Benchmarking and Boosting Multilingual Capabilities of LVLMs via OCR-Centric Reinforcement Learning

Anonymous ACL submission

Abstract

Evaluating the multilingual capabilities of Large Vision-Language Models (LVLMs) remains challenging because most benchmarks rely on non-parallel corpora, making it unclear whether cross-lingual performance gaps reflect model limitations or dataset inconsistencies. To address this, we introduce **PM^{4}Bench**, the first multimodal, multilingual, multi-task benchmark built on a strictly parallel 10-language corpus, enabling **fair, apples-to-apples cross-lingual comparison** of model performance. We further introduce a **`vision` setting** that embeds textual inputs directly into images, better approximating deployment scenarios where LVLM-driven agents interact with virtual or physical environments through unified visual observations. Experiments with 10 LVLMs reveal that **OCR is a key factor behind cross-lingual disparity** when textual content is rendered visually. Motivated by this, we design an **OCR-centric GRPO training strategy** using fully synthesized, label-free OCR data, without expensive task-specific VQA supervision. The resulting model improves general multilingual VQA capability, reduces cross-lingual disparities under the `vision` setting, and transfers gains beyond PM^{4}Bench. This methodology offers an efficient, label-free pathway toward more equitable multilingual deployment of LVLM-driven agents.

1 Introduction

Large Language Models (LLMs) excel at QA, reasoning, and instruction following, yet performance gaps persist across languages even on language-agnostic tasks like math and code. Prior work attacks the problem from multiple angles: model mechanisms (Wendler et al., 2024; Tang et al., 2024b; Zhao et al., 2024), multilingual corpora (Xue et al., 2021; Yu et al., 2025), training and inference techniques (Zhu et al., 2024a; She et al., 2024; Zhu et al., 2024b; Shi et al., 2022; Huang et al., 2023), and benchmarks (Sun et al., 2024; Zhang et al., 2024; Huang et al., 2025).

Large Vision Language Models (LVLMs) extend this picture to multimodal input and now serve as the perceptual backbone of **multimodal AI agents** that read smartphone, web, and desktop interfaces as images. LVLMs inherit cross-lingual disparities from LLMs and add new ones, notably imbalanced text recognition across scripts, which results in uneven user experience once these agents are deployed globally.

A comprehensive evaluation of LVLMs in multilingual scenarios is essential for identifying shortcomings and guiding optimization, yet most existing benchmarks fall short on two axes. (1) **Non-parallel corpora**: which introduces uncontrolled variance across languages, making it difficult to isolate whether performance gaps among languages stem from differences in corpus content or fundamental model capabilities. (2) **Text and images are processed as separate streams**, which differs from realistic deployment scenarios where AI agents must interpret unified visual observations while interacting with virtual or physical environments.

To address these gaps, we introduce **PM^{4}Bench**, to our knowledge the first multilingual, multimodal, multi-task benchmark that simultaneously satisfies: (i) a **strictly parallel 10-language corpus**, where semantically identical content is rendered in every language, enabling apples-to-apples cross-lingual comparison; (ii) a **`vision` setting** that fuses text and images into a single visual input, better approximating real-world usage scenarios; and (iii) a **multi-perspective multi-task evaluation suite** that probes diverse capability dimensions, spanning reasoning VQA, open-ended generation, multi-scale OCR, and coordinate-based MGUI grounding. Taken together, these properties let us probe per-language performance on identical content and reason about cross-lingual gaps on equal footing.

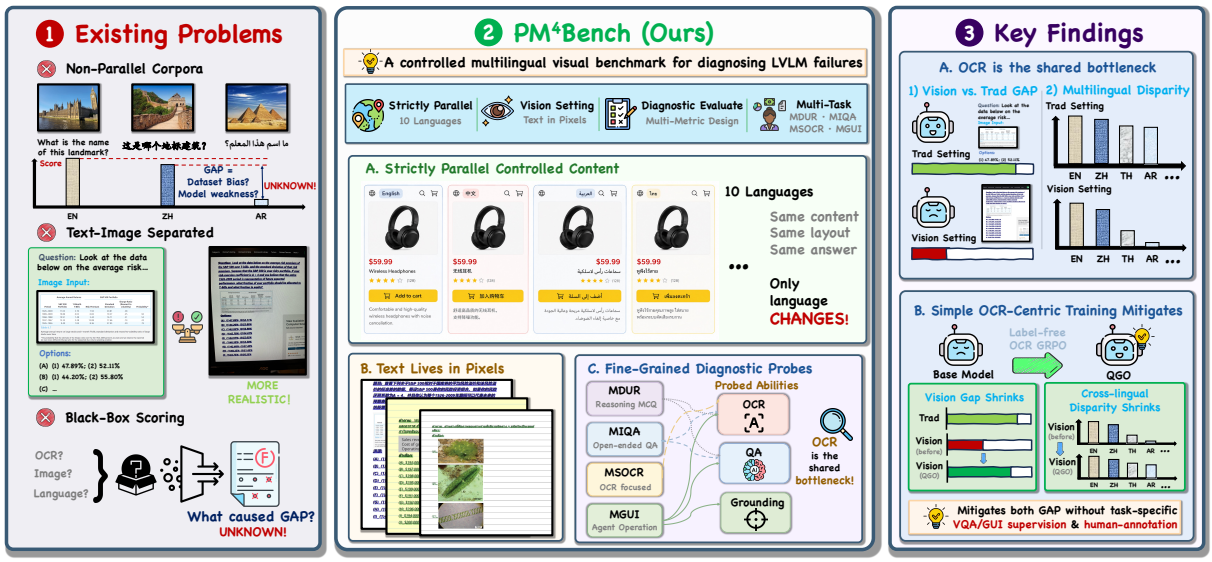


Figure 1: Overview of PM^{4}Bench. **(1) Existing problems** with current multilingual LVLM benchmarks: *non-parallel corpora* confound model capability with content variance across languages; *text-image separation* diverges from realistic deployment, where agents observe unified visual interfaces; and *lack of cause analysis* leaves the drivers of cross-lingual disparity and `vision`-setting performance drops undiagnosed. **(2) PM^{4}Bench (ours)**: (A) strictly parallel content fixes semantics, layout, and answer across 10 languages so only the language varies; (B) a `vision` setting embeds text inside the image, matching how agents perceive real interfaces; (C) four complementary tasks jointly measure cross-lingual consistency from multiple dimensions. **(3) Key findings**: OCR is a shared cross-lingual bottleneck under `vision`, motivating our label-free OCR-centric GRPO recipe whose gains transfer to external benchmarks (CC-OCR, MTVQA, OCRBench).

Using PM^{4}Bench, we evaluate 10 LVLMs covering leading commercial APIs and open-source models, and observe pronounced performance drops alongside sharper cross-lingual disparities under the `vision` setting. Since `vision` differs from `traditional` only in that text now lives *inside* the image, OCR is the natural candidate for explaining the gap. Further analysis supports this in §6.2: a controlled experiment that supplies ground-truth text alongside the vision input visibly narrows the gap, and per-language OCR task scores correlate strongly with VQA performance. Together they identify **OCR as a key bottleneck** behind cross-lingual disparity and performance drop. Base-model capability remains the upper bound, but OCR is a high-leverage axis we can intervene on cheaply.

Motivated by this, we propose an OCR-centric reinforcement learning recipe using Group Relative Policy Optimization (GRPO) (Shao et al., 2024) on `Qwen3-VL-8B-Thinking`. The core design is a **dual reward** that strengthens OCR while preserving the model's general capability: an OCR accuracy term sharpens recognition, and a length-/format term keeps the chain of thought intact, since pure-accuracy optimization otherwise collapses CoT and degrades reasoning-heavy tasks (Table 4). **Trained on fully synthesized, label-free OCR data, the resulting model improves performance across tasks, settings, and languages on PM^{4}Bench, and the gains transfer to multilingual benchmarks beyond PM^{4}Bench, evidencing genuine generalization.** Upgrading fundamental OCR skills is thus a scalable, low-cost pathway to boost the model's multilingual performance prior to release.

In summary, our work forms a complete evaluation-to-improvement loop toward more equitable global deployment of LVLMs: we provide a fair evaluation yardstick, identify a high-leverage optimization direction, and demonstrate a low-cost mitigation strategy.

- **From evaluation gap to benchmark construction.** We introduce **PM^{4}Bench**, the first 10-language strictly parallel-corpus benchmark with a **`vision` setting** that fuses text and images into a single visual input, enabling apples-to-apples cross-lingual evaluation from diverse capability dimensions.

- **From benchmark evaluation to bottleneck diagnosis.** We evaluate 10 LVLMs and

find that the `vision` setting is harder than the `traditional` interleaved-input, exhibits sharper cross-lingual disparities on all tasks and capability dimensions. Further analysis identifies **OCR capability** as a key bottleneck behind these disparities.

- **From bottleneck diagnosis to targeted mitigation.** We propose an OCR-centric GRPO training strategy using **label-free synthetic OCR data**, improving performance across tasks, settings, and languages on PM^{4}Bench while yielding transferable gains in general multilingual capabilities beyond PM^{4}Bench.

2 Related Work

LVLM Benchmark Early benchmarks (Liu et al., 2024a; Fu et al., 2023; Li et al., 2023) relied on MCQ or short VQA; recent work shifts to open-ended evaluation (MMDU (Liu et al., 2024d)) and pure visual input (MMMU-pro (Yue et al., 2024)).

Multilingual Benchmark Existing multilingual LLM benchmarks typically translate English datasets (Shi et al., 2022; Hasan et al., 2021); recent efforts like P-MMEval (Zhang et al., 2024) and BenchMAX (Huang et al., 2025) use parallel corpora for fair cross-lingual comparison.

Multilingual LVLM Benchmark Existing multilingual LVLM benchmarks evaluate visual perception (Pfeiffer et al., 2021; Liu et al., 2021; Changpinyo et al., 2022; Bugliarello et al., 2022), text-centric VQA (Tang et al., 2024a), reasoning (M3Exam (Zhang et al., 2023), EXAMS-V (Das et al., 2024)), and culture-specific knowledge (CVQA (Romero et al., 2024), M5-VLOD (Schneider and Sitaram, 2024), ALM-bench (Vayani et al., 2024)). Non-parallel corpora, however, conflate language ability with data artifacts, while parallel-corpus benchmarks (M4U (Wang et al., 2024), MMMB (Sun et al., 2025), XT-VQA (Yu et al., 2024)) rely on MCQs or short QAs and lack pure-vision input.

