zhai at ImageCLEF 2026 Task on Multimodal Reasoning: Prompt Design and Self-Consistency on a 4 B Open-Weight Thinking VLM

Jinghe Zhai[1], Huihui Chen[1,*] and Xinrui Qiu[1]

[1]Foshan University, Foshan, China

Abstract

Visual language models (VLMs) have made rapid progress on image question answering, document understanding and multimodal reasoning, yet the community still lacks a clear reference point for how far a single small open-weight thinking VLM, under a strict zero-finetuning constraint, can be pushed on multilingual examination-style reasoning. In this paper we examine this question through the ImageCLEF 2026 MultimodalReasoning Visual MCQ task, built on the EXAMS-V benchmark, which asks models to solve image-rendered multiple-choice questions across six languages. Using Qwen3-VL-4B-Thinking as the sole inference backbone — without finetuning, captioning module, or second VLM — we study the effect of prompt design, output extraction, and parallel inference-time scaling via self-consistency voting. Our results indicate that a single 4 B thinking VLM can reach competitive accuracy on this multilingual benchmark: on the official test set, our final submission attains an overall accuracy of **0.6150**. On our pre-submission slice, prompt design is the strongest screening signal we observe, while self-consistency adds a small but controlled decoding-level gain on top. The submission combines a letter-only Chain-of-Thought (CoT) prompt, five-sample self-consistency voting at low temperature, and a deterministic multi-stage answer extractor, all driven by a single Qwen3-VL-4B-Thinking backbone.

Keywords

Multimodal Reasoning, Vision-Language Model, Qwen3-VL-4B-Thinking, Self-Consistency, Visual Question Answering

1. Introduction

Recent ImageCLEF MultimodalReasoning systems have achieved strong results, but much of this progress has relied on closed-source large VLMs such as Gemini [1, 2], multi-model ensembles [2], or cascaded pipelines. This leaves a practical reproducibility gap: it remains unclear how far a small, single open-weight thinking VLM can be pushed when finetuning, auxiliary captioning modules, OCR side channels, and additional VLMs are all excluded. We study this question in the ImageCLEF 2026 MultimodalReasoning track [3, 4], focusing on a strict zero-finetuning setting with one 4 B-parameter thinking VLM.

ImageCLEF 2026 MultimodalReasoning [3] provides a direct test bed for that question: 1 117 image-rendered multiple-choice items covering six languages, with image-only input and single-letter output, and no transcribed-text side channel. On the 2025 edition, two 2025 submissions pursued complementary directions: ContextDrift swept the thinking budget of closed-source Gemini 2.5 Flash (serial inference-time scaling) [1], while MSA focused on prompt strictness and output normalisation [2]. Drawing on the prompt-strictness and inference-time-scaling intuitions of these two directions, we ask whether a single open-weight 4 B thinking VLM under a strict zero-finetuning constraint can yield strong results on the same multilingual MCQ task.

In this working notes paper, we investigate a zero-finetuning 4 B thinking-VLM pipeline on the ImageCLEF MultimodalReasoning 2026 task. We explore prompt design, output normalisation, and self-consistency decoding [5]. Focusing on the $\leq$ 7 B scale, we analyse the decoding-level accuracy gain

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

[*] Notebook for the ImageCLEF Lab at CLEF 2026.

[*]Corresponding author.

alger2020@163.com (J. Zhai); ddchh@163.com (H. Chen); qxruber@163.com (X. Qiu)

0000-0003-4465-2716 (H. Chen)

on an internal dev split sampled from EXAMS-V [6] and a set of design choices that did *not* stabilise at the probe scale, released as negative evidence. Our final submission attains 0.6150 on the official test set.

2. Related Work

Early approaches to visual question answering (VQA) relied on heavily engineered modular architectures, where separate components handled image encoding, question parsing and answer classification, and where text inside images was extracted by OCR pipelines and fused with visual features for downstream reasoning. Recent advances in vision-language models have replaced such hand-crafted systems with unified transformer-based architectures that jointly model vision and language, fusing modality-specific encoder embeddings into a shared representation space from which a decoder produces the output. Alongside state-of-the-art proprietary systems, representative recent open-weight multimodal models include Qwen3-VL [7] and InternVL 3 [8].

