UNED-Martinez at ImageCLEF 2026 Task on Multimodal Reasoning: Hybrid Per-Language Prompting for Multilingual Visual OpenQA

UNED-Martinez at CLEF 2026

Fernando Martínez Marco[1]

[1]Universidad Nacional de Educación a Distancia (UNED), Madrid, Spain

Abstract

This paper describes the UNED-Martinez system developed for the ImageCLEF 2026 Multimodal Reasoning task on Visual Open-domain Question Answering (OpenQA). The approach employs Qwen2.5-VL-7B-Instruct, a seven-billion-parameter open-weight vision-language model, with a hybrid per-language prompting strategy in a fully zero-shot, greedy-decoding setting. Rather than applying a uniform prompt across all target languages, a distinct prompting strategy was designed and experimentally evaluated for each of the six evaluation languages: Bulgarian and English receive concise direct prompts in English, while Chinese, Croatian, Italian, and Serbian each receive a prompt written in the respective native language. The system achieved an overall score of **0.5926** on the development set, and finally finished in **3rd** place on the official final-phase test leaderboard. A systematic ablation study across nine prompting configurations demonstrates that the asymmetric per-language strategy outperforms both a uniform English baseline (+4.9 points overall) and chain-of-thought approaches. The codebase and model weights are fully open-source at https://github.com/fmmarco29/imageclef-2026-qwen. Complete system prompt templates are included in the appendix to ensure reproducibility.

Keywords

Multimodal Reasoning, Visual Question Answering, Vision-Language Models, Multilingual NLP, Qwen2.5-VL, Prompt Engineering, Zero-Shot Inference, ImageCLEF 2026

1. Introduction

Vision-language models (VLMs) have matured into powerful tools for multimodal understanding, yet their systematic evaluation in multilingual settings is still poorly understood. Most existing benchmarks target a single language or assume that a uniform prompt template transfers across languages. The ImageCLEF 2026 Multimodal Reasoning task tackles this problem directly: systems must answer academic exam questions embedded as images in six typologically diverse languages, requiring simultaneous optical character recognition (OCR), visual reasoning over diagrams and tables, and answer generation in the correct target language [3, 4].

Unlike conventional VQA benchmarks where the question is provided as separate text, the question in this task is encoded entirely within the image. This forces the model to perform OCR, contextual reasoning over the visual content, and answer generation all in a single forward pass. The multilingual dimension adds a further complication: a model that successfully answers questions in English may produce incorrect or off-language responses for Cyrillic-script languages such as Bulgarian or Serbian, or for logographic scripts such as Chinese.

In this paper, we describe our approach and findings across three key aspects. First, we show that a compact, open-weight VLM (Qwen2.5-VL-7B-Instruct, 7 B parameters) ranks third overall on the official final-phase test leaderboard within the Tiny ($\leq$7 B) category without fine-tuning, proprietary APIs, or few-shot examples. Second, we present a systematic ablation study that isolates the effect of prompt language, chain-of-thought reasoning [2], and prompting asymmetry, uncovering how model-internal biases drive per-language performance gaps. Finally, we discuss a mechanistic hypothesis—attention

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany
fmartinez24@alumno.uned.es (Fernando Martínez Marco)
0009-0004-0523-0187 (Fernando Martínez Marco)
© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

dilution, VRAM fragmentation, and English as a visual bridge language—that seeks to explain the observed behaviour and points to concrete directions for future multilingual VLM prompting.

1.1. Related Work

Prompt engineering for Vision-Language Models (VLMs) has historically relied on uniform instruction templates, often in English, assuming cross-lingual capabilities are robust enough to generalize implicitly [7]. However, recent studies highlight that VLMs exhibit strong language-specific biases due to imbalances in multilingual pre-training datasets [5]. Zero-shot visual reasoning tasks are particularly sensitive to instruction language, and cross-lingual alignment can degrade when processing low-resource scripts or complex domain-specific layouts. Strategies such as translation-at-inference and bridge-language prompting have been proposed to bypass these limitations, where a high-resource language (e.g., English) guides the reasoning path before generating output in the target language. Our work builds on these insights by systematically validating an asymmetric prompting approach tailored to the specific script and resource representation of each evaluation language.