GUI-Agent Benchmarks and the Multilingual Gap Coordinate-based GUI grounding has become a core capability for multimodal agents (ScreenSpot/V2 (Cheng et al., 2024), ScreenSpot-Pro (Li et al., 2025), Mind2Web (Deng et al., 2023), VisualWebArena (Koh et al., 2024), OSWorld (Xie et al., 2024)), yet these resources are overwhelmingly English-only. MPR-GUI (Chen et al., 2025) contributes parallel multilingual GUI *perception and reasoning* but no coordinate grounding. macOSWorld (Yang et al., 2025) provides a comprehensive 5-language interactive macOS environment with coordinate sub-tasks. However, studying cross-lingual capability *imbalance* requires a corpus that is parallel both visually and linguistically across runs, which an interactive multi-step environment cannot maintain because the UI state diverges across trajectories.

3 PM^{4}Bench

3.1 Design Principles

PM^{4}Bench is built on four principles: (1) **Targeted Language Selection**, covering diverse families and writing scripts; (2) **Parallel Corpus**, semantically identical content across languages so any score gap reflects model capability rather than corpus variance; (3) **Vision Setting**, with text and queries "printed" onto images to simulate real applications; (4) **Task Diversity**, spanning competencies relevant to LVLM and agent deployment. Rendering-side confounds are controlled by sharing identical templates, font families, decorations, and backgrounds across languages; inherent script-level differences (length, density, direction) are deliberately preserved so scores reflect real-world deployment challenges.

3.2 Language Selection

PM^{4}Bench supports 10 carefully selected languages: en, zh, ko, th, vi, ru, hu, sr, cs, ar. Their graph complexity, quantified following Graph-Com (Chang et al., 2018) in Appendix C.

3.3 Task Introduction

Following Section 3.1, PM^{4}Bench includes a parallel 10-language corpus across four tasks: Multi-Discipline Understanding and Reasoning (MDUR), Multi-Image Question Answering (MIQA), Multi-Scale OCR Challenge (MSOCR), and Multilingual GUI Grounding (MGUI). Task statistics are summarized in Table 2; per-task examples are in Appendix A.

(1) Multi-Discipline Understanding and Reasoning (MDUR) assesses college-level multimodal reasoning. We source 1,730 English multiple-choice questions from open-sourced MMMU-pro (Yue et al., 2024) and translate them into 9 additional languages (details in Section 3.5). Since MMMU-pro's native vision-only variant is English-restricted, we built an HTML-to-screenshot pipeline that embeds text and images

Benchmark	#Lang.	Parallel Corpus	Pure-Vision Input	Knowledge VQA	Open-end. Gen.	OCR Challenge	Coord. GUI
xGQA (Pfeiffer et al., 2021)	8	Q ✓, A ✗*	×	✓	×	×	×
MaRVL (Liu et al., 2021)	5	×	×	✓	×	×	×
XVNLI (Bugliarello et al., 2022)	5	×	×	✓	×	×	×
xFlickrCO (Bugliarello et al., 2022)	8	×	×	×	✓	×	×
MaXM (Changpinyo et al., 2022)	7	×	×	✓	×	×	×
M3Exam (Zhang et al., 2023)	9	×	×	✓	×	×	×
EXAMS-V (Das et al., 2024)	11	×	✓	✓	×	×	×
MTVQA (Tang et al., 2024a)	9	×	×	×	×	✓	×
CVQA (Romero et al., 2024)	31	×	×	✓	×	×	×
M4U (Wang et al., 2024)	3	T ✓, I ✗	×	✓	×	×	×
MMMB (Sun et al., 2025)	6	✓	×	✓	×	×	×
M5-VLOD (Schneider and Sitaram, 2024)	12	×	×	✓	×	×	×
ALM-bench (Vayani et al., 2024)	100	×	×	✓	✓	×	×
XT-VQA (Yu et al., 2024)	3	T ✓, I ✗	×	×	×	✓	×
macOSWorld (Yang et al., 2025)	5	✓‡	×	×	×	×	✓‡
MPR-GUI (Chen et al., 2025)	6	✓	×	×	×	×	×
PM^{4}Bench (ours)	**10**	✓	✓	✓	✓	✓	✓

Table 1: Comparison of PM^{4}Bench against related multilingual multimodal benchmarks. *: questions translated, answers English. ‡: macOSWorld provides parallel instructions and coordinate sub-tasks but its multi-step trajectories cause UI state to diverge across runs.

Task	#sample (per lang.)	QA Type	Settings
MDUR	1730	MCQ	trad./vis.
MIQA	218	OE	trad./vis.
MSOCR	100	OCR	vis.
MGUI	200	Coord. Grounding	vis.

Table 2: The 4 tasks of PM^{4}Bench. Detailed explanation of trad./vis. settings is in Section 3.4.

into webpages with randomized fonts and backgrounds, guaranteeing identical layouts across languages. MDUR yields 1,730 cases per language under both `traditional` (interleaved) and `vision` (single rendered image) input formats.

(2) Multi-Image Question Answering (MIQA) targets open-ended QA over multi-image inputs. We sampled 218 QA pairs from MMDU (Liu et al., 2024d), a Wikipedia-sourced multi-turn multi-image benchmark, and translated them into 9 languages. Both `traditional` and `vision` settings are provided; for `vision`, question text and reference images are placed on a 1280-pixel-wide canvas.

(3) Multi-Scale OCR Challenge (MSOCR) probes multilingual text-recognition limits. Each image contains 20 lines of semantically meaningless words from Wikipedia parallel corpora rendered with decreasing font sizes (40 down to 2); the recognition threshold is the first line where errors occur. We provide 100 images per language, semantically identical across languages. MSOCR offers a strictly parallel visual corpus and a fully automated, easily extensible construction; in our cross-lingual-balance focus, this lets us evaluate the OCR dimension cheaply and apples-to-apples across languages.

(4) Multilingual GUI Grounding (MGUI) evaluates an LVLM as a multilingual GUI agent, following the coordinate-based grounding paradigm of ScreenSpot (Cheng et al., 2024). We hand-author 100 reusable HTML/Jinja webpage templates spanning 10 common product surfaces (*e.g.*, flight search, seat map, payment; full list in Table 6) and programmatically render each in all 10 languages on a fixed 1280×800 layout. Each image is paired with two instructions, so each language has 200 test items in total. Every natural-language instruction is indirect (*e.g.*, "*I'm emigrating abroad and won't need a return ticket, switch the trip type so I'm only booking the outbound leg.*"), without leaking the text label of the target button, and the model must emit a single click coordinate (x, y) inside the target element's bounding box. Like MSOCR, the MGUI corpus is fully automated and human-translation-free, making it cheap to scale.

3.4 Settings

To better reveal the capabilities of LVLMs and help in-depth analysis, we adopt diverse settings across tasks (see Table 2 for a summary). **(1) `Vision` setting.** The LVLM receives a single image containing all the information necessary to complete the task; the text prompt is minimal, typically a concise instruction such as "Answer the question in the image." This setup encompasses all four tasks. **(2) `Traditional` setting.** The textual content of the questions and the embedded images are provided as

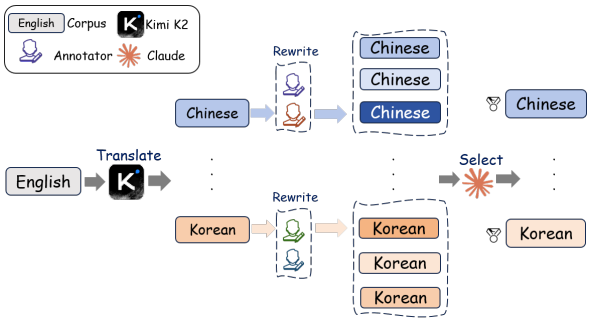


Figure 2: The translation of PM^{4}Bench's parallel texts involves three steps: 1) Kimi K2 Reference Generation; 2) Human Rewriting; 3) Claude Selection.

separate inputs (MSOCR and MGUI are excluded by construction, as they only meaningfully exist as a single image).

3.5 Translation Pipeline

To ensure data quality, we adopt an LLM-and-human-expert in-loop pipeline for the parallel corpus of MDUR and MIQA. As shown in Fig. 2, it has 3 stages: reference translation, manual translation, and selection.

We first use Kimi K2 (Team et al., 2025a), which is not in our evaluation set, to generate a reference translation. The original English plus the reference are then given to two independent groups of native-speaker annotators (also proficient in English); each group internally verifies and submits its own translation. Finally, the original English plus the 3 candidates (1 machine + 2 human) are passed to Claude Sonnet 4.5 (Anthropic, 2025) (also outside our evaluation set) to pick the best. For MIQA, 64% of selected translations were human; for MDUR, this exceeded 99%. The gap reflects task content: MDUR's items lean on rigorous numerical and domain-specific phrasing where minor mistranslations can flip correctness, so human candidates dominate; MIQA's plainer descriptive prose lets machine and human candidates compete on near-equal footing.

4 OCR-centric GRPO Reinforcement Learning

Results in Section 6.2 indicate that multilingual OCR robustness is a key factor behind cross-lingual disparity in LVLMs. To directly target this factor, we design a simple yet effective OCR-centric reinforcement learning strategy for `Qwen3-VL-8B-Thinking` based on Group Relative Policy Optimization (GRPO) (Shao et al., 2024), which boosts other non-OCR tasks and mitigates cross-lingual imbalance. We chose GRPO instead of SFT because OCR tasks are inherently unsuitable for SFT-based post-training, which can quickly degrade the model's performance on other tasks.

Synthetic multilingual OCR training data. We automatically construct a synthetic OCR dataset without any human annotation. For each sample, we first choose a random background image and resolution/aspect ratio, then render multiple lines of text, one per language. Each line is composed of semantically meaningless noun combinations from Wikipedia, so that no language receives an advantage from specific semantics. Font style and font size are also randomly perturbed. Refer to Appendix A.5 for examples.

GRPO reward design. Our reward function is designed to enhance multilingual OCR capabilities without compromising general abilities. Let s denote the full generated sequence (chain-of-thought plus final answer) with length $l = |s|$, and let y be the answer extracted from the `\boxed{}` span within s. The ground-truth is denoted by y^*. Our total reward consists of an accuracy reward R_{acc} and a formatting reward R_{fmt} that regulates output length.