Large language models have substantially improved reasoning performance by scaling test-time computation, and a parallel line of work scales test-time compute by sampling multiple reasoning paths and aggregating them via self-consistency [5]. The combination of multimodal capabilities with long chain-of-thought post-training (via SFT or RL) has more recently been realised in thinking VLMs such as the QVQ-72B-Preview and Kimi-VL-Thinking [9] families, achieving strong results on benchmarks such as MMMU [10]. Prompt design has also proven essential for eliciting structured reasoning from foundation models, from few-shot demonstrations to chain-of-thought prompting [11].

On the 2025 edition of the ImageCLEF MultimodalReasoning track, two submissions define the design space we build on. The MSA team combines three Gemini models into a caption-then-reason cascade with strict letter-only prompts and few-shot demonstrations, leading the overall and 11 single-language sub-rankings [2]. ContextDrift sweeps the thinking budget of a single closed-source Gemini 2.5 Flash and reports a non-monotonic budget–accuracy relationship, leading the English and Bulgarian sub-rankings [1].

Our system builds on these advances in a complementary direction. We take a single open-weight 4 B thinking VLM as the only inference component, add no finetuning, captioning stage or second VLM, and tune only prompt design and decoding. The Structured CoT prompt, combined with output normalisation, establishes a 4 B greedy baseline, on top of which self-consistency [5] adds parallel-sample plurality voting. Unlike ContextDrift, which scales test-time compute *serially* along the thinking-budget axis on a single trace [1], we scale it *in parallel* along the sample-count axis with a fixed maximum thinking length; unlike MSA's multi-model caption-then-reason cascade [2], we keep the engineering surface to a single-model, zero-finetuning stack. The resulting system serves as a reference point at the ≤ 7 B scale.

3. Materials and Methods

3.1. Data

The ImageCLEF 2026 MultimodalReasoning task continues the EXAMS-V setting [6] and provides the official test set `test1117`, which contains 1 117 image-rendered multiple-choice items covering six languages (English, Bulgarian, Chinese, Croatian, Italian, Serbian). Each item provides a single image that contains the complete question stem, three to five answer options and the associated metadata (language, subject, grade level, presence of tables or figures); the output space is one capital letter in A–E. The official evaluation metric is exact-match accuracy; the language and visual-element distribution is summarised in Table 1.

Table 1
Language and visual-element distribution of test1117. “Diagram”, “Graph”, “Figure” and “Table” columns count items containing at least one element of the corresponding category; an item may be counted in multiple columns.

Language	N	Diagram	Graph	Figure	Table
English	500	308	72	10	203
Chinese	342	249	55	17	51
Bulgarian	110	35	5	57	18
Croatian	57	18	15	15	10
Italian	54	17	15	15	8
Serbian	54	17	15	15	8
Total	1 117	644	177	129	298

3.2. Methodology

The overall system follows a single image-to-letter end-to-end contract. Given the image of a question, the pipeline passes through inference backbone, prompt wrapping, sampling-based decoding with self-consistency voting, and answer extraction, and emits one letter in A–E. Figure 1 gives an overview; the remainder of this section describes these stages in order.

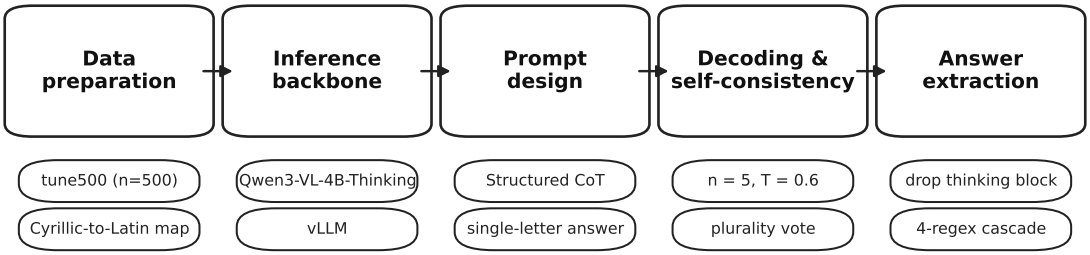

Figure 1: Experimental workflow.

3.2.1. Data preparation

For the main ablation and prompt design we sub-sample 500 items from the EXAMS-V validation split using `seed=42`, hereafter **tune500**; we never train or tune against the official test set. The language and subject distributions of tune500 are similar to those of the full validation split, and tune500 serves as a fixed development slice for local ablation and model selection. Per-language buckets become small after sub-sampling, so we treat tune500 as a development surface rather than an unbiased proxy of test1117.