2. Task Description

The ImageCLEF 2026 MR2026-OpenQA-Visual task requires generating free-text answers to exam questions presented as images. Each dataset item consists of an image containing both a question and its visual context (diagrams, graphs, tables, or photographs), together with metadata fields for language, subject area, and question identifier. No separate textual question field is provided: the model must read the question directly from the image [3, 4]. The training and development splits are available on HuggingFace Datasets (SU-FMI-AI/ImageCLEF-MR2026-OpenQA-Visual); the development set used for evaluation contains 240 items balanced across six languages: English (EN), Bulgarian (BG), Chinese (ZH), Croatian (HR), Italian (IT), and Serbian (SR). Subject areas span biology, chemistry, physics, geography, and history, with items sourced from national exam corpora including EXAMS-V [5].

Answers are evaluated by the task organizers using the COMET metric [6], a neural framework for semantic similarity evaluation. Scores range between 0 and 1, with higher values indicating greater semantic fidelity to the reference answer. The final phase imposes strict constraints: only fully open-source models are permitted, closed-API systems are prohibited, and submissions are executed by the official evaluation platform on an NVIDIA A40 40 GB GPU. Participating systems are stratified by parameter count; our submission competes in the *Tiny* ($\leq$7 B) category.

3. System Description

3.1. Model and Hardware Configuration

Our system is built on Qwen2.5-VL-7B-Instruct [1], an open-weight VLM with approximately seven billion parameters released under the Apache 2.0 licence, which builds upon the architecture of its predecessor Qwen2-VL [7]. The model processes interleaved image and text sequences using a dynamic visual resolution mechanism that adapts the number of visual tokens to the input image content.

The inference script automatically detects available hardware and adjusts its configuration. On dual NVIDIA T4 GPUs (16 GB VRAM), the model is loaded with 4-bit NF4 quantisation via BitsAndBytes [8], using double quantisation and `float16` compute dtype. On the competition's NVIDIA A40 GPU (40 GB VRAM), the model is loaded in full `float16` without quantisation. Scaled dot-product attention (`sdpa`) is used in both environments. The adaptive image resolution is set to MAX_PIXELS $= 1024\times28\times28$ on T4 hardware and $2048\times28\times28$ on A40. Decoding is greedy (`do_sample=False`, `max_new_tokens=1024`) for full reproducibility, and the random seed is fixed to 42.

3.2. Hybrid Per-Language Prompting Strategy

The key design choice in this paper is the *hybrid per-language prompting* strategy: each language receives a distinct system prompt based on experimental evidence gathered on the development set, rather than a single uniform template. Table 1 summarises the final assignment for the final submission (Experiment 9).

Table 1
Per-language prompting strategy in the final submission (Experiment 9).

Language	Script	Prompt lang.	Reasoning style
English (EN)	Latin	English	Direct, strict rules
Bulgarian (BG)	Cyrillic	English	Short direct, no CoT
Chinese (ZH)	Logographic	Chinese	Native, strict rules
Croatian (HR)	Latin	Croatian	Native, strict rules
Italian (IT)	Latin	Italian	Native, strict rules
Serbian (SR)	Cyrillic	Serbian	Native, strict rules

The English system prompt enforces six explicit rules: write in English, output only the final answer without any prefix, produce no explanations or reasoning steps, include all values for multi-part questions, be numerically precise, and inspect all visual elements carefully. The Bulgarian prompt is intentionally minimal—three sentences—because empirical evidence suggested that longer prompts may fragment the model's visual attention (Section 5):

> *"You are an expert exam solver. Answer in Bulgarian. Provide a concise and direct answer based strictly on the image. Do not explain your reasoning."*

For Chinese, Croatian, Italian, and Serbian, the system prompts are written entirely in the target language and contain four parallel strict rules: respond in the target language only; output only the final answer; include all parts for multi-part questions; be precise and concise. The complete text of these prompts is provided in Appendix A.