1) Format and Length Constraint R_{fmt} Two issues, identified through case studies on early training rollouts, motivate this term: overly long or under-structured outputs degrade general task performance, and pure OCR accuracy optimization causes the model to abandon chain-of-thought reasoning, degrading general performance. We define a target length interval $[L_{\min}, L_{\max}]$ and apply a penalty $P_{\text{len}}(l)$ that scales linearly under $L_{\min}$ and logarithmically over $L_{\max}$:

$$P_{\text{len}}(l) = \begin{cases} 1 - l/L_{\min}, & l < L_{\min}, \\ 0, & L_{\min} \le l \le L_{\max}, \\ \min\big(1, \log(1 + \frac{l - L_{\max}}{\lambda})\big), & l > L_{\max}. \end{cases} \tag{1}$$

where λ is a scale parameter. Additionally, to prevent degenerate repetition, we compute the unique 4-gram ratio $\rho(s) = |\mathcal{N}_4^{\text{uniq}}|/|\mathcal{N}_4^{\text{all}}|$. A penalty $P_{\text{rep}}(s)$ is applied if this diversity metric falls below a threshold τ:

$$P_{\text{rep}}(s) = \max\left(0,\ 1 - \frac{\rho(s)}{\tau}\right). \tag{2}$$

The combined formatting reward $R_{\text{fmt}}(s)$ encourages sequences within the target length while penalizing undesirable behaviors:

$$R_{\text{fmt}}(s) = \alpha_1 \mathbb{I}_{[L_{\min}, L_{\max}]}(l) - \alpha_2 P_{\text{len}}(l) - \alpha_3 P_{\text{rep}}(s). \tag{3}$$

where $\mathbb{I}_{[\cdot]}(\cdot)$ is the indicator function, and $\alpha_1, \alpha_2, \alpha_3$ are weighting coefficients.

2) OCR Accuracy Reward R_{acc} We measure the correctness of the extracted answer y using the normalized Levenshtein distance, denoted as $\mathcal{D}_{\text{lev}}(\cdot,\cdot)$:

$$R_{\text{acc}}(y, y^*) = 1 - \frac{\mathcal{D}_{\text{lev}}(y, y^*)}{\max(|y|, |y^*|)}. \quad (4)$$

3) Total Reward. The final reward is a weighted sum emphasizing accuracy while maintaining structural constraints:

$$R(s, y^*) = \omega_1 \cdot R_{\text{acc}}(y, y^*) + \omega_2 \cdot R_{\text{fmt}}(s), \quad (5)$$

where ω_1 and ω_2 are weighting hyper-parameters.

5 Experiment Settings

5.1 Evaluation Protocol

To comprehensively compare the performance of various kinds of LVLMs, we include the following 10 models in our experiment. Leading commercial APIs: Gemini-3.1-Pro-Preview (Google DeepMind, 2025), GPT-5 (Singh et al., 2025), GPT-5-mini (Singh et al., 2025), Doubao-Seed-1.6 (ByteDance Seed, 2025); and open-source models: Llama-4-Maverick (Meta AI, 2025), GLM-4.5v (Team et al., 2025b), and Qwen3-VL series (Bai et al., 2025). All models are evaluated with **greedy decoding** under default chat template. Training converges within 200 optimization steps, making the full post-training pipeline highly compute-efficient. Detailed inference setup, training hyper-parameters, and per-task prompts are deferred to Appendix H and Appendix B.1.

5.2 Metrics

For the performance score of MDUR, we evaluate the correct ratio $S^{\text{MDUR}} = N_{\text{cor}}/N_{\text{total}}$. For MIQA, we adopt the LLM-as-judge approach from MMDU, using DeepSeek-v4-flash (DeepSeek-AI, 2026) (outside our evaluation set and model families) to rate responses across six dimensions: *Creativity*, *Richness*, *Visual Perception*, *Logical Coherence*, *Answer Accuracy*, *Image Relationship Understanding*, against human-provided references. The final score averages these six dimensions: $S^{\text{MIQA}} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{6}\sum_{d\in D} S_i^{(d)}\right)$. Appendix E.1 confirms the reliability of this judge against the disparity of language capability of judge model; Appendix E.2 additionally reports a cross-validation with Claude-Opus-4.7 (Anthropic, 2026) as judge: per-language scores agree with DeepSeek-v4-flash judge with minor variance. The score for MSOCR is $S^{\text{MSOCR}} = \frac{1}{N}\sum_{i=1}^{N}(40 - s_i)$, where s_i is the font size of the line where image i first errs, and the constant 40 equals the maximum font size, ensuring scores start at 0. For MGUI, we measure click accuracy: each item asks the model to emit a coordinate (x, y), and the prediction is correct if it lies inside the ground-truth bounding box, $S^{\text{MGUI}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\big[(x_i, y_i) \in \text{bbox}_i\big] \times 100$, following ScreenSpot (Cheng et al., 2024). Coordinates emitted in the conventional 1000×1000 normalized space are rescaled to the 1280×800 canvas before the bbox check.

6 Results & Findings

6.1 Overall performance & cross-lingual disparity

Table 3 reports overall results, with average score ($S_{\text{avg.}}$) and coefficient of variation ($S_{\text{cv.}} = \sigma/\mu \times 100\%$, measures cross-lingual variance) computed across 10 languages. Per-language results: Appendix D.

Gemini-3.1-Pro-Preview and the Qwen3-VL-235B series lead on most tasks. Yet Fig. 3 shows **cross-lingual consistency remains a challenge**, with imbalance especially pronounced on MSOCR: every model exhibits a wide score spread across languages, and low-resource, visually dense scripts (AR, TH) consistently lag high-resource ones (EN, ZH) for both commercial APIs and open-source models. Under `vision` (Fig. 4a), $S_{\text{avg.}}$ on MDUR/MIQA drops and 90%/80% of models show higher $S_{\text{cv.}}$ than under `traditional`, so the **`vision` setting compromises both performance and cross-lingual balance**.

We transparently report an anomaly: GPT-5 and GPT-5-mini lead on MDUR/MIQA yet their MGUI accuracy collapses to 10-15%. We rule out base-capability deficit, evaluation bug, and data issue: a **MGUI$_{\text{text}}$** sanity check (Appendix D.1) shows GPT-5 reaches 94.0 when asked for the visible text of the click target instead of a coordinate, and GPT-5's median bbox-edge distance is only 33 px (∼2.2% of diagonal). These signals confirm GPT-5 *understands* the MGUI task and *correctly locates* the target; the failure is solely in emitting precise pixel coordinates, consistent with its strong MDUR/MIQA results.

Model	MDUR				MIQA				MSOCR		MGUI	
	$S_{avg.}\uparrow$		$S_{cv.}\downarrow$		$S_{avg.}\uparrow$		$S_{cv.}\downarrow$		$S_{avg.}\uparrow$	$S_{cv.}\downarrow$	$S_{avg.}\uparrow$	$S_{cv.}\downarrow$
	trad.	vision	trad.	vision	trad.	vision	trad.	vision	vision	vision	vision	vision
Gemini-3.1-Pro-Preview	**80.24**	**78.45**	0.018	**0.024**	60.27	53.11	0.097	0.077	**30.38**	**0.238**	87.10	0.026
GPT-5	72.89	69.03	**0.010**	0.037	54.82	55.95	0.057	0.046	16.31	0.594	10.56	0.264
GPT-5-mini	69.78	69.13	0.013	0.028	**62.98**	**61.55**	**0.032**	**0.044**	13.58	0.676	14.83	0.185
Llama-4-Maverick	57.50	48.29	0.026	0.077	46.07	44.05	0.071	0.080	8.27	0.943	54.80	0.214
GLM-4.5v	55.03	46.96	0.054	0.272	50.66	43.37	0.060	0.296	7.70	1.398	74.40	0.186
Doubao-Seed-1.6	59.59	54.40	0.054	0.176	55.06	47.04	0.065	0.233	8.58	1.100	81.45	0.109
Qwen3VL-235B-Thinking	65.04	61.53	0.024	0.062	61.56	54.35	0.041	0.126	15.99	0.605	**89.15**	**0.024**
Qwen3VL-235B-Instruct	61.07	56.49	0.037	0.083	61.96	59.52	0.131	0.162	18.34	0.569	82.10	0.077
Qwen3VL-32B-Thinking	64.04	60.46	0.017	0.098	54.98	49.46	0.062	0.160	11.37	0.794	86.35	0.048
Qwen3VL-8B-Thinking	38.55	34.88	0.061	0.218	53.63	47.69	0.072	0.216	1.53	1.145	78.30	0.130
QGO-8B (ours)	46.82	40.84	0.080	0.190	55.24	51.06	0.060	0.174	8.17	0.759	80.00	0.095

Table 3: Overall model performance comparison on MDUR, MIQA, MSOCR and MGUI, where $S_{avg.}$ is the average score across 10 languages, and $S_{cv.}$ refers to the coefficient of variance among 10 languages. Best values in **bold**, second best underlined. The Qwen3-VL-8B-Thinking model after OCR-centric reinforcement learning is denoted as QGO-8B. Qwen3-VL-235B-A22B-Thinking/Instruct is referred as Qwen3VL-235B-Thinking/Instruct for brevity.

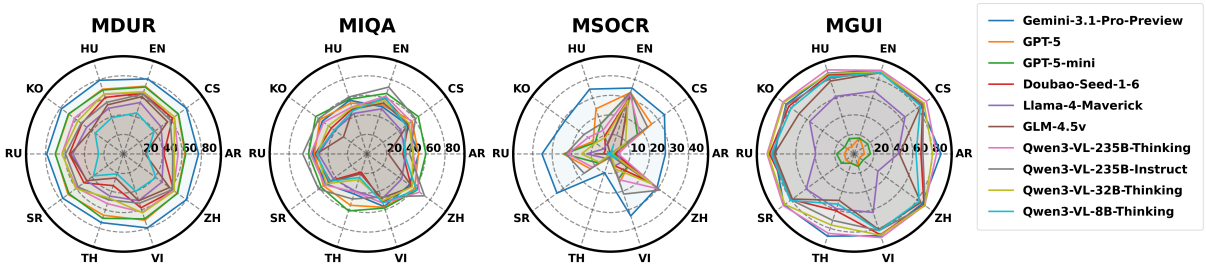
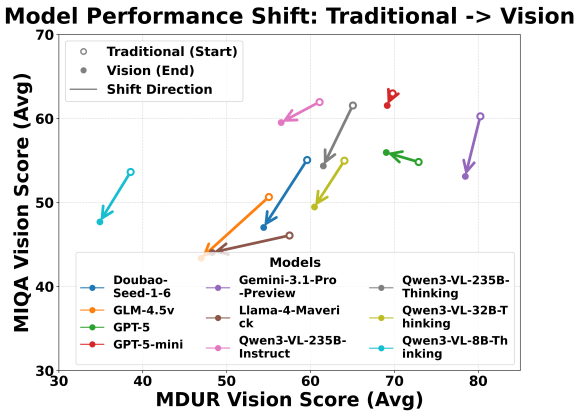
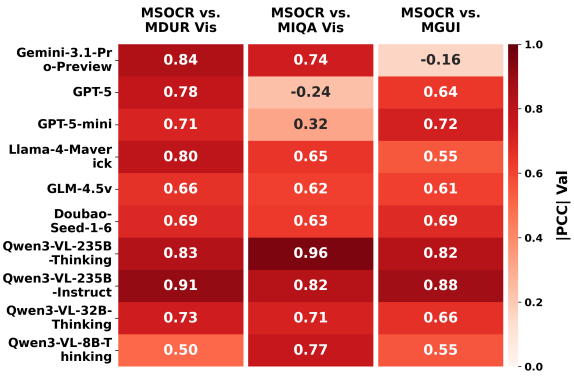


Figure 3: Scores (`vision` setting) of each model and each language across all four tasks of PM^{4}Bench.