A single answer-space normalisation is applied during extraction: the Bulgarian and Serbian items label their options with the first five letters of the Cyrillic alphabet, and we map them positionally to the Latin A/B/C/D/E so that the predicted letters can be compared directly against the gold labels. The mapping does not modify the question image and is implemented as a string substitution at the post-processing stage.

3.2.2. Inference backbone

The inference backbone is Qwen3-VL-4B-Thinking (4 B parameters, BF16). The checkpoint applies a thinking chat template that wraps the reasoning trace in a dedicated thinking block before producing the final answer. The model is served by vLLM [12].

The visual input receives one preprocessing step: images whose longest edge exceeds 1 024 pixels are downsampled to that limit with aspect ratio preserved, comfortably fitting within vLLM's 16 K context window (8 K reserved for generation).

3.2.3. Prompt design

The Structured CoT prompt has a single design goal: to pack the image of any item and its option set into a single user message, and to constrain the model to emit exactly one Latin capital letter in A–E. The complete prompt is an English instruction that unfolds in six steps; it contains no examples and is not accompanied by a separate *system* message.

Structured CoT prompt

```
You are an expert exam solver. The image shows a multiple-choice question that may
be in ANY language (English, Chinese, French, Bulgarian, Arabic, etc.).

Follow this process:
1. Read ALL text in the image carefully, including the question and every option.
2. If the text is not in English, first understand it in its original language.
3. Identify any visual elements: diagrams, graphs, tables, formulas, maps.
4. For EACH option, determine if it is correct or incorrect, and why.
5. Eliminate clearly wrong options first, then choose among remaining.
6. Output ONLY the letter of the correct answer (A, B, C, D, or E).
```

The prompt is used together with the thinking chat template introduced in §3.2.2, which wraps the model's reasoning in a thinking block; only the text after the block is treated as the final output. The "output only the letter" constraint therefore applies to this suffix, which in practice contains a single letter, sometimes followed by trailing punctuation or a short comment. The downstream extractor locates the letter on the suffix.

The same English instruction is applied uniformly to all six languages; language-aware prompt variants explored in §4.1 did not yield a stable improvement and are not used.

3.2.4. Decoding and self-consistency

Given the conversation packaged in §3.2.3 we collect $n{=}5$ independent generations at temperature $T{=}0.6$. The remaining sampling parameters are top_p=0.95 and top_k=20, identical to the defaults shipped in Qwen3-VL-Thinking's `generation_config.json`; the maximum generation length is max_tokens=8192, which contains a complete thinking block followed by the letter output. The five samples are submitted concurrently to a single vLLM instance.

For each item we cast the five extracted letters (§3.2.5) through a plurality vote to obtain the final prediction. With five samples over up to five options the vote can produce a tie (e.g. a 2–2–1 split). We break ties deterministically by indexing into the tied candidates with a SHA-256 hash of the image path, so tie resolution is fully deterministic given the five extracted letters; we use a hash rather than a fixed letter preference (e.g. "always pick A") to avoid coupling the tiebreak to a particular option distribution. Sampling itself is not seeded across vLLM workers, so the five generations are not bit-reproducible across runs; reproducibility therefore refers to the post-extraction voting and tiebreak layer.

3.2.5. Answer extraction

Each generated sample is a raw text string. The extractor first replaces the five Cyrillic option letters that may appear in non-English outputs with their visually corresponding Latin counterparts (positions 1–5: A, Be, Ve, Ge, De → A, B, C, D, E), so that all subsequent regexes can be written against the Latin alphabet alone. It then searches for the closing thinking tag and takes the text after the *last* such tag as the extraction surface; if that surface consists of a single Latin letter — the ideal case under Structured CoT prompt— the letter is returned in upper case.

Otherwise the extractor applies four regex patterns in cascade on the extraction surface:

```
(i)   ANSWER\s*(?:IS|:|=)\s*([A-E])\b
(ii)  \b([A-E])\s*[.)]*\s*$
(iii) (?:^|\s)([A-E])\s*[.):]
(iv)  (?:^|\s)([A-E])(?:\s|$|[.,;!?])
```

Any match returns immediately. If none of the four matches, a final fallback fires: a reverse scan from the end of the extraction surface returns the first character in A–E it encounters; if that too fails, the extractor returns the placeholder `"A"` and sets a `fallback` flag to `True`. The placeholder `"A"` exists only to keep the submission JSON structurally complete; its firing count is tracked separately so that it cannot inflate accuracy unnoticed.

