Nika at ImageCLEF 2026 Task on Multimodal Reasoning: More Tokens, Fewer Trees — Test-Time Scaling for Small VLMs on Multilingual Visual MCQ

ImageCLEF Lab at CLEF 2026

Spiros Baxevanakis[1,†], Peng-Jian Yang[1,†]

[1]University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands

Abstract

Test-time scaling (TTS) reliably improves reasoning in large language models, but whether it transfers to small open vision-language models remains unclear. We examine this on EXAMS-V, a multilingual visual multiple-choice benchmark, comparing self-consistency, describe-then-reason with PRM-guided beam search, and two post-hoc selectors across Qwen2.5-VL-7B-Instruct and Qwen3.5-4B. What matters is the conditions under which TTS runs, not the search or verification machinery. The largest factor is parseability: an early prompt format left many chains reasoning correctly yet never committing to an answer letter, which a standard answer cue and a guided repair step largely remove. A larger decoding budget removes the rest: raising the per-chain token limit from 1k to 2k recovers 3.7 pp, whereas sampling more chains (8 to 16) adds only 0.15 pp. Once chains have room to finish, elaborate methods contribute little: PRM-guided beam search trails plain self-consistency by 0.39 pp at over eight times the cost, and neither a training-free generative critic nor a trained multimodal PRM beats majority vote across both policies. The largest gain comes instead from the policy model itself (+11.4 pp). Our best configuration reaches 84.1% on the held-out ImageCLEF 2026 test split, ranking first on the Visual MCQ leaderboard.

Keywords

test-time scaling, vision-language models, multimodal reasoning, self-consistency, process reward models,

1. Introduction

Multimodal reasoning remains a challenge for vision–language models (VLMs) whenever questions combine structured visual content with multi-step inference across languages. Real exam questions are a natural stress test of this combination, and the EXAMS-V benchmark [1] spans thirteen languages and twenty school subjects rendered as text-in-image panels. The 2026 ImageCLEF multimodal reasoning shared task [2, 3] fixes the practical constraints: a single A40 GPU, open-weight policies with at most 7B parameters, and a reproducibility mandate.

The natural question is whether the inference-time techniques that advanced text-only language models [4, 5, 6] transfer to open VLMs under this budget. The evidence is cautionary: self-refinement degrades open-source VLMs at this scale [7], long sequential "thinking" often underperforms parallel sampling on small policies [8, 9], and discriminative reward models trained on mathematics fail to transfer to humanities and non-English content [10, 11]. A viable pipeline must externalise selection, prefer parallel sampling to long single-chain reasoning, and avoid iterative self-refinement loops.

We investigate whether describe-then-reason with PRM-guided search beats flat self-consistency; how accuracy scales along two axes (chain count and per-chain token budget) under a single-A40 envelope [6]; where TTS helps and fails across subjects and languages [7]; and whether a training-free generative critic can beat majority vote on low-agreement questions, with a cross-model PRM [12] controlling for sycophantic self-scoring [13]. The latter extends generative-verifier findings [14, 15, 16] to a multimodal, multilingual setting under a $\leq$7B budget not previously studied to our knowledge.

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

[†]These authors contributed equally.

spiros.baxevanakis@student.uva.nl (S. Baxevanakis); lesterpjy@gmail.com (P. Yang)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Summary of findings. Across all four dimensions, the policy substrate and engineering details (parser format, decoding budget) dominate any search or selector method. Our best configuration, Qwen3.5-4B at SC-N=16 with a 2,048-token budget, reaches 81.6% on the full validation set. [1]

2. Related Work

Test-time scaling: theory and limits. Snell et al. [6] formalize TTS as a two-axis allocation (width: parallel chains; depth: per-chain revision or search), proving compute-optimal allocation can match naive Best-of-N at less cost [17, 18]. Setlur et al. [19] show the verifier advantage grows with reasoning depth, motivating an external selector; converging findings [8, 9, 20, 7] argue for parallel sampling and against self-correction at our policy scale.

Multimodal verification and PRM generalization. A *process reward model* (PRM) assigns each reasoning step a probability $P(+)$ that the step is correct [5]; it is useful only when $P(+)$ sharply separates correct from incorrect steps. Two multimodal PRMs supersede VisualPRM-8B [21]: Athena-PRM-7B [22] and Qwen-VL-PRM-7B [12], which separates perception from reasoning errors. On the text side, discriminative math-trained PRMs fail out of distribution [10, 11, 23] while generative verifiers transfer more reliably [14, 15, 16]. Prior EXAMS-V work [24, 25] informs our scaffold; on the search axis, PRM-BAS [26] dominates step-level Best-of-N at matched compute.

3. Method

We evaluate two pipelines: describe-then-reason with PRM-guided beam search and a three-way selector contrast, and flat self-consistency with a large token budget.

Policy and data. We evaluate two open-weights policies: `Qwen/Qwen2.5-VL-7B-Instruct` (7B parameters, bfloat16) for the baseline and ablation configurations, and `Qwen/Qwen3.5-4B` (4B parameters, bfloat16), a smaller but newer-generation policy, for the headline configuration. We cap image resolution at `max_pixels`=1,003,520 (≈4k vision tokens after Qwen smart-resize), balancing legibility of text-in-image exam panels against context budget: at `max_model_len`=16,384, this leaves ∼12k tokens for the prompt and generated reasoning. Inference uses vLLM [27] with batched parallel decoding (`SamplingParams(n=N)`) and automatic prefix caching. All experiments run on a single A100-40 GB GPU, which matches the shared task's A40 memory constraint. We evaluate on the full EXAMS-V validation split (4,651 questions, 11 languages) for headline numbers and a 200-question (language, subject)-stratified subset (seed 42, minimum ten per language) for ablations whose full-validation cost is prohibitive.