(a) Performance shift from `traditional` to `vision` setting.



(b) Per-model Pearson Correlation Coefficient between MSOCR and three downstream-task score vectors across 10 languages. Computation details: Appendix G

6.2 OCR is a key bottleneck causing performance gaps

Since `vision` differs from `traditional` mainly in that text now lives inside the image, OCR is the natural candidate for explaining the gap. We test this hypothesis with three converging signals.

(i) Controlled experiment. Supplying ground-truth texts alongside `vision` input (Table 5) lifts $S_{avg.}$ and lowers $S_{cv.}$ for both commercial APIs and the Qwen3-VL series, with larger effects on smaller models. **(ii) Per-language correlation.** Across the 10 languages, **most evaluated models exhibit** $|\mathbf{PCC}| > 0.5$ between MSOCR and other-task score vectors, indicating a strong corre-

Model	MDUR				MIQA				MSOCR		MGUI	
	$S_{\text{avg.}} \uparrow$		$S_{\text{cv.}} \downarrow$		$S_{\text{avg.}} \uparrow$		$S_{\text{cv.}} \downarrow$		$S_{\text{avg.}} \uparrow$	$S_{\text{cv.}} \downarrow$	$S_{\text{avg.}} \uparrow$	$S_{\text{cv.}} \downarrow$
	trad.	vision	trad.	vision	trad.	vision	trad.	vision	vision	vision	vision	vision
Original Qwen3	38.55	34.88	0.061	0.218	53.63	47.69	0.072	0.216	1.53	1.145	78.30	0.130
QGO-8B (only-fmt)	42.05	36.81	0.073	0.219	**55.25**	49.48	0.062	0.207	2.35	1.244	78.90	0.121
QGO-8B (only-ocr)	38.89	26.67	**0.051**	0.273	53.90	47.40	0.066	0.217	8.13	**0.738**	68.50	0.225
QGO-8B (full)	**46.82**	**40.84**	0.080	**0.190**	55.24	**51.06**	**0.060**	**0.174**	**8.17**	0.759	**80.00**	**0.095**

Table 4: Ablation study of the proposed reward functions on the MDUR, MIQA, MSOCR and MGUI tasks across 10 languages. "only-fmt" and "only-ocr" denote models trained exclusively with the Format and Length Constraint (R_{fmt}) or the OCR Accuracy Reward (R_{acc}), respectively.

Model	w_o_text	w_text	Δ Avg	Δ CV
GPT-5	69.03	71.10	+2.07	-2.46%
GPT-5-mini	69.13	70.71	+1.58	-1.39%
Gemini-3.1-Pro-Preview	78.45	80.45	+1.99	-0.48%
Qwen3-VL-235B-Thinking	61.53	66.97	+5.44	-4.42%
Qwen3-VL-8B-Thinking	34.88	41.75	+6.87	-17.64%

Table 5: MDUR vision performance comparison. **w_o_text**: Original vision input. **w_text**: With text reference provided. Δ Avg shows score improvement, Δ CV shows change in cross-language consistency (lower is better).

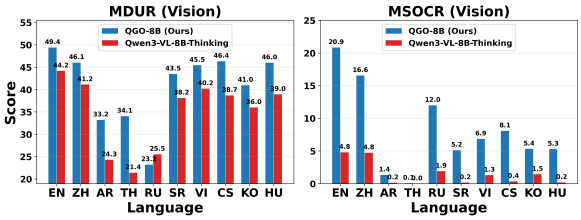


Figure 5: Per-language comparison of QGO-8B against the Qwen3-VL-8B-Thinking baseline on MDUR and MSOCR (`vision`).

lation between text recognition and downstream VQA (Fig. 4b). **(iii) Causal closure via intervention.** The OCR-centric RL of §4 lifts non-OCR tasks, identifying **OCR as a key factor driving performance drop and cross-lingual disparity on `vision` set**.

6.3 Effect of OCR-centric GRPO reinforcement learning

Against Qwen3-VL-8B-Thinking, the trained model **QGO-8B improves $S_{\text{avg.}}$ and lowers $S_{\text{cv.}}$ (more balanced among languages) on every task and setting** (Table 3). The largest per-language gains concentrate on the baseline's weakest languages AR and TH (MDUR and MSOCR shown in Fig. 5; MIQA and MGUI deferred to Fig. 13 in Appendix A.6). Notably, only MSOCR shares its OCR objective with our synthetic training data; MDUR, MIQA, and MGUI tasks are not trained during RL, so the uniform across-task improvement reflects genuine transfer.

External-benchmark transfer. On three external suites (Appendix F), QGO-8B beats baseline on **CC-OCR** (Liu et al., 2024c) (NED +4.7 overall, +4.0 on six unseen languages FR/DE/IT/JA/PT/ES), matches baseline on **OCRBench** (Liu et al., 2024b) (791 vs 789), and gains +0.5 $S_{\text{avg.}}$ on **MTVQA** (Tang et al., 2024a) trained languages. Gains thus generalize beyond PM^{4}Bench.

6.4 Ablation Study

We isolate R_{fmt} and R_{acc} to verify both are essential. Table 4 shows full QGO-8B wins overall: R_{fmt} keeps the CoT alive (without it MDUR collapses to 26.67) and R_{acc} supplies perception (alone lifts MSOCR 1.53 → 8.13). Crucially, R_{acc} alone *destroys* generalization: MDUR drops below baseline and the MGUI task falls from 78.30 to 68.50, because pure OCR accuracy optimization collapses the chain-of-thought. Only the **joint** reward wins on every task: on the MGUI task, baseline 78.30 → *only-fmt* 78.90 / *only-ocr* 68.50, while full QGO-8B reaches 80.00, confirming the format reward is co-essential.

7 Conclusion

We introduce PM^{4}Bench, a 10-language strictly parallel-corpus benchmark challenging multiple capability dimensions under a `vision` setting that simulates real-world applications such as AI agents. Across 10 LVLMs it exposes notable cross-lingual disparities in which OCR serves as a key bottleneck. Our label-free OCR-centric GRPO lifts non-OCR tasks across languages, with gains transferring to external benchmarks.

8 Limitations

Scope of RL gains. Our OCR-centric GRPO yields meaningful but bounded improvements: it lifts `vision`-setting performance on PM^{4}Bench across all four tasks and transfers to CC-OCR/MTVQA/OCRBench, yet on the latter two suites the gains are modest (matching baseline on OCRBench, $+0.5$ on MTVQA trained-language $S_{\text{avg.}}$). Genuine cross-lingual capability ultimately requires base-model improvements; OCR-centric RL is a high-leverage but partial intervention.

Synthetic-data scope. MSOCR and the MGUI task are intentionally small, fully synthetic decoupling probes designed to isolate cross-lingual capability differences from data artifacts (cultural variance, translation drift, content-coverage imbalance). They are not a replacement for specialized OCR or GUI benchmarks (*e.g.*, dedicated text-spotting/KIE suites, interactive multi-step environments such as macOSWorld) and should not be read as such. PM^{4}Bench complements rather than supersedes those resources.

Judge reliability. MIQA evaluation uses an LLM-as-judge protocol with a single judge model. We provide both a translate-to-English consistency check (Appendix E.1) and a Claude-Opus-4.7 cross-judge cross-validation (Appendix E.2); residual per-dimension and inter-judge variability is left for future work.

Synthetic-to-real gap. We acknowledge that MSOCR and MGUI inevitably differ from real-world OCR and GUI scenarios, and cannot fully simulate the complexity of real deployments. These synthetic tasks are designed primarily to fairly measure cross-lingual disparity under controlled conditions, rather than to precisely estimate absolute real-world performance. Our goal is therefore to make the performance patterns on synthetic data approximate real-world distributions as closely as possible, while leaving more faithful real-world evaluation to other benchmarks.

Potential risks. PM^{4}Bench is intended as a research artifact for diagnosing and mitigating cross-lingual capability disparities of LVLMs, but we acknowledge the existence of potential risks. **Data misuse.** Although our parallel corpus is sourced from publicly available datasets and rendered with synthetic templates, redistribution or downstream usage outside the intended evaluation context could amplify any residual biases in the source data. We therefore release PM^{4}Bench strictly for non-commercial research and encourage users to audit cultural and demographic balance before any deployment-oriented use. **Evaluation misinterpretation.** Scores on PM^{4}Bench reflect controlled, parallel-corpus settings and should not be over-generalized as an absolute measure of real-world multilingual deployment readiness; we encourage users to pair our diagnostics with deployment-domain evaluation before drawing operational conclusions.

References

Anthropic. 2025. Claude sonnet 4.5. https://www.anthropic.com/news/claude-sonnet-4-5. Accessed: 2026-05-25.

Anthropic. 2026. Claude opus 4.7 system card. https://www.anthropic.com/claude-opus-4-7-system-card. Accessed: 2026-05-25.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Emanuele Bugliarello, Fangyu Liu, Jonas Pfeiffer, Siva Reddy, Desmond Elliott, Edoardo Maria Ponti, and Ivan Vulić. 2022. IGLUE: A benchmark for transfer learning across modalities, tasks, and languages. In *ICML*.

ByteDance Seed. 2025. Doubao seed 1.6: Latest general-purpose model series. https://seed.bytedance.com/en/seed1_6. Accessed: 2026-03-02.

Li-Yun Chang, Yen-Chi Chen, and Charles A Perfetti. 2018. Graphcom: A multidimensional measure of graphic complexity applied to 131 written languages. *Behavior research methods*, 50:427–449.