3.3. Answer Normalisation and Validation

Raw model outputs frequently contain spurious answer prefixes in all six languages, markdown code fences, internal chain-of-thought blocks enclosed in `<think>...</think>` tags, and extraneous punctuation. The normalisation pipeline applies the following steps in sequence:

1. Remove markdown code fences.
2. Remove `<think>...</think>` blocks via two-pass regex (closed tags first, then unclosed tags caused by truncation).
3. Strip internal-thought headers via 8 regex patterns covering EN, BG, and ZH variants of `reasoning:`, `thinking:`, `let's think:`, `step by step:`, and their Cyrillic and Chinese equivalents.
4. Strip multilingual answer prefixes via 20 regex patterns covering EN, BG, HR, IT, SR, and ZH scripts.
5. Remove surrounding quotation marks (straight and typographic).
6. Remove isolated trailing periods.
7. Normalise internal whitespace.

The validation step checks whether the normalised string is in a set of high-frequency invalid tokens (empty string, bare function words in any of the six languages, single-character strings). If the first generation attempt fails validation, a simplified emergency prompt without any system message is used as a fallback. If both attempts fail, a single-space placeholder is written to the output.

3.4. Inference Pipeline

The system is containerised in a Docker image that pre-downloads the model weights at build time for offline inference on the official evaluation platform. The pipeline processes items sequentially and maintains a JSON checkpoint every 10 items. VRAM pressure is controlled by flushing the GPU cache every 20 items. The end-to-end flow is summarised in Figure 1 and consists of four steps:

1. Select system prompt based on language; construct messages (system + user with embedded image).
2. Run Qwen2.5-VL-7B with greedy decoding; normalise; validate. If valid, return answer.
3. If invalid: flush GPU cache; run emergency retry prompt; normalise; validate. Return answer or space placeholder.
4. Output: `[{"question_id": "...", "answers": ["..."], "language": "..."}]`.