Describe-then-reason. Following Ahmed et al. [24], the describe-then-reason pipeline first samples N image-grounded descriptions of the question at T=0.7. For each description, with the image removed, the policy generates M reasoning chains in the flat variant, or reasoning step-by-step via PRM-BAS-style beam annealing [26] in the search variant: an initial beam of width B_0 is expanded with B continuations per surviving beam, each step is scored by Qwen-VL-PRM-7B [12], which retains access to the original image, and the beam width is halved whenever the score range among the top beams exceeds a threshold τ. Beams terminate on the answer cue or at a maximum depth d. Across all (description × terminal beam) candidates, a final selector picks the answer. In our experiments the flat variant uses N=4, M=4 (20 calls/q, bracketing SC-N=8 from above to test whether structured decomposition helps at higher compute); the search variant halves descriptions to N=2 because beam expansion is expensive, and adopts the hyperparameters reported in Hu et al. [26]: B_0=4, B=2, τ=0.05, d=6 (∼13 calls/q).

[1] Code available at: https://github.com/lesterpjy/tts-small-vlm.

Selectors. Three selectors are applied post hoc to the same chain pool. *Majority vote*: most common answer letter with log-probability tie-breaking (no extra inference). A *training-free generative critic* inspired by Zhang et al. [15] scores each chain on three axes (Appendix F). A *discriminative PRM* (Qwen-VL-PRM-7B; 12), a separately trained model, scores each chain in one-shot mode (Appendix E), controlling for sycophantic self-preference [13]. Both models co-reside on the A100-40 GB at bfloat16. Each selector uses a skip rule [28]: high-confidence majorities ($\geq 5/8$ agreement) pass through; only the weak ($\leq 4/8$) and all-unparseable tiers are rescored.

Self-consistency. The simpler pipeline samples $N \in \{8, 16\}$ chains from Qwen3.5-4B at $T{=}0.7$ with `max_new_tokens`$=2048$, conditioned on a CoT prompt ending with the MMMU closer [29, 30]. A regex extracts the answer letter (mapping Cyrillic labels to A–E for multilingual coverage). The final answer is the majority letter across parseable chains, with log-probability tie-breaking. We additionally sweep $T \in \{0.3, 0.5, 0.7, 0.9\}$ at full validation scale.

Guided parse repair. When every chain for a question fails to produce an answer letter (2–18% of questions depending on budget), each chain's reasoning is concatenated with the cue `\n\nAnswer:` and a single token is decoded under vLLM's `guided_choice` constraint over $\{A, B, C, D, E\}$. The majority across all committed letters (original plus repaired) is the final answer. This eliminates parse failures by construction.

Stratification and evaluation. We partition accuracy by subject and language; per-subject results are reported only for cells with $n \geq 50$. Pairwise method comparisons use McNemar's paired test with Bonferroni correction for the number of reported comparisons.

4. Experiments

We evaluate seventeen configurations across two policies, three search strategies, three selectors, and two scaling axes. Table 1 summarises them; the remaining sections report findings ordered by headline result rather than by configuration. Unless noted, all headline numbers are on the full EXAMS-V validation set ($n{=}4{,}651$); configurations marked "dev-200" use a 200-question stratified subset.

5. Results and Discussion

We report results grouped by headline finding.

5.1. The TTS progression and the parser artifact

The three-stage TTS progression (zero-shot, chain-of-thought, self-consistency) holds for both policies at full validation scale (Table 2). Each step yields a significant gain under Qwen2.5-VL-7B, and the progression is more pronounced under Qwen3.5-4B, where the full span from zero-shot to SC-$N{=}8$ (2k) exceeds 24 pp.

However, the relative contribution of each stage shifts with the policy. Under Qwen2.5, the CoT-to-SC margin is modest ($+3.3$ pp), suggesting that eight samples add limited signal once CoT prompting is in place. Under Qwen3.5, the margin widens markedly: CoT accounts for roughly half the total gain, and the remaining half comes from parallel sampling and the token-budget increase (Section 5.3).

A parser artifact masked part of the CoT gain. An early prompt format left $\sim$9% of development chains unparseable. Switching to the MMMU-standard closer [29, 30] cut parse failures below 2% and raised dev accuracy by $\sim$6 pp; the fix corroborates at val scale ($+0.82$ pp on Q2.5 SC-$N{=}8$). The methodological lesson: what looks like reasoning failure can be an extraction failure, and TTS studies that do not control for parseability risk attributing engineering artifacts to the scaling method.

Table 1
Experimental configurations. "Policy" is Q2.5=Qwen2.5-VL-7B-Instruct, Q3.5=Qwen3.5-4B. Calls/q is the typical end-to-end VLM forward count. Header rows group configurations by theme. All full-validation configurations use n=4,651.

Method	Policy	Search / sampling	Selector	Calls/q	Scale
Baselines					
B0 zero-shot	Q2.5	guided_choice, 1 sample	—	1	val-full
B1 chain-of-thought	Q2.5	1 sample, MMMU closer	post-hoc regex	1	val-full
B2 self-consistency	Q2.5	N=8, T=0.7, 1024-tok	majority vote	8	val-full
B0 zero-shot	Q3.5	guided_choice, 1 sample	—	1	val-full
B1 chain-of-thought	Q3.5	1 sample, MMMU closer	post-hoc regex	1	val-full
B2 self-consistency	Q3.5	N=8, T=0.7, 1024-tok	majority vote	8	val-full
Search-strategy contrasts (dev-200)					
S-DTR	Q2.5	DTR N=4,M=4	majority vote	20	dev-200
S-PRM-BAS	Q2.5	DTR N=2 + PRM-BAS (B_0=4, B=2, d=6)	majority vote	∼13	dev-200
Scaling (Q3.5, val-full)					
SC 1k-tok	Q3.5	N=8, T=0.7, 1024-tok	majority vote	8	val-full
SC 2k-tok	Q3.5	N=8, T=0.7, 2048-tok	majority vote	8	val-full
SC N=16	Q3.5	N=16, T=0.7, 2048-tok	majority vote	16	val-full
Temp. sweep	Q3.5	N=8, $T\in\{.3, .5, .7, .9\}$, 2k	majority vote	8	val-full
Selectors (full val, both pools)					
R1 generative critic	Q2.5	SC pool	GM-PRM-style critic	+3/ch	val-full
R2 Qwen-VL-PRM	Q2.5	SC pool	PRM 1-shot + skip	+1/ch	val-full
R1 generative critic	Q3.5	SC-2k pool	GM-PRM-style critic	+3/ch	val-full
R2 Qwen-VL-PRM	Q3.5	SC-2k pool	PRM 1-shot + skip	+1/ch	val-full
Best configuration					
S-best	Q3.5	N=16, T=0.7, 2048-tok	majority + guided repair	≤16	val-full

Table 2
TTS progression on full validation (n=4,651). All Q2.5 pairs are Bonferroni-significant at α=0.017. Q3.5 SC rows show the effect of doubling the token budget. P.-fail: fraction of questions with all chains unparseable.