Soravit Changpinyo, Linting Xue, Michal Yarom, Ashish V Thapliyal, Idan Szpektor, Julien Amelot, Xi Chen, and Radu Soricut. 2022. Maxm: Towards multilingual visual question answering. *arXiv preprint arXiv:2209.05401*.

Chen and 1 others. 2025. MPR-GUI: Benchmarking and enhancing multilingual perception and reasoning in GUI agents. *arXiv preprint arXiv:2512.00756*.

Kanzhi Cheng, Qiushi Sun, Yougang Chu, Fangzhi Xu, Yantao Li, Jianbing Zhang, and Zhiyong Wu. 2024. Seeclick: Harnessing gui grounding for advanced visual gui agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Rocktim Jyoti Das, Simeon Emilov Hristov, Haonan Li, Dimitar Iliyanov Dimitrov, Ivan Koychev, and Preslav Nakov. 2024. Exams-v: A multi-discipline multilingual multimodal exam benchmark for evaluating vision language models. *arXiv preprint arXiv:2403.10378*.

DeepSeek-AI. 2026. Deepseek-v4: Technical report. https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro/blob/main/DeepSeek_V4.pdf.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Sam Stevens, Boshi Wang, Huan Sun, and Yu Su. 2023. Mind2web: Towards a generalist agent for the web. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and 1 others. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.

Google DeepMind. 2025. Gemini 3.1 pro: Model information. https://deepmind.google/models/gemini/pro/. Accessed: 2026-05-25.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Wayne Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in llms: Improving multilingual capability by cross-lingual-thought prompting. *arXiv preprint arXiv:2305.07004*.

Xu Huang, Wenhao Zhu, Hanxu Hu, Conghui He, Lei Li, Shujian Huang, and Fei Yuan. 2025. Benchmax: A comprehensive multilingual evaluation suite for large language models. *arXiv preprint arXiv:2502.07346*.

Jing Yu Koh, Robert Lo, Lawrence Jang, Vikram Duvvur, Ming Chong Lim, Po-Yu Huang, Graham Neubig, Shuyan Zhou, Ruslan Salakhutdinov, and Daniel Fried. 2024. Visualwebarena: Evaluating multimodal agents on realistic visual web tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.

Kaixin Li, Ziyang Meng, Hongzhan Lin, Ziyang Luo, Yuchen Tian, Jing Ma, Zhiyong Huang, and Tat-Seng Chua. 2025. ScreenSpot-Pro: Gui grounding for professional high-resolution computer use. *Preprint*, arXiv:2504.07981.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually grounded reasoning across languages and cultures. In *EMNLP*.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024a. Mmbench: Is your multi-modal model an all-around player? *Preprint*, arXiv:2307.06281.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. 2024b. OCRBench: On the hidden mystery of OCR in large multimodal models. *arXiv preprint arXiv:2305.07895*.

Zhibo Liu, Jiawei Chen, Mingxin Huang, Yuliang Liu, and Lianwen Jin. 2024c. CC-OCR: A comprehensive and challenging OCR benchmark for evaluating large multimodal models in literacy. *arXiv preprint arXiv:2412.02210*.

Ziyu Liu, Tao Chu, Yuhang Zang, Xilin Wei, Xiaoyi Dong, Pan Zhang, Zijian Liang, Yuanjun Xiong, Yu Qiao, Dahua Lin, and 1 others. 2024d. Mmdu: A multi-turn multi-image dialog understanding benchmark and instruction-tuning dataset for lvlms. *arXiv preprint arXiv:2406.11833*.

Meta AI. 2025. Llama 4 maverick: The beginning of a new era of natively multimodal models. https://ai.meta.com/blog/llama-4-multimodal-intelligence/. Accessed: 2026-03-02.

Jonas Pfeiffer, Gregor Geigle, Aishwarya Kamath, Jan-Martin O Steitz, Stefan Roth, Ivan Vulić, and Iryna Gurevych. 2021. xgqa: Cross-lingual visual question answering. *arXiv preprint arXiv:2109.06082*.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, and 1 others. 2024. Cvqa: Culturally-diverse multilingual visual question answering benchmark. *arXiv preprint arXiv:2406.05967*.

Florian Schneider and Sunayana Sitaram. 2024. M5–a diverse benchmark to assess the performance of large multimodal models across multilingual and multicultural vision-language tasks. *arXiv preprint arXiv:2407.03791*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y.K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Shuaijie She, Wei Zou, Shujian Huang, Wenhao Zhu, Xiang Liu, Xiang Geng, and Jiajun Chen. 2024. Mapo: Advancing multilingual reasoning through multilingual alignment-as-preference optimization. *arXiv preprint arXiv:2401.06838*.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2024. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2022. Language models are multilingual chain-of-thought reasoners. *Preprint*, arXiv:2210.03057.

Aaditya Singh and 1 others. 2025. Openai gpt-5 system card. *Preprint*, arXiv:2601.03267.

Hai-Long Sun, Da-Wei Zhou, Yang Li, Shiyin Lu, Chao Yi, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, De-Chuan Zhan, and 1 others. 2025. Parrot: Multilingual visual instruction tuning. In *ICML*.

Jiaxing Sun, Weiquan Huang, Jiang Wu, Chenya Gu, Wei Li, Songyang Zhang, Hang Yan, and Conghui He. 2024. Benchmarking chinese commonsense reasoning of llms: From chinese-specifics to reasoning-memorization correlations. *arXiv preprint arXiv:2403.14112*.

Jingqun Tang, Qi Liu, Yongjie Lin, Yunshui Liu, Yangjie Han, Tao Lin, and 1 others. 2024a. MTVQA: Benchmarking multilingual text-centric visual question answering. *arXiv preprint arXiv:2405.11985*.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024b. Language-specific neurons: The key to multilingual capabilities in large language models. *arXiv preprint arXiv:2402.16438*.

Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, Zhuofu Chen, Jialei Cui, Hao Ding, Mengnan Dong, Angang Du, Chenzhuang Du, Dikang Du, Yulun Du, Yu Fan, and 150 others. 2025a. Kimi k2: Open agentic intelligence. *Preprint*, arXiv:2507.20534.

V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, and 69 others. 2025b. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *Preprint*, arXiv:2507.01006.

Ashmal Vayani, Dinura Dissanayake, Hasindri Watawana, Noor Ahsan, Nevasini Sasikumar, Omkar Thawakar, Henok Biadglign Ademtew, Yahya Hmaiti, Amandeep Kumar, Kartik Kuckreja, and 1 others. 2024. All languages matter: Evaluating lmms on culturally diverse 100 languages. *arXiv preprint arXiv:2411.16508*.

Hongyu Wang, Jiayu Xu, Senwei Xie, Ruiping Wang, Jialin Li, Zhaojie Xie, Bin Zhang, Chuyan Xiong, and Xilin Chen. 2024. M4u: Evaluating multilingual understanding and reasoning for large multimodal models. *arXiv preprint arXiv:2405.15638*.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in english? on the latent language of multilingual transformers. *arXiv preprint arXiv:2402.10588*.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, and 1 others. 2024. OSWorld: Benchmarking multimodal agents for open-ended tasks in real computer environments. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. *Preprint*, arXiv:2010.11934.

Yang and 1 others. 2025. macOSWorld: A multilingual interactive benchmark for GUI agents. *arXiv preprint arXiv:2506.04135*.

Jia Yu, Fei Yuan, Rui Min, Jing Yu, Pei Chu, Jiayang Li, Wei Li, Ruijie Zhang, Zhenxiang Li, Zhifei Ren, Dong Zheng, Wenjian Zhang, Yan Teng, Lingyu Meng, ZhenJiang Jin, Jiantao Qiu, ShaSha Wang, Zhongying Tu, Dahua Lin, and 4 others. 2025. Wanjuansilu: A high-quality open-source webtext dataset for low-resource languages. *Preprint*, arXiv:2501.14506.

Xinmiao Yu, Xiaocheng Feng, Yun Li, Minghui Liao, Ya-Qi Yu, Xiachong Feng, Weihong Zhong, Ruihan Chen, Mengkang Hu, Jihao Wu, and 1 others. 2024. Cross-lingual text-rich visual comprehension: An information theory perspective. *arXiv preprint arXiv:2412.17787*.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, and 1 others. 2024. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*.

Wenxuan Zhang, Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. *Advances in Neural Information Processing Systems*, 36:5484–5505.

Yidan Zhang, Boyi Deng, Yu Wan, Baosong Yang, Haoran Wei, Fei Huang, Bowen Yu, Junyang Lin, and Jingren Zhou. 2024. P-mmeval: A parallel multilingual multitask benchmark for consistent evaluation of llms. *arXiv preprint arXiv:2411.09116*.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024. How do large language models handle multilingualism? *arXiv preprint arXiv:2402.18815*.

Wenhao Zhu, Shujian Huang, Fei Yuan, Cheng Chen, Jiajun Chen, and Alexandra Birch. 2024a. The power of question translation training in multilingual reasoning: Broadened scope and deepened insights. *arXiv preprint arXiv:2405.01345*.

Wenhao Zhu, Shujian Huang, Fei Yuan, Shuaijie She, Jiajun Chen, and Alexandra Birch. 2024b. Question translation training for better multilingual reasoning. *arXiv preprint arXiv:2401.07817*.

A Vision-Setting Input Samples

Unlike traditional interleaved text-image input paradigms, the **Vision Setting** in PM^{4}Bench consolidates all textual information and reference images into a single, unified image as the input for Large Vision-Language Models (LVLMs). This design specifically emulates the authentic operational environments likely to be encountered by AI Agents and Embodied AI Robots.

During autonomous GUI navigation, these agents may receive smartphone or computer interfaces as pure pixel-level visual inputs (images or videos). Consequently, the agent is forced to rely exclusively on visual perception to parse interface semantics, particularly the embedded on-screen text. **This visual parsing becomes exceptionally challenging when the target interface operates in non-English or multilingual contexts.** The dedicated MGUI task in PM^{4}Bench (Table 13) directly probes this regime through coordinate-based grounding on a parallel multilingual UI corpus.

As the pioneering benchmark evaluating multilingual capabilities under vision-only inputs, and, to our knowledge, the first to evaluate *coordinate-based MGUI grounding on a strictly parallel multilingual corpus*, PM^{4}Bench bridges a critical gap in the field. Our data, findings and OCR-centric reinforcement learning method provide essential references for the future development and optimization of AI Agents/Embodied AI Robots, particularly those targeting global, multilingual user bases.