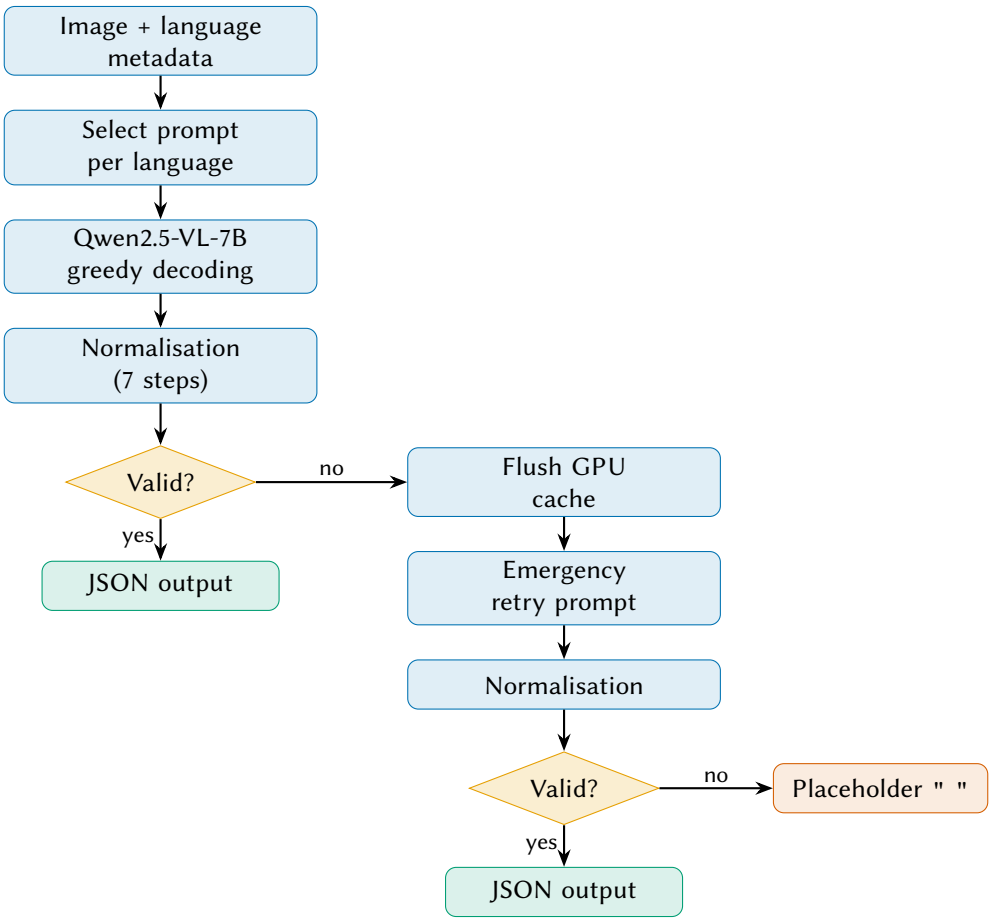

Figure 1: End-to-end inference pipeline. A language-specific prompt is selected and the image is passed to Qwen2.5-VL-7B with greedy decoding. If normalisation and validation fail, the GPU cache is flushed and an emergency prompt without a system message is used as fallback; if both attempts fail, a single-space placeholder is written.

4. Experiments

4.1. Experimental Setup

All experiments evaluate on the development set (240 items, 40 per language). The baseline (Experiment 1) uses a uniform English system prompt with a few-shot pool of 1,026 examples built from the training split and EXAMS-V [5], selected by a hierarchical relevance score prioritising language and subject

matching. All nine configurations share the same normalisation and validation pipeline; only the prompting strategy changes.

4.2. Ablation Study

Table 2 presents the complete ablation study; Figure 2 visualises the per-language breakdown for five representative configurations. Each configuration is described in order of increasing overall score.

Table 2
Ablation study on the development set (240 items). Bold marks the best score per column. The final submission is Experiment 9.

Config	Overall	EN	BG	ZH	HR	IT	SR	Description
Experiment 1	0.5433	0.5512	0.5345	0.5249	0.5356	0.5544	0.5592	Uniform EN prompt + few-shot
Experiment 2	0.5005	0.5300	0.4551	0.4645	0.5159	0.5337	0.5038	CoT + JSON structured output
Experiment 3	0.5287	0.5339	0.4991	0.5248	0.5140	0.5800	0.5203	EN reasoning bridge (CoT all)
Experiment 4	0.5399	0.5619	0.4903	0.5350	0.5411	0.5531	0.5577	CoT + `###FINAL###` marker
Experiment 5	0.5468	0.5622	0.4848	0.5608	0.5697	0.5465	0.5566	Bilingual: EN base + native reinf.
Experiment 6	0.5540	0.5622	0.4615	0.5631	0.5858	0.5771	0.5740	100% native prompt per language
Experiment 7	0.5673	0.5409	0.5623	0.5838	0.5718	0.5737	0.5711	Selective CoT: BG=CoT EN, rest=native
Experiment 8	0.5834	0.5866	0.5364	0.5856	0.6208	0.5825	0.5885	Hybrid: BG=CoT EN, EN=direct, rest=native
Experiment 9	**0.5926**	**0.5849**	**0.5714**	**0.5848**	**0.6240**	0.5788	**0.6117**	Hybrid: BG=brief direct, EN=direct, rest=native