Policy	Method	Acc.	95% CI	P.-fail	C/q
Q2.5	Zero-shot	56.83	[55.4, 58.2]	0.0%	1
Q2.5	CoT	63.10	[61.7, 64.5]	2.0%	1
Q2.5	SC-N=8 (1k)	66.42	[65.0, 67.8]	0.2%	8
Q3.5	Zero-shot	57.11	[55.7, 58.5]	0.0%	1
Q3.5	CoT	69.66	[68.3, 71.0]	16.2%	1
Q3.5	SC-N=8 (1k)	77.81	[76.6, 79.0]	18.4%	8
Q3.5	SC-N=8 (2k)	81.49	[80.4, 82.6]	2.7%	8
Q3.5	**SC-N=16 (2k)**	**81.64**	**[80.5, 82.7]**	**1.9%**	16

Under Q3.5, every step is significant: CoT over zero-shot gains +12.6 pp (non-overlapping CIs), the 1k→2k budget increase recovers +3.7 pp (171 additional correct answers), and the N=8 → N=16 step adds only +0.15 pp (7 questions, within sampling noise). Q3.5 CoT itself suffers 16.2% parse-fail from the same truncation that affects SC (18.4% at 1k), so the reported 69.66% underestimates single-chain accuracy and the CoT-to-SC(1k) gain conflates truncation absorption with diversity benefit.

5.2. Structured search underperforms flat sampling

Dev-200 CIs overlap between PRM-BAS, DTR, and flat SC-N=8 (Table 11, Appendix H). A val-scale run of PRM-BAS under Qwen3.5-4B (N=2, B_0=4, B=2, τ=0.05, d=6; n=4,319, 93% of validation) resolves the ranking in favour of SC: PRM-BAS reaches 80.74% against SC majority at 81.13% on the same questions (−0.39 pp). Of the 14.6% of questions where PRM-BAS produces a different answer than SC, 233 are corrections and 250 are regressions, for a net loss of 17 questions. Two bottlenecks interact: in the DTR pipeline the reasoning stage operates on the text description alone (no image

Table 3
Two-axis scaling on Qwen3.5-4B, MMMU closer. *Top*: dev-200 exploration. *Bottom*: val-full (n=4,651) validation of the key points. Dev-200's +2 pp N-scaling signal collapses at val-full.

Scale	Config.	Acc.	P.-fail	s/q
Dev-200 exploration				
dev	N=8, T=0.7, 1024-tok	∼65	18.0%	∼7
dev	N=8, T=0.7, 2048-tok	80.0	2.5%	∼14
dev	N=8, T=0.7, 3072-tok	81.0	1.5%	∼21
dev	N=16, T=0.7, **2048-tok**	**82.0**	**2.5%**	**∼25**
Val-full validation (n=4,651)				
val	N=8, T=0.7, 1024-tok	77.81	18.4%	8.8
val	N=8, T=0.7, 2048-tok	81.49	2.4%	11.8
val	N=16, T=0.7, **2048-tok**	**81.64**	**1.9%**	**14.6**
Temperature sweep (val-full, N=8, 2048-tok)				
val	T=0.3	80.84	2.7%	11.2
val	T=0.5	81.01	2.6%	11.5
val	T=0.7	**81.49**	**2.4%**	**11.8**
val	T=0.9	81.04	3.0%	12.3

access), so visual detail lost at the description step cannot be recovered downstream; beam search compounds this loss because the PRM scoring each step (though it retains image access) lacks sufficient discriminative signal to steer search toward the missing detail.

Diagnosis. Hu et al. [26] show that PRM-guided beam-annealing search dominates flat Best-of-N on multimodal math benchmarks. Our result inverts this: on multilingual visual MCQ, PRM-BAS underperforms flat SC by -0.39 pp at $8.7\times$ the cost. Two factors explain the discrepancy. First, the PRM's per-step P(+) saturates (0.962 on descriptions, 0.849 on reasoning), collapsing search diversity: 71.9% of val questions have all surviving beams agreeing on the same letter (Shannon entropy $0.259/2.322$ bits, 89% reduction from uniform); the $8.7\times$ cost yields near-zero answer diversity. Second, slicing by SC agreement tier (Appendix A) reveals that on high-confidence questions ($\geq 7/8$, 91.8% SC accuracy), PRM-BAS regresses by -2.16 pp (48 corrections vs. 119 regressions), dominating the overall deficit. Only on the 108 all-unparseable questions does PRM-BAS produce a clear positive (37 correct, 34.3%). The P(+) asymmetry identifies reasoning as the structural bottleneck (discussed in Section 7).

5.3. Token budget dominates chain count

We sweep two scaling axes (per-chain token budget and number of sampled chains), first on the 200-question development subset, then on the full validation set (Table 3).

Token budget dominates chain count. Doubling the budget from 1k to 2k tokens recovers +3.7 pp (Figure 1a): chains generate valid reasoning but are truncated before emitting an answer letter, producing parse failures rather than reasoning failures. Doubling chains (N=8 → 16) adds only +0.15 pp. The asymmetry is structural: ∼60% of questions already reach strong agreement at N=8, while the weak-agreement tail ($\leq 4/8$, 23%) has below-chance accuracy that additional sampling does not correct. (Appendix A). For compute-constrained TTS, ensuring chains run to completion matters more than sampling more of them.

Temperature sensitivity. An inverted-U optimum at T=0.7 replicates from dev-200 to full validation (Figure 1b), but the spread compresses to 0.65 pp at scale, suggesting the policy's sampling diversity is already near optimal.