The cascade is needed because Structured CoT prompt only requests "Output ONLY the letter" and the model in practice produces loosely formatted outputs such as "Answer: B."; the reverse-scan fallback resolves these without dropping to the placeholder. On the $1\,117 \times 5 = 5\,585$ generations of the test1117 submission, the `fallback` flag fired for 0 samples, i.e. every generation was resolved to an explicit A–E letter before the placeholder branch.

4. Experiments and Results

All experiments use the Qwen3-VL-4B-Thinking backbone described in §3.2; the prompt is the Structured CoT prompt (§3.2.3) by default, with prompt variants explored in §4.1. The evaluation metric is exact-match accuracy between the predicted letter and the gold letter, averaged over items. The official test set is evaluated only once at submission time, with no selective re-runs, and every number reported below matches the official leaderboard.

4.1. Pre-submission evaluations

Evaluations cover four inference-time axes — prompt, input resolution, decoder, and two "other" directions (few-shot, OCR-then-reason) — plus two parameter-efficient training paths. All screening runs use a 100-item EN-balanced slice of tune500 (seed 42; Wilson 95% CI ± 9 pp), summarised in Table 2; the decoding choice is then corroborated on the full tune500 (Table 3).

Table 2
Pre-submission inference-time screening on the 100-item slice. Each row varies exactly one axis off the baseline (P0 + greedy + 1 024 px); "Δ" is the signed deviation from the baseline of 71% on the same slice. Prompt codes: P0 = Structured CoT prompt (six-step structured CoT); P1 = Direct answer with no CoT; P2 = 8-step instruction; P3 = Language-aware (instruction in question language); P4 = Thinking-style prompt variants (single-paragraph reasoning, multiple revisions).

Prompt	Resolution	Decoder	Other	Acc.	Δ
P0	1 024 px	greedy	—	71	—
P1	1 024 px	greedy	—	66	−5
P2	1 024 px	greedy	—	64	−7
P3	1 024 px	greedy	—	66	−5
P4	1 024 px	greedy	—	64–65	−7 to −6
P0	1 280 px	greedy	—	69	−2
P0	2 048 px	greedy	—	66	−5
P0	1 024 px	SC= 3 (T=0.6)	—	68	−3
P0	1 024 px	**SC= 5 (T=0.6)**	—	**72**	**+1**
P0	1 024 px	greedy	Few-shot (k=2)	59	−12
P0	1 024 px	greedy	OCR-then-reason (R7)	67	−4

Prompt. P1–P4 all drop 5–7 pp below P0 (Table 2); the six-step structured CoT carries most of the prompt-side signal at 4 B scale.

Table 3
Decoding selection on tune500. Both paths use Structured CoT prompt (P0); see §3.2.4 for decoding parameters.

Decoder	tune500 Acc. (%)	Δ (pp)
greedy	61.0	—
SC= 5 (T=0.6, submitted)	**69.4**	**+8.4**

Resolution. Raising the longest-edge cap to 1 280 / 2 048 px monotonically loses 2 / 5 pp; we lock preprocessing at 1 024 px.

Decoder. Among decoders, only SC= 5 trends positive (+1 pp); SC= 3 does not. Option-position permutation voting (2/4-way) also failed to improve over P0 on the same slice and is omitted from Table 2 for brevity. SC= 5 is the only candidate carried to tune500.

Few-shot and OCR-then-reason. Few-shot (k=2) collapses by 12 pp, plausibly from long multi-image context at 4 B scale; OCR-then-reason (Tesseract transcript prepended to the prompt) loses 4 pp. Both discarded.

Training. Two parameter-efficient training paths (GRPO; LoRA SFT) were also screened. Neither improved over the inference-only baseline.

On tune500, switching from greedy to SC= 5 lifts accuracy from 61.0% to 69.4% (Table 3; Δ=+8.4 pp, Wilson ±4 pp at n=500). Both numbers are from the same vLLM backend; see §3.2.4. The direction matches the slice-level +1 in Table 2, so we carry P0 + SC= 5 at T=0.6 to the official test set.

4.2. Official submission results

Submitting Structured CoT prompt + SC= 5 on the 1 117-item official test set, we obtain an overall accuracy of **0.6150** (Table 4). The four intermediate languages cluster within ∼ 4 pp, while the EN–ZH spread is 14 pp (Wilson half-widths ±9–13 pp at $n \leq 110$); this is consistent with the thinking block being emitted predominantly in English and with tune500 itself skewing English-heavy. Predicted-letter counts are 313 (A), 277 (B), 267 (C), 257 (D), 3 (E), consistent with EXAMS-V's mostly four-option format. The 7.9 pp gap between tune500 (69.4) and the test set (61.5) is examined below.