Input examples under both the `traditional` and `vision` settings, for each task and each language, are illustrated in Fig. 6–Fig. 11.

A.1 MDUR Input Samples

The `vision` setting (Fig. 6) renders the question, options, and reference images into one page-level image; the `traditional` setting (Fig. 7) supplies them as text plus separate image files. All 10 languages share the same underlying problem semantics and reference image; only the textual portion is localized.

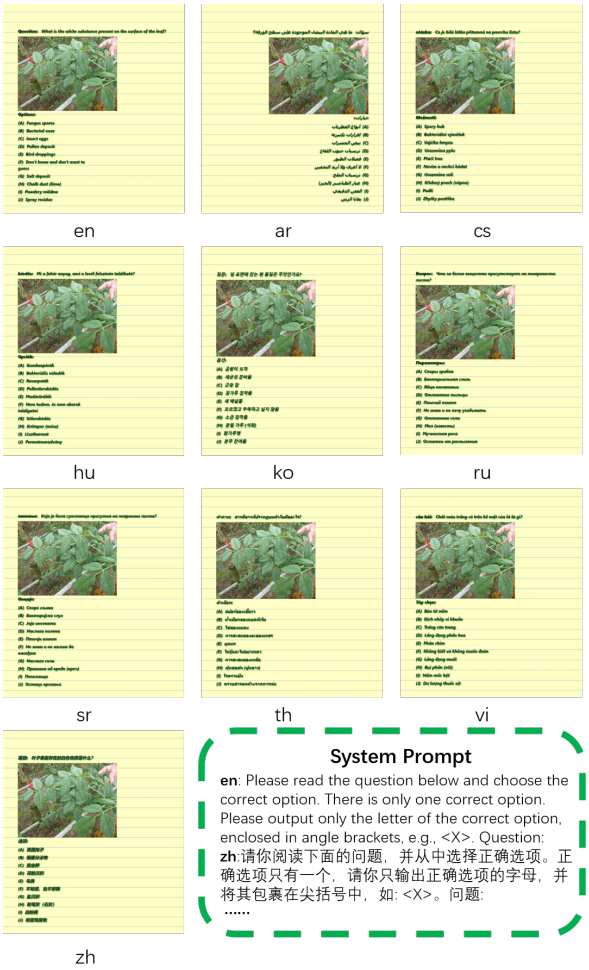


Figure 6: Examples of MDUR's `vision` setting input.

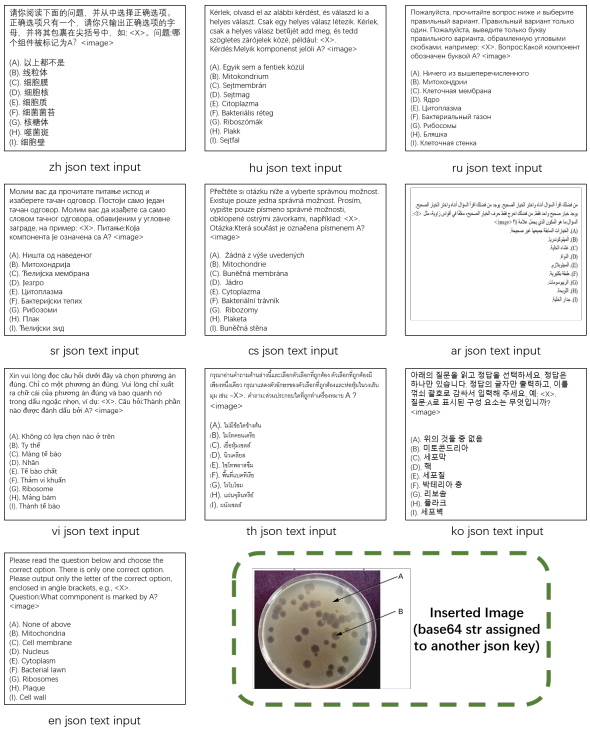


Figure 7: Examples of MDUR's `traditional` setting input.

A.2 MIQA Input Samples

Each MIQA item asks a free-form question with multiple reference images. The `vision` setting (Fig. 8) consolidates the question text and all reference images into a single composite image; the `traditional` setting (Fig. 9) supplies the question as text and the images as separate inputs. The 10 language variants share the same reference images and only differ in question wording.

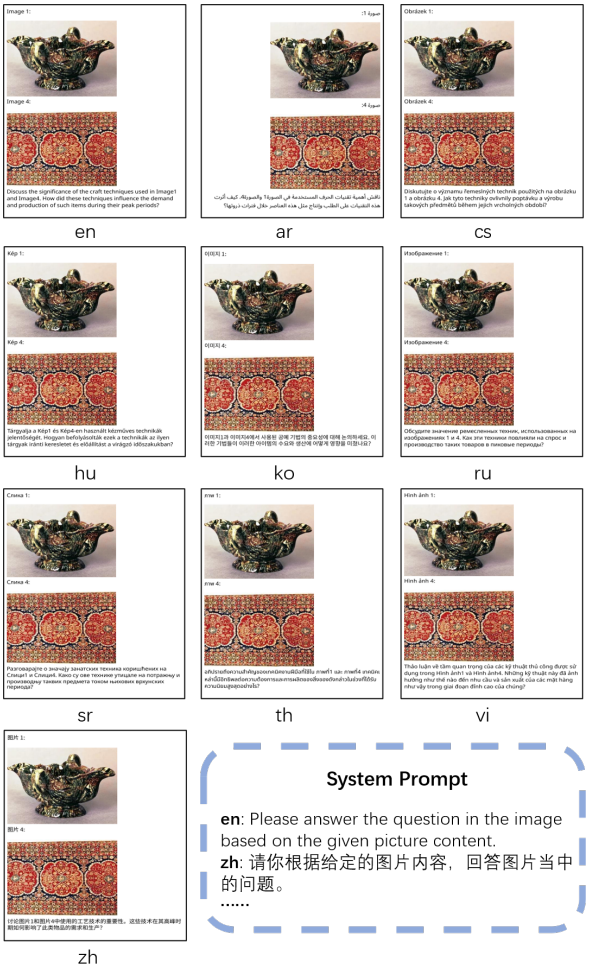


Figure 8: Examples of MIQA's `vision` setting input.

Figure 9: Examples of MIQA's `traditional` setting input.

A.3 MSOCR Input Samples

Each MSOCR sample is a single rendered image containing 20 lines of text in a single target language. The model is asked to transcribe the full content top-to-bottom; we score by exact-match accuracy at the line level. Fig. 10 shows representative examples spanning the 10 supported languages.

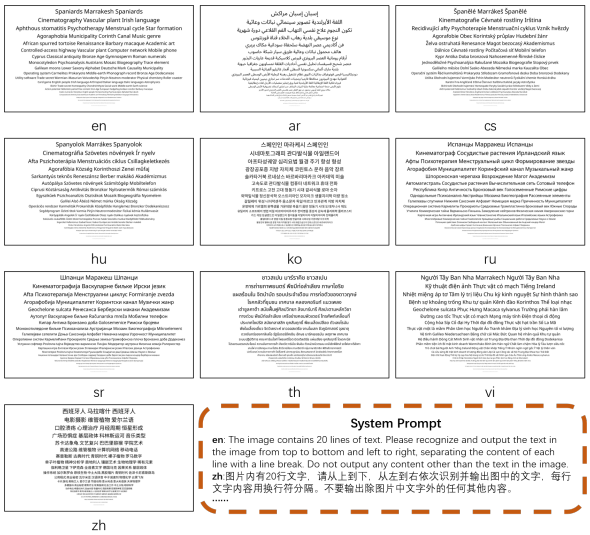


Figure 10: Examples of MSOCR input.

A.4 MGUI Input Samples

The MGUI task is constructed by programmatically rendering a fixed set of HTML/Jinja webpage templates with the same layout in all 10 languages. As shown in Fig. 11, the visual structure (positions and sizes of buttons, fields, banners, etc.) is byte-identical across languages: only the visible text is swapped, so any cross-lingual difference in click accuracy is attributable to the model's multilingual visual understanding rather than dataset bias. The model is asked to emit a single (x, y) coordinate per natural-language instruction; the prediction is judged correct if the point falls inside the ground-truth bounding box of the target interface element.

The 100 hand-authored templates are organized into 10 product-surface *topics* (10 sub-page templates per topic), spanning the most common GUI flows that a multilingual visual agent would encounter in real deployment. Table 6 enumerates the full topic-to-sub-page mapping.

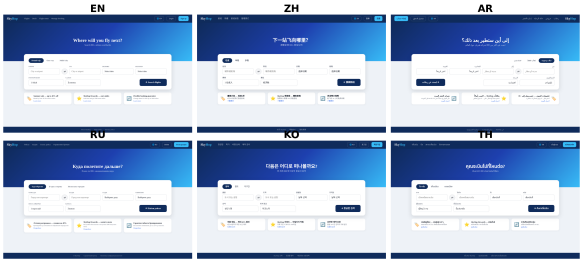


Figure 11: Examples of the MGUI task. The same flight-search webpage template (`a01_search`) is rendered identically in 6 of the 10 supported languages (EN, ZH, AR, RU, KO, TH); only the visible text changes.

Topic (prefix)	10 sub-page templates
Flight booking (a)	Flight Search; Search Results; Flight Detail; Seat Selection; Booking Form; Payment Page; Boarding Pass; Flight Status; Manage Booking; Loyalty Program.
Online banking (b)	Dashboard; Money Transfer; Transaction History; Card Management; Bill Payments; Investment Portfolio; Loan Overview; Budget Planner; Security Center; Support & Help.
Code editor / IDE (c)	Main Editor; Integrated Terminal; Git Panel; Extension Marketplace; Debug View; Global Search; Settings; Merge Conflict Resolver; Refactor Tools; Project Dashboard.
Admin dashboard (d)	Overview; Orders; Customers; Products; Inventory; Analytics; Marketing; Team; Settings; Reports.
E-commerce (e)	Product Detail; Shopping Cart; Checkout; Search Results; Order History; Wishlist; Product Comparison; Reviews; Return/Refund Flow; Account Settings.
Food delivery (f)	Home; Restaurant Menu; Cart; Checkout; Order Tracking; Search Results; Reviews; User Profile; Deals & Coupons; App Settings.
Maps (g)	Explore; Route Planner; Place Detail Card; Public Transit; Saved Places; Traffic & Incidents; Street View; Nearby Businesses; Offline Maps; Map Settings.
Email client (m)	Inbox; Compose; Read View; Folder Management; Advanced Search; Settings; Contacts; Calendar Invite; Label Management; Spam Folder.
Social media (s)	Feed; Profile; Notifications; Messages; Compose; Trending; Groups; Story Viewer; Privacy Settings; Analytics.
Weather (w)	Current Dashboard; 7-Day Forecast; Radar Map; Alerts; Saved Locations; Astronomy; Air Quality; Precipitation Detail; App Settings; Historical Comparison.