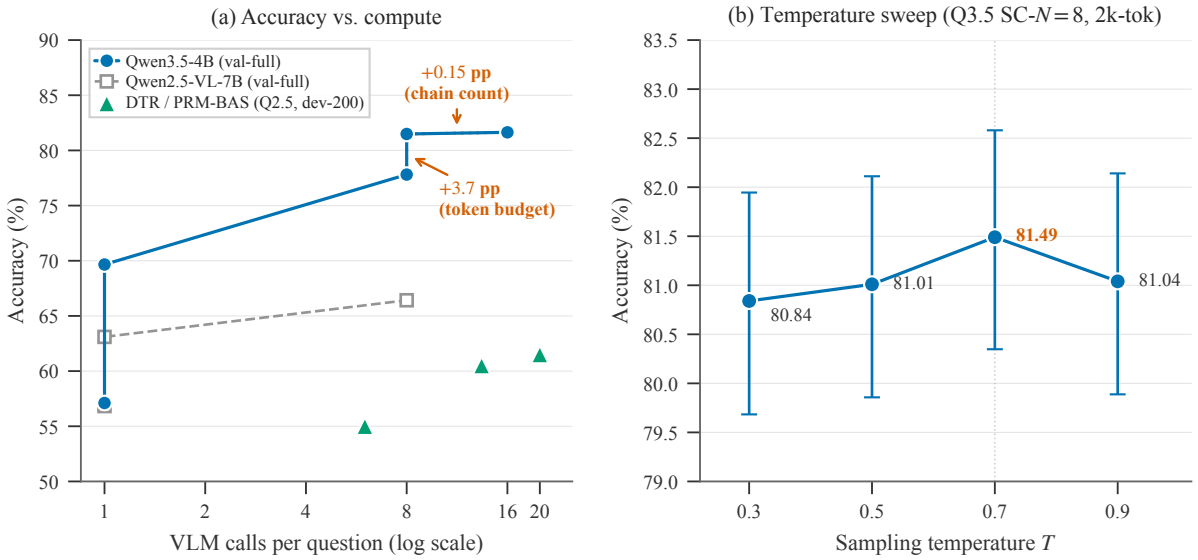

Figure 1: (a) Accuracy vs. VLM calls per question. At $N{=}8$, the token-budget arrow marks $+3.7$ pp from 1k→2k tokens at constant call count; the chain-count arrow marks $+0.15$ pp from doubling N at constant budget. DTR search points (dev-200, Q2.5) sit below flat SC at higher compute. (b) Temperature sweep on Q3.5 SC-$N{=}8$ (2k-tok, val-full). Error bars are 95% Wilson CIs.

5.4. Guided repair closes the parse-fail residue

After the budget fix, val SC-$N{=}8$ still has 124 all-unparseable questions (2.67%); the test split ($n{=}1{,}117$) has 82 (7.34%), elevated because the test mix is 44.8% English and 30.6% Chinese. Guided repair drives both rates to zero at one decoded token per affected chain. The repair targets valid reasoning that was never committed to an answer letter, a failure mode that chain scaling cannot reach, recovering more than doubling N ($+0.15$ pp val) at a fraction of the compute.

5.5. Policy choice dominates search and selection

Switching from Qwen2.5-VL-7B to Qwen3.5-4B at matched settings yields $+11.4$ pp (Table 2, SC-$N{=}8$ rows); with the budget increase the gap widens to 15.1 pp and all eleven languages improve by $\geq$3 pp (Table 10). Qwen3.5-4B is *smaller*; the gain reflects newer-generation training, consistent with Ahmadpour et al. [7]. At our budget scale, upgrading the policy dominates any inference-time strategy.

5.6. Selectors yield a null result on both pools

We test whether a post-hoc selector improves on majority vote. Two selectors are compared: a *training-free generative critic* (three rubric axes; Appendix F) and a *discriminative PRM* (Qwen-VL-PRM-7B; one-shot scoring). Both apply a skip rule preserving high-confidence majorities. We evaluate on two pools of different baseline strength (66% Q2.5, 81% Q3.5) to control for ceiling effects (Table 4).

Cross-pool replication. The null replicates across both policies: on Q2.5 (66%) the critic is an exact null and the PRM gains $+0.45$ pp (n.s.); on Q3.5 (81%) both turn negative (PRM: 51 corrected vs. 55 regressed; critic: 37 vs. 59). As pool accuracy grows, the majority-wrong tail shrinks and a near-balanced selector's net effect turns negative (Appendix B).

Why both selectors fail. The critic exhibits self-recognition bias [13]: 40% of chains receive the modal score 0.75, and the correct/incorrect gap is negligible (0.713 vs. 0.687; Appendix B). The PRM

Table 4
Selectors on full validation (n=4,651). Both selectors are tested on two candidate pools (Q2.5 SC-N=8 and Q3.5 SC-N=8 at 2048 tokens). All use the skip-on-high rule; R2 uses one-shot PRM scoring. The Q2.5 pool predates the MMMU closer (Section 5.1); its majority anchor is therefore 65.60% rather than the post-closer 66.42% in Table 2.

Pool	Selector	Acc.	Δ
Q2.5	SC majority (anchor)	65.60%	—
Q2.5	R1 generative critic	65.60%	+0.00 pp
Q2.5	R2 PRM (1-shot)	66.05%	+0.45 pp[n.s.]
Q3.5	SC majority (anchor)	81.49%	—
Q3.5	R2 PRM (1-shot)	81.40%	−0.09 pp
Q3.5	R1 generative critic	81.01%	−0.47 pp

[n.s.] Q2.5 R2: McNemar χ^2=1.80, p≈0.18. Q3.5: both negative, not tested.

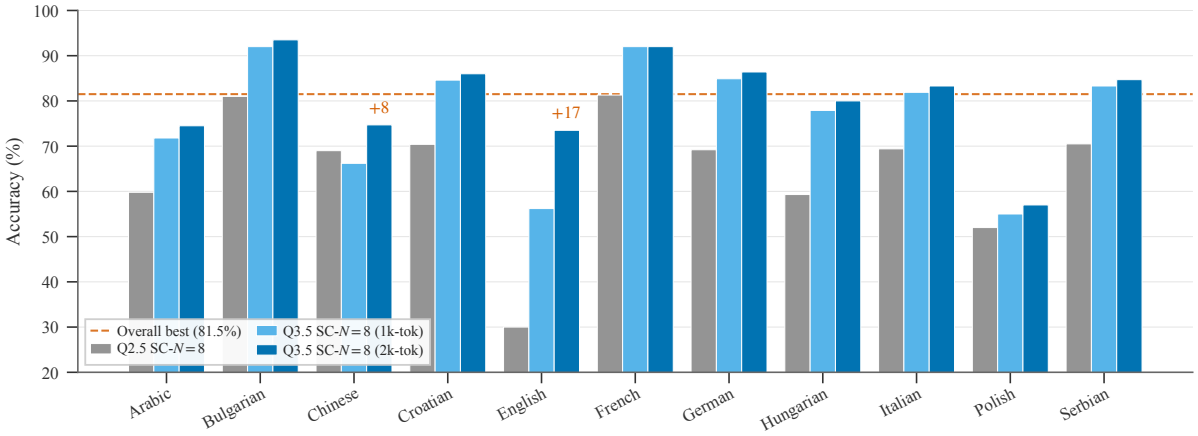

Figure 2: Per-language accuracy on full validation. Red annotations mark the accuracy gain from the budget increase (1k→2k) for languages where it exceeds +5 pp. The dashed line is the overall best (81.5%).