Table 4
Per-language accuracy on the official test set for Structured CoT prompt + SC= 5. "n" is the number of items in that language. Bulgarian and Serbian predicted letters have been normalised from Cyrillic to Latin by the positional mapping of §3.2.5.

	EN	BG	HR	IT	SR	ZH	**Overall**
n	500	110	57	54	54	342	**1 117**
Acc.	0.6760	0.6182	0.6140	0.5926	0.5741	0.5351	**0.6150**

4.3. Error Analysis: Subject-Level Failure Patterns

Once the official test labels became available, we ran a per-subject breakdown of accuracy on the same submission file.

Figure 2 shows substantial per-subject spread: the strongest bucket (Fine Arts) is roughly 17 pp above the weakest (Physics); the three single-discipline science buckets (Physics, Chemistry, Biology) cluster around or slightly below the overall mean, whereas the cross-disciplinary "Biology + Chemistry" and "Physics + Chemistry" buckets sit ∼ 8 pp above it. The 4 B Qwen3-VL-4B-Thinking backbone shows its largest weaknesses on single-discipline physics and chemistry, where heavy use of equations and figure-based quantitative reasoning is most likely.

This pattern suggests that the observed errors are unlikely to be explained by multilingual reading alone, and also involve visually grounded scientific reasoning under small-model capacity. In particular, physics and chemistry questions often require reading formulas, aligning diagram elements with the

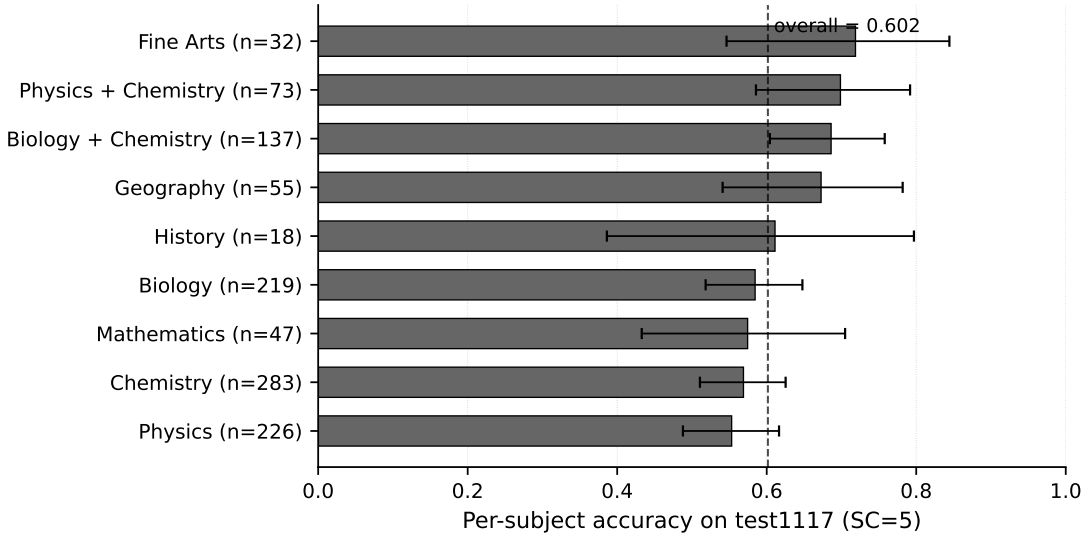

Figure 2: Per-subject accuracy of Structured CoT prompt + SC$=5$ on the official test set, with 95% Wilson confidence intervals. Subject labels span six languages and are normalised by keyword matching; only buckets with $n \geq 18$ are shown. The dashed line marks the overall mean across these buckets.

question text, and carrying out multi-step quantitative or symbolic inference. Such cases are harder to resolve with prompt-only inference than questions that rely mainly on conceptual recognition or direct visual-text matching. Therefore, although self-consistency improves the decoding robustness of the submitted system, it does not fully address errors caused by weak visual-symbolic grounding or insufficient domain-specific reasoning.