Table 6: The 10 product-surface *topics* that anchor PM^{4}Bench's MGUI task, with the 10 hand-authored HTML/Jinja sub-page templates per topic ($10 \times 10 = 100$ templates total). Each template is rendered with byte-identical layout in all 10 languages and paired with two indirect instructions, yielding $100 \times 10 \times 2 = 2000$ language-balanced (200 per language) coordinate-grounding test items.

A.5 Synthetic OCR Training Data

A representative example of the multilingual synthetic data used to train QGO-8B (Section 4) is shown in Fig. 12.

Figure 12: Example of synthesized OCR-centric GRPO training data.

A.6 QGO-8B Per-language Performance on MIQA and MGUI

Fig. 13 provides the per-language counterpart to Fig. 5 in the main text, extending the comparison from MDUR and MSOCR to MIQA and MGUI under the `vision` setting. QGO-8B preserves or improves over the Qwen3-VL-8B-Thinking baseline on every one of the 10 languages for both tasks, with the largest absolute gains on the visually hardest languages (TH, AR), confirming that OCR-centric GRPO transfers beyond OCR transcription to free-form open-ended VQA and coordinate-based MGUI grounding.

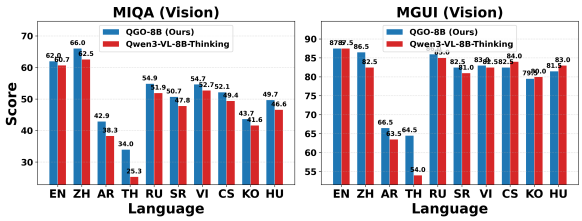


Figure 13: Per-language comparison of QGO-8B against the Qwen3-VL-8B-Thinking baseline on MIQA and MGUI (`vision`). Companion to Fig. 5 in the main text.

B Prompts

This section collects the prompts we used to query the evaluated LVLMs (one figure per task) and the prompts used by the LLM judge for MIQA. The translation prompt used to localise the corpus into the 9 non-English languages is given in Table 7.

B.1 Task Prompts

EN:
Please read the image content and the multiple-choice question, and choose the correct option after careful consideration. There is only one correct option. Please output the letter of the correct option in the last line, enclosed in angle brackets, e.g., <X>.

ZH:
请你阅读图片内容和其中的选择题，并从中选择正确选项，逐步思考后作答。正确选项只有一个，请你在最后一行输出正确选项的字母，并将其包裹在尖括号中，如: <X>

AR:
في الصحيح الخيار حرف ضع فضلك من ,فقط واحد صحيح خيار يوجد .التدريجي التفكير بعد الصحيح الخيار واختر الاختيارات، متعدد والسؤال الصورة محتوى اقرأ فضلك من: مثال زاوية، أقواس في مغلقًا الأخير، السطر <X>.

CS:
Přečtěte si obsah obrázku a výběrovou otázku a po postupném přemýšlení vyberte správnou možnost. Existuje pouze jedna správná možnost. V posledním řádku uveďte písmeno správné možnosti a obklopte ho ostrými závorkami, například: <X>.

HU:
Kérlek, olvasd el a kép tartalmát és a választós kérdést, majd fokozatos gondolkodás után válaszd ki a helyes választ. Csak egy helyes válasz létezik. Kérlek, az utolsó sorban add meg a helyes válasz betűjét, és tedd szögletes zárójelek közé, például: <X>.

KO:
이미지 내용을 읽고 객관식 질문을 살펴본 후 신중하게 선택하여 정답을 선택해주세요. 정답은 하나만 있습니다. 마지막 줄에 정답의 글자를 출력하고, 이를 꺾쇠 괄호로 감싸서 입력해 주세요, 예: <X>.

RU:
Пожалуйста, прочитайте содержание изображения и вопрос с выбором ответа, а затем выберите правильный вариант, обдумав его. Правильный вариант только один. Пожалуйста, выведите букву правильного варианта в последней строке, обрамленную угловыми скобками, например: <X>.

SR:
Молим вас да прочитате садржај слике и питање са избором одговора, а затим изаберете тачан одговор након пажљивог разматрања. Постоји само један тачан одговор. Молим вас да у последњем реду изађете са словом тачног одговора, обавијеним у угловне заграде, на пример: <X>.

TH:
กรุณาอ่านเนื้อหารูปและคำถามแบบเลือกตอบ จากนั้นเลือกคำเลือกที่ถูกต้องหลังจากคิดอย่างรอบคอบ คำเลือกที่ถูกต้องมีเพียงหนึ่งเดียว กรุณาแสดงตัวอักษรของคำเลือกที่ถูกต้องในบรรทัดสุดท้าย โดยห่อหุ้มในวงเล็บมุม เช่น: <X>

VI:
Xin vui lòng đọc nội dung hình ảnh và câu hỏi trắc nghiệm, sau đó lựa chọn phương án đúng sau khi suy nghĩ kỹ lưỡng. Chỉ có một phương án đúng. Vui lòng xuất ra chữ cái của phương án đúng ở dòng cuối cùng, và bao quanh nó trong dấu ngoặc nhọn, ví dụ: <X>.

Figure 14: Prompts used for MDUR.

EN:
Please think step by step and answer the question in the image. Place your answer between <Start of Answer> and <End of Answer>. The final response format should be as follows: <Start of Answer>This is your answer<End of Answer>.

ZH:
请你逐步思考，回答图片当中的问题，并把答案放在<Start of Answer>和<End of Answer>之间，最终回答格式应该如下: <Start of Answer>这是你的回答<End of Answer>。

AR:
يلي كما النهائية الإجابة تنسيق يكون أن يجب .<End of Answer> و <Start of Answer> بين إجابتك ضع ,الصورة في السؤال على والإجابة بخطوة خطوة التفكير يرجى <Start of Answer>إجابتك هي هذه<End of Answer>.

CS:
Postupujte krok za krokem a odpovězte na otázku na obrázku. Umístěte svou odpověď mezi <Start of Answer> a <End of Answer>. Konečný formát odpovědi by měl vypadat takto: <Start of Answer>Toto je vaše odpověď<End of Answer>.

HU:
Kérjük, gondolkozzon lépésről lépésre, és válaszoljon a képen szereplő kérdésre. Helyezze a választ a <Start of Answer> és a <End of Answer> közé. A végső válaszformátumnak így kell kinéznie: <Start of Answer>Ez a válaszod<End of Answer>.

KO:
단계적으로 생각하여 이미지 속 질문에 답하세요. 답변을 <Start of Answer>와 <End of Answer> 사이에 넣으세요. 최종 답변 형식은 다음과 같아야 합니다: <Start of Answer>이것이 당신의 답변입니다<End of Answer>.

RU:
Пожалуйста, подумайте шаг за шагом и ответьте на вопрос на изображении. Поместите свой ответ между <Start of Answer> и <End of Answer>. Окончательный формат ответа должен выглядеть следующим образом: <Start of Answer>Это ваш ответ<End of Answer>.

SR:
Молимо вас да размишљате корак по корак и одговорите на питање на слици. Ставите ваш одговор између <Start of Answer> и <End of Answer>. Коначни формат одговора треба да изгледа овако: <Start of Answer>Ово је ваш одговор<End of Answer>.

TH:
โปรดคิดทีละขั้นตอนและตอบคำถามในภาพ โดยวางคำตอบของคุณระหว่าง <Start of Answer> และ <End of Answer> รูปแบบคำตอบสุดท้ายควรเป็นดังนี้: <Start of Answer>นี่คือคำตอบของคุณ<End of Answer>.

VI:
Hãy suy nghĩ từng bước và trả lời câu hỏi trong hình. Đặt câu trả lời của bạn giữa <Start of Answer> và <End of Answer>. Định dạng câu trả lời cuối cùng phải như sau: <Start of Answer>Đây là câu trả lời của bạn<End of Answer>.

Figure 15: Prompts used for MIQA.

EN:
The image contains 20 lines of text. Please recognize and output the text in the image from top to bottom and left to right, separating the content of each line with a line break. You should output the text in the image at the end of your response. You should place the text in the image between the <start> and <end> marks.

ZH:
图像包含20行文本。请按从上到下、从左到右的顺序识别并输出图像中的文本，每行内容用换行符分隔。请在响应末尾提取图像中的文本，并将其放置在 <start> 和 <end> 标记之间。

HU:
A kép 20 sor szöveget tartalmaz. Kérjük, ismerje fel és adja ki a képen lévő szöveget felülről lefelé és balról jobbra haladva, a sorok tartalmát sortöréssel elválasztva. A kép szövegét a válasz végén, a <start> és <end> jelzők között kell elhelyezni.

RU:
Изображение содержит 20 строк текста. Распознайте и выведите текст на изображении сверху вниз и слева направо, разделяя содержимое каждой строки переносом строки. Текст изображения должен быть выведен в конце ответа между метками <start> и <end>.

AR:
إخراج يجب .أسطر بفاصل سطر كل محتوى وفصل اليمين، إلى اليسار ومن أسفل إلى أعلى من الصورة في النص على التعرف يرجى .النص من سطرًا 20 على الصورة تحتوي علامتي بين ووضعه الرد، نهاية في الصورة في النص <start> و <end>.

CS:
Obrázek obsahuje 20 řádků textu. Prosím rozpoznejte a vypište text v obrázku shora dolů a zleva doprava, oddělujte obsah každého řádku zalomením řádku. Text z obrázku by měl být uveden na konci odpovědi mezi značkami <start> a <end>.

VI:
Hình ảnh chứa 20 dòng văn bản. Vui lòng nhận dạng và xuất văn bản trong hình ảnh từ trên xuống dưới và từ trái sang phải, phân tách nội dung mỗi dòng bằng dấu ngắt dòng. Bạn nên xuất văn bản trong hình ảnh ở cuối phản hồi và đặt giữa các dấu <start> và <end>.

KO:
이미지에는 20줄의 텍스트가 포함되어 있습니다. 이미지의 텍스트를 위에서 아래로, 왼쪽에서 오른쪽으로 인식하고 각 줄의 내용을 줄 바꿈으로 구분하여 출력하십시오. 이미지의 텍스트는 응답의 끝에 <start>와 <end> 표시 사이에 배치해야 합니다.