improves the correction rate (37.6% vs. 31.8%), partially mitigating self-preference, but cannot distinguish right from wrong overrides: the score gap between its top-ranked chain and runner-up is 0.049 on corrections and 0.048 on degradations, yielding a net effect indistinguishable from noise.

5.7. Stratified analysis

Figure 2 and Table 10 (Appendix G) report per-language accuracy under both policies, with the Q3.5 column split by token budget to isolate the effect of the budget increase.

The budget increase concentrates gains on English and Chinese. English gains +17.3 pp and Chinese +8.5 pp when the budget doubles; under Q2.5, over a third of English questions had all eight chains unparseable. Most other languages gain only 1–2 pp and French (92%) is unchanged, confirming a targeted rather than general effect.

Persistent below-average strata. Chinese, Polish, English, and Arabic remain below 81.5% even after the budget increase. Under Q2.5, SC *regresses* on Chinese by 2.8 pp: a chain-agreement audit (Appendix C) shows 22.3% of Chinese questions reach unanimous 8/8 agreement (vs. ≤15% elsewhere). SC gains power when chains err independently; when errors are highly correlated, the vote amplifies the shared mistake. Languages benefiting most from SC (Bulgarian, Croatian, Hungarian, Serbian) sit at mid-accuracy where the independence condition holds.

Per-subject patterns. Per-subject accuracy (Appendix D) reveals a split: STEM subjects benefit most from TTS (Physics $+14$ pp, Mathematics $+20$ pp over zero-shot), while the weakest Q3.5 subjects (Social 43%, Islamic Studies 57.8%) are text-heavy. Using subject as a proxy for visual complexity (EXAMS-V lacks content-type annotations), the pattern suggests two regimes: in STEM, reasoning errors vary across chains and majority vote aggregates them away; in the weakest subjects, the bottleneck is missing domain knowledge that no amount of sampling can compensate.

5.8. Held-out test performance

Our best configuration (Qwen3.5-4B at SC-N=16, 2,048-token budget, with guided repair) ranks first on the official Visual MCQ leaderboard at 84.06% overall, and is first on each of the six test languages (Table 12, Appendix I).

The test split (n=1,117) is 44.8% English and 30.6% Chinese, and both exceed their validation accuracy: English rises to 85.6% (from 73.5% val, $+12.1$ pp) and Chinese to 78.1% (from 74.7%, $+3.4$ pp). These are the two languages where the token-budget and guided-repair fixes (Sections 5.3, 5.4) had the most headroom on validation, so the test result tracks with the validation diagnosis.

6. Conclusion

We studied test-time scaling for open VLMs on multilingual visual MCQ, comparing structured search against flat self-consistency across two policies on the full EXAMS-V validation set. TTS improves accuracy: the full progression from zero-shot to self-consistency spans over 24 pp on Qwen3.5-4B, but the gains concentrate on the per-chain token budget rather than on structured search, larger chain counts, or trained selectors: neither a generative critic nor a trained PRM beats majority vote across two policy models and two pool strengths. Stratified analysis confirms that TTS gains concentrate on mid-accuracy languages where the policy is miscalibrated rather than weak, and that persistent below-average strata (Polish, Islamic Studies) resist SC because chains converge on the same wrong answer, leaving no diversity for the majority vote to correct.

The overarching lesson is that at our budget scale, the engineering substrate (parser format, decoding budget) and the policy choice dominate any search or selector method. Our best configuration, Qwen3.5-4B at SC-N=16 with a 2,048-token budget and guided repair (Section 5.4), reaches 81.6% on full validation and 84.1% on the held-out ImageCLEF 2026 test split, ranking first on the Visual MCQ leaderboard (Section 5.8). Future work should investigate whether decoupling perception from reasoning via a dedicated reasoning model can address the P(+) asymmetry identified in Section 5.2, and whether better-calibrated PRMs trained on multilingual non-mathematical content would break the selector null.

7. Limitations

Our findings rest on two policy models, Qwen2.5-VL-7B-Instruct and Qwen3.5-4B, both from the same family, so we cannot separate behavior typical of small open VLMs from behavior specific to Qwen models. The parseability effect in particular depends on how a model emits answer letters under a given prompt format, which may differ for other open source VLM families. The 11.4 pp gain from switching policy supports our claim that the policy substrate dominates inference-time strategy, but also shows that a third model could shift the picture; whether the pattern holds across families and across the sub-7B range is untested. Ablations whose full-validation cost was prohibitive used the dev-200 subset, whose ± 7 pp Wilson CIs preclude ranking claims, so all headline findings rest on val-full re-runs (n=4,651). The segmented PRM null appears structural, since P(+) ≈ 0.85 on reasoning steps leaves little signal once aggregated, though better per-step calibration on multilingual mixed-subject content might change this. Following the 2025 ImageCLEF winners [24], we did not fine-tune, use few-shot exemplars, or apply test-time image augmentation, all of which remain open.

Acknowledgments

Experiments were conducted on the Snellius national supercomputer, provided by the University of Amsterdam.

Declaration on Generative AI

During the preparation of this work, the authors use AI for proofreading, polishing and rephrasing sentences and paragraphs in the manuscript.

References

[1] 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, 2024. URL: https://arxiv.org/abs/2403.10378. arXiv:2403.10378.

[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] 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] 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, 2023. URL: https://arxiv.org/abs/2203.11171. arXiv:2203.11171.