5. Conclusion

Our final system, Structured CoT prompt + SC$=5$, reaches 0.6150 overall accuracy on the $1\,117$-item official test set, with the multi-stage extractor never falling back to the placeholder across $5\,585$ generations. On the 500-item pre-submission split, SC$=5$ at $T = 0.6$ lifts greedy accuracy from 61.0% to 69.4% ($+8.4$ pp, Wilson ± 4 pp). The other directions screened in §4.1 (direct prompt, eight-step reasoning, language-aware adaptation, OCR-then-reason) all fall within ± 9 pp Wilson of the baseline, giving no stable positive signal. The post-submission per-subject reading (Figure 2) shows a ~ 17 pp spread, with cross-disciplinary buckets sitting ~ 8 pp above single-discipline ones.

Our analysis has four limitations. The main ablation runs on n=500 with a single seed, so the $+8.4$ pp SC gain is an engineering-direction observation rather than a statistically conclusive gap; all experiments use a single 4 B backbone, so conclusions do not extrapolate to 7–72 B thinking VLMs; the cross-language pattern of Table 4 is reported descriptively because per-language sample sizes are not balanced; and the $5\,585$-generation budget was not measured against wall-clock, energy or inference cost. Future work includes stratified per-language evaluation on a balanced split, multi-seed re-runs to characterise SC variance, and a comparison against a 7 B or larger thinking-style VLM under the same Structured CoT prompt + SC$=5$ recipe.

Acknowledgments

We thank the ImageCLEF 2026 MultimodalReasoning organisers, the vLLM project, the Qwen team, and the MBZUAI EXAMS-V authors for releasing their resources.

Declaration on Generative AI

During the preparation of this work, the author used GitHub Copilot Chat (Anthropic Claude family) in order to: Grammar and spelling check, paraphrase and reword, and code refactoring. After using these tool(s), the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] V. T. Krazheva, D. Markova, D. I. Dimitrov, I. Koychev, P. Nakov, ContextDrift at ImageCLEF 2025 multimodal reasoning: Evaluating VLMs' multimodal, multilingual and multidomain reasoning capabilities via thinking budget variations and textual augmentation, in: CLEF 2025 Working Notes, CEUR Workshop Proceedings, 2025, pp. 2486–2499. URL: https://ceur-ws.org/Vol-4038/paper_194.pdf, paper 194.

[2] S. Ahmed, M. Younes, A. Moustafa, A. Allam, H. Moustafa, MSA at ImageCLEF 2025 multimodal reasoning: Multilingual multimodal reasoning with ensemble vision-language models, in: CLEF 2025 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, 2025, pp. 2275–2283. URL: https://ceur-ws.org/Vol-4038/paper_176.pdf, paper 176.

[3] 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.

[4] B. Ionescu, H. Müller, D.-C. Stanciu, A. Radu, R.-G. Bolborici, M. Negru, A.-F. Ene, V.-M. Vasilescu, A.-A. Nicolae, L.-D. Ştefan, M.-G. Constantin, M. Dogariu, A.-G. 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.-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.-N. 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.

[5] 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, in: International Conference on Learning Representations (ICLR), 2023. URL: https://openreview.net/forum?id=1PL1NIMMrw.

[6] R. J. 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), Association for Computational Linguistics, 2024, pp. 7768–7791. URL: https://aclanthology.org/2024.acl-long.420. doi:10.18653/v1/2024.acl-long.420.

[7] Qwen Team, Qwen3-VL technical report, 2025. URL: https://arxiv.org/abs/2511.21631. arXiv:2511.21631.

[8] J. Zhu, W. Wang, Z. Chen, Z. Liu, S. Ye, L. Gu, H. Tian, Y. Duan, W. Su, et al., InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models, Technical Report arXiv:2504.10479, arXiv, 2025. Technical Report.

[9] Kimi Team, Kimi-VL Technical Report, Technical Report arXiv:2504.07491, arXiv, 2025. Technical Report.

[10] X. Yue, Y. Ni, K. Zhang, T. Zheng, R. Liu, G. Zhang, S. Stevens, D. Jiang, W. Ren, et al., MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI, in:

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9556–9567.

[11] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, in: Advances in Neural Information Processing Systems, volume 35, 2022, pp. 24824–24837. URL: https://openreview.net/forum?id=_VjQlMeSB_J.

[12] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. E. Gonzalez, H. Zhang, I. Stoica, Efficient memory management for large language model serving with PagedAttention, in: Proceedings of the 29th ACM Symposium on Operating Systems Principles (SOSP), 2023, pp. 611–626. doi:10.1145/3600006.3613165.