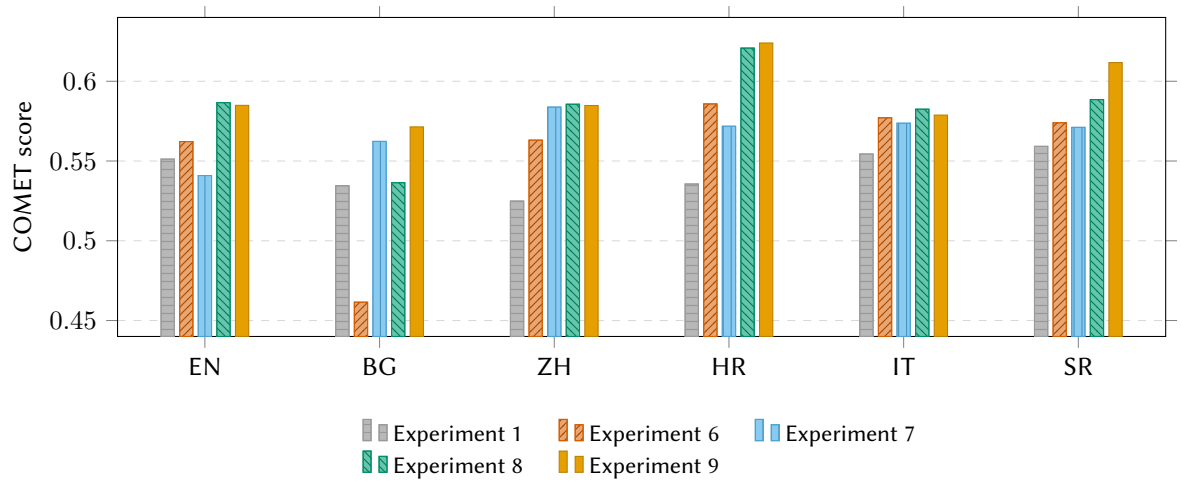

Figure 2: Per-language COMET scores for five key configurations on the development set. The Bulgarian (BG) anomaly is noticeable: native-language prompts (Experiment 6, vermillion, diagonal hatch) result in a score of 0.4615, while selective CoT (Experiment 7, sky blue, vertical hatch) recovers BG substantially. The final submission Experiment 9 (orange, solid fill) achieves the best overall balance, with Croatian (HR) reaching 0.6240.

Experiment 1 — Uniform English baseline. A single English system prompt for all six languages establishes a baseline reference (0.5433 overall).

Experiment 2 and Experiment 4 — Chain-of-thought (uniform). Applying CoT instructions uniformly degraded performance. The FINAL-marker variant (Experiment 4) fell below baseline (−0.003), while the JSON structured-output variant (Experiment 2) dropped to 0.5005 (−0.043), with severe losses for Bulgarian (−0.079) and Chinese (−0.060). Structured constraints interfere with the model's ability to generate fluent non-English output.

Experiment 3 — English reasoning bridge. Using English CoT reasoning for all languages while requiring output in the native language also underperformed (0.5287, −0.015). Increased prompt length appears to divert attention from visual tokens without providing commensurate reasoning benefit.

Experiment 5 and Experiment 6 — Prompt language variation. Switching to fully native prompts (Experiment 6) improved Croatian (+0.050), Chinese (+0.038), Italian (+0.023), and Serbian (+0.015) over baseline, while Bulgarian dropped to 0.4615 (−0.073). This revealed a strong language-specific bias in the model's instruction-following for Cyrillic-script languages.

Experiment 7 — Selective CoT. Restricting CoT to Bulgarian (via an English CoT prompt) while keeping native prompts for all other languages substantially recovered Bulgarian to 0.5623 (+0.028 over baseline) and raised overall to 0.5673 (+0.024). Chinese reached a local peak of 0.5838; English regressed slightly (−0.010), confirming that CoT is suboptimal for English.

Experiment 8 — Zero-shot hybrid. Removing the entire few-shot pool while keeping language-specific prompts produced a large gain: 0.5834 overall (+0.016 over Experiment 7, +0.040 over baseline). Croatian reached 0.6208 and English 0.5866, confirming that zero-shot inference outperforms few-shot for this task.

Experiment 9 — Final submission. The sole change from Experiment 8 is the Bulgarian prompt: the multi-sentence CoT English prompt is replaced by the three-sentence minimalist direct prompt. Bulgarian improved from 0.5364 to 0.5714 (+0.035), Serbian from 0.5885 to 0.6117 (+0.022 as a side effect), and overall to **0.5926** (+0.009 over Experiment 8, +0.049 over baseline).