[5] H. Lightman, V. Kosaraju, Y. Burda, H. Edwards, B. Baker, T. Lee, J. Leike, J. Schulman, I. Sutskever, K. Cobbe, Let's verify step by step, 2023. URL: https://arxiv.org/abs/2305.20050. arXiv:2305.20050.

[6] C. Snell, J. Lee, K. Xu, A. Kumar, Scaling llm test-time compute optimally can be more effective than scaling model parameters, 2024. URL: https://arxiv.org/abs/2408.03314. arXiv:2408.03314.

[7] M. Ahmadpour, A. Meighani, P. Taebi, O. Ghahroodi, A. Izadi, M. S. Baghshah, Limits and gains of test-time scaling in vision-language reasoning, 2025. URL: https://arxiv.org/abs/2512.11109. arXiv:2512.11109.

[8] S. S. Ghosal, S. Chakraborty, A. Reddy, Y. Lu, M. Wang, D. Manocha, F. Huang, M. Ghavamzadeh, A. S. Bedi, Does thinking more always help? mirage of test-time scaling in reasoning models, 2025. URL: https://arxiv.org/abs/2506.04210. arXiv:2506.04210.

[9] A. P. Gema, A. Hägele, R. Chen, A. Arditi, J. Goldman-Wetzler, K. Fraser-Taliente, H. Sleight, L. Petrini, J. Michael, B. Alex, P. Minervini, Y. Chen, J. Benton, E. Perez, Inverse scaling in test-time compute, 2025. URL: https://arxiv.org/abs/2507.14417. arXiv:2507.14417.

[10] T. Zeng, S. Zhang, S. Wu, C. Classen, D. Chae, E. Ewer, M. Lee, H. Kim, W. Kang, J. Kunde, Y. Fan, J. Kim, H. I. Koo, K. Ramchandran, D. Papailiopoulos, K. Lee, Versaprm: Multi-domain process reward model via synthetic reasoning data, 2025. URL: https://arxiv.org/abs/2502.06737. arXiv:2502.06737.

[11] D. B. Lee, S. Lee, S. Park, M. Kang, J. Baek, D. Kim, D. Wagner, J. Jin, H. Lee, T. Bocklet, J. Wang, J. Fu, S. J. Hwang, J. Bian, L. Song, Rethinking reward models for multi-domain test-time scaling, 2025. URL: https://arxiv.org/abs/2510.00492. arXiv:2510.00492.

[12] B. Ong, T. D. Pala, V. Toh, W. C. Tjhi, S. Poria, Training vision-language process reward models for test-time scaling in multimodal reasoning: Key insights and lessons learned, 2025. URL: https://arxiv.org/abs/2509.23250. arXiv:2509.23250.

[13] A. Panickssery, S. R. Bowman, S. Feng, Llm evaluators recognize and favor their own generations, 2024. URL: https://arxiv.org/abs/2404.13076. arXiv:2404.13076.

[14] J. Zhao, R. Liu, K. Zhang, Z. Zhou, J. Gao, D. Li, J. Lyu, Z. Qian, B. Qi, X. Li, B. Zhou, Genprm: Scaling test-time compute of process reward models via generative reasoning, 2025. URL: https://arxiv.org/abs/2504.00891. arXiv:2504.00891.

[15] J. Zhang, Y. Yan, K. Zheng, X. Zou, S. Dai, X. Hu, Gm-prm: A generative multimodal process reward model for multimodal mathematical reasoning, 2025. URL: https://arxiv.org/abs/2508.04088. arXiv:2508.04088.

[16] P. Kuang, X. Wang, W. Liu, J. Dong, K. Xu, Tim-prm: Verifying multimodal reasoning with tool-integrated prm, 2025. URL: https://arxiv.org/abs/2511.22998. arXiv:2511.22998.

[17] R. Liu, J. Gao, J. Zhao, K. Zhang, X. Li, B. Qi, W. Ouyang, B. Zhou, Can 1b llm surpass 405b llm? rethinking compute-optimal test-time scaling, 2025. URL: https://arxiv.org/abs/2502.06703. arXiv:2502.06703.

[18] Y. Wu, Z. Sun, S. Li, S. Welleck, Y. Yang, Inference scaling laws: An empirical analysis of compute-optimal inference for problem-solving with language models, 2025. URL: https://arxiv.org/abs/2408.00724. arXiv:2408.00724.

[19] A. Setlur, N. Rajaraman, S. Levine, A. Kumar, Scaling test-time compute without verification or rl is suboptimal, 2025. URL: https://arxiv.org/abs/2502.12118. arXiv:2502.12118.

[20] J. Huang, X. Chen, S. Mishra, H. S. Zheng, A. W. Yu, X. Song, D. Zhou, Large language models cannot self-correct reasoning yet, 2024. URL: https://arxiv.org/abs/2310.01798. arXiv:2310.01798.

[21] W. Wang, Z. Gao, L. Chen, Z. Chen, J. Zhu, X. Zhao, Y. Liu, Y. Cao, S. Ye, X. Zhu, L. Lu, H. Duan, Y. Qiao, J. Dai, W. Wang, Visualprm: An effective process reward model for multimodal reasoning, 2025. URL: https://arxiv.org/abs/2503.10291. arXiv:2503.10291.

[22] S. Wang, Z. Liu, J. Wei, X. Yin, D. Li, E. Barsoum, Athena: Enhancing multimodal reasoning with data-efficient process reward models, 2026. URL: https://arxiv.org/abs/2506.09532. arXiv:2506.09532.

[23] Z. Chen, Y. Wang, T. Xiao, R. Zhou, X. Yang, W. Wang, Z. Sui, J. Wang, From mathematical reasoning to code: Generalization of process reward models in test-time scaling, 2025. URL: https://arxiv.org/abs/2506.00027. arXiv:2506.00027.

[24] 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, 2025. URL: https://arxiv.org/abs/2507.11114. arXiv:2507.11114.

[25] 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 Vol. 4038, 2025.

[26] P. Hu, Z. Zhang, Q. Chang, S. Liu, J. Ma, J. Du, J. Zhang, Q. Liu, J. Gao, F. Ma, Q. Liu, Prm-bas: Enhancing multimodal reasoning through prm-guided beam annealing search, 2025. URL: https://arxiv.org/abs/2504.10222. arXiv:2504.10222.

