer is wrong. Here the model's sample space contains no correct trajectory, so neither RL nor output-form shaping can help. Solving them would require capability brought in from outside; one option is distilling from a larger model into the smaller one, in which case the distilled traces should be filtered for correctness (keeping only chains the teacher answered correctly) and checked by human, so the student is not taught the teacher's mistakes.

Second, a closer look at model behavior. Our per-question analysis so far rests on only a few samples. Future work could inspect more outputs to surface possible negative patterns introduced by GRPO (e.g. weaker image reading), and use out-of-distribution datasets to better map the boundary of the model's ability.

Second, a closer look at model behavior. Inspecting outputs across our experimental conditions revealed no clear patterns tied to the training conditions; we leave open how to systematically analyze such behavioral effects, and how to combine that with out-of-distribution evaluation to map the model's ability boundary.

Acknowledgments

We thank the Data Science at Georgia Tech (DS@GT) CLEF competition group for their support. This research was supported in part through research cyber infrastructure resources and services provided by the Partnership for an Advanced Computing Environment (PACE) at the Georgia Institute of Technology, Atlanta, Georgia, USA[11].

Declaration on Generative AI

During the preparation of this work, the author(s) used Claude and GPT in order to check grammar and spelling. 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] P. Jindal, Best open-source vision language models of 2026, 2026. URL: https://www.labellerr.com/blog/top-open-source-vision-language-models/, accessed: 2026-03-07.

[2] Z. Lin, X. Chen, D. Pathak, P. Zhang, D. Ramanan, Revisiting the role of language priors in vision-language models, 2024. URL: https://arxiv.org/abs/2306.01879. arXiv:2306.01879.

[3] 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.

[4] 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.

[5] C. Team, Contextdrift at imageclef 2025 multimodal reasoning: Evaluating vlms' multimodal, multilingual and multidomain reasoning capabilities via thinking budget variations and textual augmentation, in: Working Notes of CLEF 2025, 2025.

[6] M. Team, Msa at imageclef 2025 multimodal reasoning: Multilingual multimodal reasoning with ensemble vision language models, in: Working Notes of CLEF 2025, 2025.

[7] D. Dimitrov, M. S. Hee, Z. Xie, R. J. Das, M. Ahsan, S. Ahmad, N. Paev, I. Koychev, P. Nakov, Overview of imageclef 2025 – multimodal reasoning, in: Working Notes of CLEF 2025, CEUR-WS, 2025.

[8] ImageCLEF Multimodal Reasoning Organizers, Imageclef 2025-2026 multimodal reasoning lab website, 2026. URL: https://mbzuai-nlp.github.io/ImageCLEF-MultimodalReasoning/2026/, accessed: 2026-05-26.

[9] DeepSeek-AI, text authors, Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL: https://arxiv.org/abs/2501.12948. arXiv:2501.12948.

[10] R. J. Das, S. E. Hristov, H. Li, D. I. 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), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 7547–7571. URL: https://aclanthology.org/2024.acl-long.420.

[11] PACE, Partnership for an Advanced Computing Environment (PACE), 2017. URL: http://www.pace.gatech.edu.

A. Gemma 4 E2B Prompt Templates

This appendix lists the prompt templates used for the Gemma 4 E2B student model. The filed Gemma submission used `reasoning_commit`. The other two prompts were used as diagnostic ablation variants reported in Tables 16 and 17.

A.1. `reasoning_commit`: Filed Gemma 4 E2B Submission Prompt

This prompt was used by the submitted Gemma 4 E2B + LoRA SFT system.

> Solve this multiple-choice question from the image. Reason in English in 3 to 5 short numbered steps. Strict rules:
>
> - Each step must follow logically from the previous.
> - DO NOT use "Wait", "Let me reconsider", "Actually", "On second thought", "Hmm", or any self-correction.
> - DO NOT loop back or re-analyze options you already discussed.
> - Commit to your reasoning and pick one answer.
>
> Even if the image labels options as 1/2/3/4, Cyrillic letters, Arabic letters, or other symbols, map them to A/B/C/D/E in order. The final answer must be exactly one Latin letter.
>
> Your response must end with exactly this line: `Answer: X` (where X is A, B, C, D, or E). If you run out of space, immediately write `Answer: X` with your best guess.

A.2. `reasoning_direct_choice`: Gemma 4 E2B Diagnostic Prompt

This prompt was used only for diagnostic ablations. Unlike `reasoning_commit`, it constrains the output space to A–D and therefore applies only to the evaluated four-choice subset.

> Solve this multiple-choice question from the image. Use direct reasoning in 3 to 5 short numbered steps: read the key facts from the image and question, identify the relevant concept or calculation, compare the plausible options, and choose the option that matches your conclusion.
>
> This test has exactly four answer choices. If the image labels options as 1/2/3/4, Arabic letters, Cyrillic letters, or other non-Latin symbols, map them by order: first option = A, second = B, third = C, fourth = D. Never choose E.
>
> After the numbered steps, write one short conclusion and the matching option text. Use exactly this ending format:
>
> `Final conclusion: <short conclusion>`
>
> `Chosen option: X. <option text>`
>
> `Answer: X`
>
> X must be one Latin letter from A, B, C, or D. No text after the final answer.

A.3. `reasoning_direct_choice_ocr`: Gemma 4 E2B OCR Diagnostic Prompt

This prompt was used only for OCR ablations with Gemma 4 E2B. It uses OCR text as auxiliary evidence while explicitly treating the original image as the source of truth.

Solve this multiple-choice question using the image and the OCR text. Use direct reasoning in 3 to 5 short numbered steps: read the question and answer options from the OCR when it is reliable, check the image for diagrams, graphs, formulas, tables, labels, and spatial relationships, identify the relevant concept or calculation, compare the plausible options, and choose the option that matches your conclusion.

OCR may contain recognition errors or extra text, so the original image is the source of truth.

This test has exactly four answer choices. If the image labels options as 1/2/3/4, Arabic letters, Cyrillic letters, or other non-Latin symbols, map them by order: first option = A, second = B, third = C, fourth = D. Never choose E.

After the numbered steps, write one short conclusion and the matching option text. Use exactly this ending format:

```
Final conclusion: <short conclusion>

Chosen option: X. <option text>

Answer: X
```

X must be one Latin letter from A, B, C, or D. No text after the final answer.

B. Qwen3.5 Prompt

This prompt is used in all Qwen3.5 experiments:

分析这道题的题目、所有选项和存在的视觉元素（如有），仅回答一个选项：A、B、C、D或E。

Translation: "Analyze this question's text, all options, and any visual elements (if present); answer with exactly one option: A, B, C, D, or E."

Except that EXAMS-V validation evaluations use a variant that inserts 进行思考，并作答 ("think it through, then answer") before the final clause.