5. Results and Analysis

5.1. Official Leaderboard Evaluation

On the official final hidden-test leaderboard, the UNED-Martinez system ranked **3rd overall** among all submitted systems in the final evaluation phase. While the official hidden-test set COMET scores are not released individually to participants, the 3rd place ranking confirms that the prompting insights validated on the development set transfer successfully to the hidden evaluation data.

5.2. Zero-Shot Outperforms Few-Shot

Removing the 1,026-example few-shot pool between Experiment 7 and Experiment 8 improved overall performance by 0.016 points. We hypothesise that the degradation arises from a *multimodal hallucination* effect: when the model receives textual descriptions of answers referencing images it cannot see, it attempts to locate the described visual features in the current test image and misaligns its attention. Zero-shot inference eliminates this cross-image contamination, in line with prior work on prompt grounding in VLMs.

5.3. Native Prompts Improve Non-Cyrillic Languages but Harm Bulgarian

Experiment Experiment 6 shows that native-language prompts substantially improve Croatian (+0.050), Chinese (+0.038), Italian (+0.023), and Serbian (+0.015) over the uniform English baseline. These gains are consistent with the hypothesis that native-language instructions reduce language-switching artifacts in generated text.

Bulgarian is the critical exception: native Cyrillic prompts cause a significant drop of -0.073 to 0.4615. This behavior suggests a potential *training data imbalance*: the Qwen2.5-VL model has been exposed to substantially fewer image–Bulgarian text pairs than image–English text pairs during pre-training. English instructions reliably activate the model's visual reasoning pathways, whereas Bulgarian instructions do not. Using English as the instruction language for Bulgarian therefore acts as a *visual bridge language*: the model reasons in English and delegates only the final token generation to its cross-lingual transfer capabilities. Interestingly, Serbian (which also uses the Cyrillic script) did not exhibit this degradation under native prompting (Experiment 6). This difference suggests a disparity in the pre-training alignment of specific Cyrillic-script languages within Qwen2.5-VL, where Serbian instruction-following is more robustly aligned than Bulgarian.

5.4. Attention Dilution Hypothesized to Explain CoT Failure for Bulgarian

Selective CoT (Experiment 7) improved Bulgarian to 0.5623 using an English CoT prompt. Yet Experiment 9 shows that a three-sentence direct English prompt yields 0.5714, exceeding the CoT score by 0.009 points. We hypothesize that this is due to *attention dilution*: in a 7 B parameter model, the attention budget per generation step is limited. A lengthy CoT system prompt containing explicit reasoning steps, XML tags, and negative constraints consumes a disproportionately large fraction of the context attention, leaving less attention capacity for the visual token sequence. The minimalist Bulgarian prompt concentrates the model's attention on the image content, resulting in more accurate visual question answering.

5.5. Serbian Improvement Hypothesized as a VRAM Side Effect

Serbian improved by 0.022 points between Experiment 8 and Experiment 9 despite no change to the Serbian prompt. We hypothesize that this improvement may be correlated with system-level memory behavior or caching allocator patterns. Speculatively, the CoT prompt for Bulgarian in Experiment 8 generates substantially longer sequences, which could cause the PyTorch memory allocator to fragment VRAM across multiple arenas. This fragmentation might prevent the dynamic resolution scaler from consistently achieving maximum pixel resolution ($2048 \times 28 \times 28$) for subsequent items, potentially downsampling Serbian images and degrading OCR quality. Replacing the long CoT for Bulgarian with a short direct prompt reduces VRAM pressure, allowing the allocator to maintain peak resolution throughout the inference run, though further study is needed to isolate this effect from other image resolution behaviors.

5.6. English Prefers Direct Prompts

English consistently scores best with a direct, rule-based prompt and no CoT. The highest EN score (0.5866 in Experiment 8) is achieved without any reasoning steps. Adding CoT to English (Experiment 7) reduces its score to 0.5409 (-0.046). This suggests that, for a well-resourced language, explicit reasoning steps introduce format errors and elaboration that lower semantic similarity relative to concise reference answers.