[27] 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, 2023. URL: https://arxiv.org/abs/2309.06180. arXiv:2309.06180.

[28] J. Wang, K. Q. Lin, J. Cheng, M. Z. Shou, Think or not? selective reasoning via reinforcement learning for vision-language models, 2025. URL: https://arxiv.org/abs/2505.16854. arXiv:2505.16854.

[29] X. Yue, Y. Ni, K. Zhang, T. Zheng, R. Liu, G. Zhang, S. Stevens, D. Jiang, W. Ren, Y. Sun, C. Wei, B. Yu, R. Yuan, R. Sun, M. Yin, B. Zheng, Z. Yang, Y. Liu, W. Huang, H. Sun, Y. Su, W. Chen, Mmmu:

A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi, 2024. URL: https://arxiv.org/abs/2311.16502. arXiv:2311.16502.

[30] X. Yue, T. Zheng, Y. Ni, Y. Wang, K. Zhang, S. Tong, Y. Sun, B. Yu, G. Zhang, H. Sun, Y. Su, W. Chen, G. Neubig, Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark, 2025. URL: https://arxiv.org/abs/2409.02813. arXiv:2409.02813.

A. Chain-Agreement Analysis

Table 5 reports per-tier accuracy on the Q3.5 SC-$N{=}8$ (2k-tok) pool. The agreement tier is the number of chains (out of 8) that commit the same answer letter.

Table 5
Chain-agreement tiers on Q3.5 SC-$N{=}8$ (2k-tok, $n{=}4{,}651$). Weak-agreement questions ($\leq 4/8$) have accuracy below chance for 4-option MCQ.

Tier	n	% of val	Accuracy
Unanimous ($8/8$)	1437	30.9%	90.2%
Strong (6–$7/8$)	1364	29.3%	70.9%
Moderate ($5/8$)	637	13.7%	53.5%
Weak ($\leq 4/8$)	1089	23.4%	41.7%
All unparseable	124	2.7%	—

The skip rule used by both selectors (Section 5.6) passes through every question with at least $5/8$ agreement, i.e. the unanimous, strong, and moderate tiers (73.9% of validation). Only the weak and all-unparseable tiers (26.1%) are rescored. On this scored subset the base accuracy is $\sim$37%, leaving limited room for any selector to improve on majority vote.

B. Generative Critic Score Distribution

The training-free generative critic scores each chain on three axes (step intent, visual alignment, logical soundness), each rated 1–5, mapped to $[0, 1]$ via $(s{-}1)/4$, and mean-aggregated. Table 6 shows the score distribution across 5,710 scored chains from the Q2.5 SC-$N{=}8$ pool.

Table 6
Generative critic score distribution. The modal score 0.75 ($4/4/4$ axis pattern) accounts for 40% of chains. The calibration slope is nearly flat: a perfect 1.0 is correct only 40% of the time.

Score	% of chains	Acc. at score
0.00	0.1%	14.3%
0.25	5.7%	27.4%
0.50	5.0%	29.9%
0.75 (mode)	**39.7%**	**33.6%**
0.83	8.8%	30.5%
0.92	5.3%	33.0%
1.00	9.8%	39.7%

The correct-vs-incorrect chain score gap is negligible: correct chains average 0.713 vs. 0.687 for incorrect (gap 0.026), with an identical median of 0.75. In 59% of questions, multiple chains are tied at the top score, forcing arbitrary tie-breaking. The pattern is consistent with the self-recognition bias reported by Panickssery et al. [13]: the policy rates its own outputs uniformly high regardless of correctness.

C. Per-Language Chain Agreement

Table 7 breaks down chain agreement by language for the Q2.5 SC-N=8 pool, revealing the structural differences that drive per-language selector performance.

Table 7
Per-language chain agreement on Q2.5 SC-N=8 (n=4,651). English has 34% all-unparseable questions; Arabic and Hungarian have the largest weak-agreement tails.

Lang	n	Unan.	Strong	Weak	Unparse.
Arabic	517	19.0%	29.2%	36.4%	0.4%
Bulgarian	400	40.8%	34.5%	13.2%	0.0%
Chinese	600	22.3%	29.3%	30.2%	0.2%
Croatian	585	39.1%	29.1%	18.5%	0.0%
English	347	14.1%	10.1%	36.0%	**34.0%**
French	224	43.3%	35.3%	11.6%	0.0%
German	279	30.5%	39.4%	15.8%	0.4%
Hungarian	535	20.4%	32.0%	30.1%	0.4%
Italian	562	45.0%	28.5%	15.3%	0.0%
Polish	100	41.0%	25.0%	20.0%	0.0%
Serbian	502	35.7%	29.7%	19.3%	0.0%

English performs worst under Q2.5: only 14% unanimous, 34% all-unparseable, 36% weak agreement; 70% of English questions fall in the worst two tiers, explaining why English is the only language where SC *regresses* against zero-shot under Q2.5. Arabic and Hungarian have the largest weak-agreement tails (36% and 30%), making them the languages where selectors have the most questions to rescore and the least signal to work with. Bulgarian, French, and Italian show consistent agreement distributions (39–45% unanimous, small weak tails, 80+ % SC accuracy).

D. Per-Subject Accuracy

Table 8 reports accuracy for the five weakest and five strongest subjects under the best Q3.5 and Q2.5 configurations respectively.

Table 8
Per-subject accuracy for selected subjects ($n \geq 50$). Bottom 5 and top 5 under Q3.5 SC-N=8 (2k-tok). Q2.5 columns: zero-shot (ZS), CoT, SC-N=8.

		Q2.5			Q3.5
Subject	n	ZS	CoT	SC	**SC 2k**
Bottom 5 (Q3.5)					
Social	100	28.0	39.0	49.0	**43.0**
Professional	100	47.0	51.0	52.0	**57.0**
Islamic Studies	83	21.7	45.8*	33.7↓	**57.8**
Agriculture	100	47.0	48.0	42.0↓	**59.0**
Tourism	76	55.3	63.2	61.8	**67.1**
Top 5 (Q2.5 SC, for contrast)					
Psychology	81	80.2	91.4	85.2	—
Politics	135	76.3	76.3	80.0	—
Sociology	190	66.8	72.1	77.4	—
Biology	594	65.3	70.2	72.6	—
Mathematics	100	49.0	60.0	69.0	—

*CoT > SC under Q2.5 (regression). ↓SC regresses vs. ZS. "–" = Q3.5 per-subject data unavailable for non-bottom-5 subjects at time of writing.

E. PRM Scoring Mode (One-shot vs Per-step)

On a DTR $N{=}4, M{=}4$ pool under Qwen2.5-VL-7B (dev-200), per-step (segmented) PRM scoring is an exact null ($63.0\% \rightarrow 63.0\%$) while one-shot scoring shows a directional $+3.0$ pp gain ($63.0\% \rightarrow 66.0\%$) and is $9.4\times$ faster (Table 9). However, the ± 7 pp Wilson CIs at $n{=}200$ overlap fully, so the difference is not significant. One-shot evaluation gives the PRM maximal context to judge chain quality holistically, avoiding the per-step error compounding diagnosed in Section 5.2.

Table 9
PRM scoring mode on a DTR Q2.5 pool (dev-200). CIs overlap; the one-shot gain is directional only.

Selector	Acc.	95% CI	Δ	s/q
DTR majority (anchor)	63.0%	[56.1, 69.4]	—	—
PRM segmented (per-step)	63.0%	[56.1, 69.4]	$+0.0$ pp	56.3
PRM flat (one-shot)	66.0%	[59.2, 72.2]	$+3.0$ pp	6.0

F. Generative Critic Prompt

The critic scores each chain independently on three axes, one inference call per axis. The prompt template is:

```
You are a rigorous evaluator reviewing a reasoning chain written by a student
answering a multiple-choice exam question. The question and its answer options
are shown in the image.

Rate the chain on ONE axis: {axis_name}.

Axis definition: {axis_definition}

Scale: 1 (severely deficient), 2 (weak), 3 (adequate), 4 (strong), 5 (excellent).

Reasoning chain to evaluate:
—
{chain_text}
—

Output ONLY a JSON object on a single line with exactly two fields:
{"score": <integer 1-5>, "reason": "<brief one-sentence justification>"}
Do not output anything else.
```

The three axes are:

Step intent. Does each reasoning step address the question directly and build toward a commitment to one of the answer options? Penalise irrelevant tangents, repetition, self-doubt loops, or failure to commit to a letter.

Visual alignment. Does the chain's reasoning match what is actually visible in the image? Penalise hallucinated elements, misread values, missing visual evidence, or contradictions with the options rendered in the image.

Logical soundness. Does the final answer follow from the premises via valid logical, mathematical, or scientific steps? Penalise unjustified leaps, false equivalences, arithmetic errors, or a mismatch between the concluded letter and the reasoning that preceded it.

Per-axis scores are mapped to $[0, 1]$ via $(s-1)/4$ and mean-aggregated to a single chain score.

Table 10
Per-language accuracy on full validation (n=4,651). Q2.5 columns: ZS, CoT, SC-N=8 (MMMU closer). Q3.5 columns: SC-N=8 at 1024 and 2048 tokens. $\Delta_{1k\to2k}$ is the gain from the budget increase. Down-arrow marks a regression.

		Q2.5			Q3.5 SC		
Lang	n	ZS	CoT	SC	1k	**2k**	$\Delta_{1k\to2k}$
Arabic	517	40.8	56.7	59.8	71.8	**74.5**	+2.7
Bulgarian	400	58.8	76.8	81.0	92.0	**93.5**	+1.5
Chinese	600	71.8	60.8$^{\downarrow}$	69.0	66.2	**74.7**	+8.5
Croatian	585	60.7	67.0	70.4	84.6	**86.0**	+1.4
English	347	36.3	38.0	30.0$^{\downarrow}$	56.2	**73.5**	+17.3
French	224	69.6	78.1	81.3	92.0	**92.0**	+0.0
German	279	61.7	68.1	69.2	84.9	**86.4**	+1.5
Hungarian	535	49.4	57.8	59.3	77.9	**80.0**	+2.1
Italian	562	64.4	69.0	69.4	81.9	**83.3**	+1.4
Polish	100	47.0	51.0	52.0	55.0	**57.0**	+2.0
Serbian	502	56.6	66.3	70.5	83.3	**84.7**	+1.4

G. Per-Language Accuracy

H. Dev-200 Search Comparison

Table 11
Search and verification on dev-200 under Q2.5. All rows use the same pre-MMMU closer for a matched comparison; the MMMU closer was adopted later and applied at val scale (Table 2). At n=200 the ±7 pp Wilson intervals overlap, so dev-200 alone cannot rank the more expensive methods against SC.

Method	**Acc.**	**95% CI**	**Calls/q**	s**/q**
SC-N=8	**65.5%**	**[58.6, 71.8]**	**8**	**7.4**
DTR N=2,M=2	55.0%	[48.1, 61.7]	6	14.1
DTR N=4,M=4	61.5%	[54.6, 68.0]	20	28.5
DTR-N=2 + PRM-BAS	60.5%	[53.6, 67.0]	13.4	42.3

I. Official Leaderboard

Table 12
Official ImageCLEF 2026 Visual MCQ leaderboard (test split, n=1,117): overall accuracy and per-language breakdown across the six test languages. Our submission (leaderboard handle `spirosbax`) ranks first overall and on every language.

#	**Participant**	**Overall**	**EN**	**BG**	**ZH**	**HR**	**IT**	**SR**
1	**spirosbax (ours)**	**84.06**	**85.60**	**89.09**	**78.07**	**87.72**	**90.74**	**87.04**
2	DS@GT	79.86	85.20	86.36	67.25	85.96	87.04	83.33
3	FAU	71.08	74.80	69.09	63.45	73.68	81.48	75.93
4	zhaijinghe	61.50	67.60	61.82	53.51	61.40	59.26	57.41
5	sjaini	59.27	69.20	50.91	44.15	59.65	72.22	66.67
6	begyed	57.74	60.00	66.36	47.08	66.67	70.37	64.81
7	linxiaocan	55.60	60.40	54.55	51.17	50.88	59.26	42.59
8	pratikpriyanshu	50.76	57.20	43.64	41.81	56.14	55.56	51.85
9	wether	46.73	54.60	56.36	31.87	45.61	48.15	48.15