5.7. Qualitative Failure Analysis

To better understand the limitations of the proposed approach, a qualitative analysis of incorrect generations was performed. Failures are primarily classified into three categories:

1. **Cyrillic OCR Limitations**: While Qwen2.5-VL performs well on Latin and Han characters, it occasionally misidentifies Cyrillic script letters in diagrams under low-contrast or stylized fonts. This directly affects Bulgarian and Serbian items.
2. **Complex Scientific Diagrams**: The model struggles with diagrams requiring spatial coordinate projections (e.g., parsing specific force vectors in physics or molecular bonds in chemistry).
3. **Multi-part Output Format Failures**: Despite validation fallbacks, the model occasionally outputs a single concatenated string instead of separated letters (e.g., ‘A, B‘ instead of a list structure), which results in semantic alignment penalties in the COMET scoring.

5.8. Cumulative Gains Summary

Table 3 summarises the stepwise contribution of each major design decision over the baseline.

Table 3
Cumulative gain over the Experiment 1 baseline for each design decision.

Design decision	Overall	Δ vs. baseline
Experiment 1: Uniform EN prompt, few-shot	0.5433	—
+ Native prompts ZH/HR/IT/SR (Experiment 6)	0.5540	+0.011
+ EN override for Bulgarian (Experiment 2-based)	0.5626	+0.019
+ Selective CoT for BG (Experiment 7)	0.5673	+0.024
+ Remove few-shot pool (Experiment 8)	0.5834	+0.040
+ Brief direct BG prompt (Experiment 9)	**0.5926**	**+0.049**

6. Limitations and Future Work

The zero-shot hybrid prompting strategy has two clear limitations. The prompting configurations were tuned on a 240-item development set, which may not transfer to out-of-domain subjects or scripts not covered in the ablation. The pipeline also lacks a semantic verification step, so factually wrong answers that survive the regex validation pass undetected.

Three directions follow from these gaps. First, replacing the static few-shot pool with a dynamic RAG framework—indexing training questions in a semantic vector space and retrieving the top-k most relevant examples at inference time—would supply domain-specific context without the cross-image visual noise that degraded performance in the few-shot experiments. Second, the language biases documented here suggest a multi-agent consensus approach: generating candidate answers in parallel across several visual bridge languages (e.g., English, Italian, Chinese) and using a judge model to pick the most translation-consistent response before outputting the native-script answer. Third, parameter-efficient fine-tuning with LoRA, combined with visual attention masks that focus the encoder on text-heavy image regions, could reduce the calibration burden currently carried by hand-tuned prompts.

7. Conclusions

This paper has described a system for multilingual visual open-domain question answering that achieves third place on the official ImageCLEF 2026 Visual OpenQA final-phase test leaderboard within the Tiny category using a compact, fully open-source vision-language model. The main result is that a *hybrid per-language prompting strategy*—assigning each language an independently tuned system prompt—outperforms uniform English prompting by 4.9 points overall and beats chain-of-thought approaches applied uniformly.

The ablation study yields three concrete findings. Few-shot examples that reference images the model cannot see introduce multimodal hallucination and hurt performance, so zero-shot inference is the better

choice for this task. Attention dilution in 7 B models means long reasoning prompts reduce accuracy on visually grounded tasks rather than improving accuracy. English can also serve as a reliable visual bridge language for Cyrillic-script outputs when the model's native-language instruction-following is unreliable due to training data imbalance.

Future work will focus on systematic LoRA fine-tuning with properly calibrated quantisation, automated per-language prompt optimisation using development-set feedback, and dynamic routing strategies that assign prompting modes based on predicted item difficulty or script type.

Acknowledgments

The author is a student of the Master's Programme in Artificial Intelligence Research (Máster en Investigación en Inteligencia Artificial) at UNED, enrolled in the Fundamentals of Linguistic Processing course. This work was carried out independently as part of the course practicum. Compute resources were provided by Kaggle (NVIDIA T4 GPUs), Google Cloud (TPU v3-8), and the official evaluation platform (NVIDIA A40 GPU).

Declaration on Generative AI

During the preparation of this work, the author used generative AI tools for content enhancement (exploring methodological alternatives and technical solutions), paraphrasing, writing style improvement, grammar and spelling checks, and manuscript formatting. The author reviewed and edited all AI-assisted content and takes full responsibility for the publication.

A. System Prompt Templates

This appendix presents the complete, exact templates of the system prompts used for each language in the final submission (Experiment 9).

A.1. English (EN)

```
You are an expert exam solver. Answer the visual question directly.
Follow these strict rules:
1. Respond in English only.
2. Output ONLY the final answer. Do NOT include any conversational text, pleasantries
    , or explanations.
3. If the question asks for multiple parts or sub-questions (e.g., A, B, C), list the
     answers for each part clearly separated.
4. Be numerically precise and double-check any mathematical calculations.
5. Carefully inspect all diagram labels, graphs, and tables before choosing.
6. Under no circumstances write reasoning paths or explain your choice.
```

A.2. Bulgarian (BG)

```
You are an expert exam solver. Answer in Bulgarian. Provide a concise and direct
    answer based strictly on the image. Do not explain your reasoning.
```

A.3. Chinese (ZH)

```
You are an expert exam solver. Answer the visual question directly.
Follow these strict rules:
1. Respond in Chinese only (using Simplified Chinese).
2. Output ONLY the final answer. Do NOT include any conversational text, pleasantries
    , or explanations.
3. If the question asks for multiple parts, list them clearly.
4. Be precise and concise.
```

A.4. Croatian (HR)

```
You are an expert exam solver. Answer the visual question directly.
Follow these strict rules:
1. Respond in Croatian only.
2. Output ONLY the final answer. Do NOT include any conversational text, pleasantries
    , or explanations.
3. If the question asks for multiple parts, list them clearly.
4. Be precise and concise.
```

A.5. Italian (IT)

```
You are an expert exam solver. Answer the visual question directly.
Follow these strict rules:
1. Respond in Italian only.
2. Output ONLY the final answer. Do NOT include any conversational text, pleasantries
    , or explanations.
3. If the question asks for multiple parts, list them clearly.
4. Be precise and concise.
```

A.6. Serbian (SR)

```
You are an expert exam solver. Answer the visual question directly.
Follow these strict rules:
1. Respond in Serbian only (using Cyrillic script).
2. Output ONLY the final answer. Do NOT include any conversational text, pleasantries
    , or explanations.
3. If the question asks for multiple parts, list them clearly.
4. Be precise and concise.
```

References

[1] Qwen Team, Qwen2.5-VL technical report, 2025. URL: https://arxiv.org/abs/2502.13923. arXiv:2502.13923.

[2] 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 (NeurIPS 2022), volume 35, 2022, pp. 24824–24837.

[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] R. J. Das, S. E. Zlataev, F. D. Schmidt, A.-I. Mihailescu, P. Nakov, EXAMS-V: A multi-discipline multilingual multimodal exam benchmark for evaluating and analyzing 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. 7711–7728. URL: https://aclanthology.org/2024.acl-long.420. doi:10.18653/v1/2024.acl-long.420.

[6] R. Rei, C. Stewart, A. C. Farinha, A. Lavie, COMET: A neural framework for MT evaluation, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 6685–6701. doi:10.18653/v1/2020.emnlp-main.536.

[7] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, Y. Fan, K. Dang, M. Du, X. Ren, R. Men, D. Liu, C. Zhou, J. Zhou, J. Lin, Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution, arXiv preprint arXiv:2409.12191 (2024). URL: https://arxiv.org/abs/2409.12191.

[8] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer, QLoRA: Efficient finetuning of quantized LLMs, in: Advances in Neural Information Processing Systems (NeurIPS 2023), volume 36, 2023.