TH:
ภาพประกอบด้วยข้อความ 20 บรรทัด โปรดจดจำและแสดงข้อความในภาพจากบนลงล่างและซ้ายไปขวา แยกเนื้อหาของแต่ละบรรทัดด้วยการขึ้นบรรทัดใหม่ ควรแสดงข้อความในภาพที่ท้ายคำตอบ โดยวางไว้ระหว่างเครื่องหมาย <start> และ <end>

SR:
Слика садржи 20 редова текста. Препознајте и испишите текст на слици одозгора надоле и с лева на десно, раздвајајући садржај сваког реда преласком у нови ред. Текст са слике треба да се налази на крају одговора између ознака <start> и <end>.

Figure 16: Prompts used for MSOCR.

B.2 LLM-as-Judge Prompt for MIQA

You are an assistant skilled at evaluating the quality of creative text.
Please act as an impartial judge and evaluate the quality of the response provided by an AI assistant to the user question displayed below. You'll need to assess the response on the following dimensions: Creativity, Richness, Visual Perception, Logical Coherence, Answer Accuracy and Image Relationship Understanding. We will provide you with a creative question and the AI model's response and a reference answer for your evaluation. As you begin your assessment, follow this process:
1. Evaluate the AI model's answers on different dimensions, pointing out its strengths or weaknesses in each dimension and assigning a score of 1 to 10 for each.
2. Finally, based on the assessments across dimensions, provide an overall score of 1 to 10 for the AI model's response.
3. Your scoring should be as stringent as possible and follow the scoring rules below:

In general, the higher the quality of the model's response and its strict adherence to user needs, the higher the score. Responses that do not meet user needs will receive lower scores.

Scoring rules:
Creativity:
Scores 1-2 when there is no innovation or uniqueness in the content.
Scores 3-4 when providing partially original content but with low creative quality.
Scores 5-6 when mostly creative but lacks significant novelty, with moderate quality.
Scores 7-8 when having novelty and high-quality content.
Scores 9-10 when highly novel and of exceptional quality compared to the reference answer.

Richness:
Scores 1-2 when lacking depth and breadth, with very limited information.
Scores 3-4 when limited in depth and breadth, with fewer explanations and examples, showing low diversity.
Scores 5-6 when limited in depth and breadth but provides basic necessary information.
Scores 7-8 when providing depth and useful additional information.
Scores 9-10 when providing exceptional depth, breadth, and high diversity compared to the reference answer.

Visual Perception:
Scores 1-2 when the description of the visual information in the image contains errors or is significantly inconsistent with the content of the image.
Scores 3-4 When the description of the visual information in the image reflects only a small amount of the image's information and contains some errors.
Scores 5-6 when the description of the visual information in the image includes the basic information of the image but contains minimal information.
Scores 7-8 when the description of the visual information in the image matches the image well and is rich in content, providing a substantial amount of information about the image.
Scores 9-10 when the description of the visual information in the image not only matches the image but also is more detailed and informative compared to the reference answer, providing more information about the image.

Logical Coherence:
Scores 1-2 when entirely incoherent, lacking any logic, and not matching the question or known information.
Scores 3-4 when somewhat coherent but with many logical errors or inconsistencies.
Scores 5-6 when mostly coherent, with few errors, but may struggle to maintain complete coherence in complex situations.
Scores 7-8 when excellent logical handling, very few errors.
Scores 9-10 when flawless logic, impeccable in handling complexity, and significantly higher logical coherence compared to the reference answer.

Answer Accuracy
Scores 1-2 when the answer is significantly inconsistent with the question or contains obvious errors.
Scores 3-4 when the answer is partially correct but contains some errors or is incomplete.
Scores 5-6 when the answer is basically correct but lacks details or is not sufficiently detailed.
Scores 7-8 when the answer is accurate and detailed, fully corresponding to the question.
Scores 9-10 when the answer is not only accurate and detailed but also provides additional useful information, exceeding expectations.

Image Relationship Understanding:
Scores 1-2 when there are significant errors or confusion in distinguishing and describing different images, unable to correctly identify and relate the content of the images.
Scores 3-4 when the description of different images reflects only minimal distinguishing information, contains some errors and confusion, and fails to clearly differentiate and relate the images.
Scores 5-6 when the description of different images includes basic distinguishing information, is able to correctly identify and relate the images in a basic manner, but the information provided is minimal and lacks detail.
Scores 7-8 when the description of different images is accurate and detailed, clearly distinguishing and relating the images, with rich content that points out the main commonalities and differences between the images.
Scores 9-10 when the description of different images is not only accurate and detailed but also provides richer information and analysis, clearly distinguishing and relating the images, more comprehensively pointing out the commonalities and differences between the images compared to the reference answer.

Overall Score:
Scores 1-2 when irrelevant to the question, factually incorrect, or generates harmful content.
Scores 3-4 when no serious errors, mostly harmless, but of low quality and does not meet requirements.
Scores 5-6 when basically meeting requirements but performing poorly in some dimensions, with moderate quality.
Scores 7-8 when performing well in all dimensions.
Scores 9-10 when fully addressing user questions and all requirements, significantly surpassing the reference answer.

Please remember, you must evaluate and explain before scoring. After your explanation for each dimension, add the score for that dimension. Finally, at the end of your response, in the format of the dictionary (including brackets), return all your scoring results, ensuring your scores are integers:
{'Dimension One': Score, 'Dimension Two': Score, ..., 'Overall Score': Score}, for example: {'Creativity': 9, 'Richness': 6, ..., 'Overall Score': 7}.

[Question]
{question}

[The Start of Reference Answer]
{ref_answer}
[The End of Reference Answer]

[The Start of Assistant's Answer]
{pred_answer}
[The End of Assistant's Answer]

Figure 17: Prompt used by the LLM judge to score MIQA responses.

B.3 Translation Prompt

Table 7 shows the two prompts used by our human-expert-in-the-loop translation pipeline (§3.5). The *Translation Prompt* asks an LLM to render the source content into the target language while preserving variables, numerals, mathematical notation, code, and image / option placeholders verbatim. For each target-language item, three independent translations are produced; the *Claude Selection Prompt* then asks a second LLM to choose the best of the three by index. The selected candidate is finally cross-checked by a native-speaker annotator before entering the benchmark.

Translation Prompt
You are a language expert specialized in {*lang*}. Please translate the following content {*lang*}. Please organize your output according to the format of the input strictly. Do not output anything other than translation. Do not translate any variables, Arabic numerals, mathematical symbols Python code or people's names. Only translate the textual description. Do not translate image marks or option indexes included by '⟨⟩', '()', for instance, '$\langle ImageHere \rangle$', '(b)'. Input: {*input*} Your translation:

Claude Selection Prompt
You are a language expert specialized in {*lang*}. Given 3 translations of an English text, you are to choose one translation which has the best quality. A good translation should accurately preserve the numbers, variables or Python code in the English text, and correctly translate other content. You can only output the index of translation, '1', '2' or '3', do not output anything else. $\langle startofEnglishtext \rangle$ {*Original English text*} $\langle EndofEnglishtext \rangle$ $\langle startoftranslation1 \rangle$ {*translation_1*} $\langle endoftranslation1 \rangle$ $\langle startoftranslation2 \rangle$ {*translation_2*} $\langle endoftranslation2 \rangle$ $\langle startoftranslation3 \rangle$ {*translation_3*} $\langle endoftranslation3 \rangle$ Your selection:

Table 7: Prompt templates for human-expert in loop translation. The Italic {*text*} in curly braces represents variables that need to be replaced.

C Language Selection

The languages we selected cover multiple language families and script systems, ensuring a certain degree of diversity. We additionally analyzed quantitative metrics that capture the graphical complexity of different writing systems, which helps explain why languages such as Thai and Arabic underperform in the `vision` setting. Despite having the highest graphical complexity among the chosen languages, Chinese still achieves relatively strong performance in `vision` settings due to the abundance of training resources for it.

As shown in Table 8 and Table 9, we report the following metrics:

- **Perimetric Complexity (PC)**:
  - Formula: $PC = P^2/(4\pi A)$.
  - P: total perimeter of the shape (in pixels).
  - A: pixel area of the foreground (the shape itself).
  - Reflects the spatial density of the shape, independently of size.
- **Number of Disconnected Components (DC)**:
  - Counts the independent, non-connected parts of the shape.
  - For example, the letter "i" has two disconnected components (the dot and the vertical stroke), while "T" has one continuous part.
  - Reflects the discontinuity of the shape, indicating how fragmented it appears visually.
- **Number of Connected Points (CP)**:
  - Counts the intersection points where multiple segments meet.
  - For instance, the letter "T" has one connected point, while "F" has two.
  - Reflects the cohesion of the shape, indicating how well its strokes are interconnected.
- **Number of Simple Features (SF)**:
  - Counts the basic elements that make up the shape, such as strokes, lines, dots or circles.
  - For example, the letter "L" consists of two simple features (a vertical and a horizontal line).
  - Closely related to stroke count; especially useful for evaluating complex writing systems such as Chinese characters or Japanese kana.

Language	ISO	GI	GC score ↑	Language Family	Script System
Chinese	zh	6097	10.014	Sino-Tibetan	Chinese Characters
Thai	th	102	1.084	Kra–Dai	Thai
Korean	ko	40	-0.840	Koreanic	Hangul / Chosŏn'gŭl
Arabic	ar	28	-1.532	Afro-Asiatic	Arabic alphabet
Hungarian	hu	46	-1.567	Uralic	Latin
Czech	cs	42	—	Indo-European	Latin
Russian	ru	33	-2.159	Indo-European	Cyrillic
Serbian	sr	30	-2.298	Indo-European	Serbian Cyrillic
Vietnamese	vi	29	—	Austroasiatic	Latin
English	en	26	-2.703	Indo-European	Latin

Table 8: Language Information Table. — indicates that GraphCom does not provide specific numerical values. However, by comparing the number of characters, language families, and other aspects of the script systems, we have identified the rankings of Czech and Vietnamese in the table.

Languages	GI	PC Mean	DC Mean	CP Mean	SF Mean	GC score↑
Chinese	6097	32.47	4.55	11.64	12.5	**10.014**
Thai	102	14.88	1.68	4.54	6.24	**1.084**
Korean	40	14.71	1.38	2.15	3.4	**-0.840**
Arabic	28	8.78	1.82	1.36	3.07	**-1.532**
Hungarian	46	9.09	1	2.85	3.7	**-1.567**
Czech	42	-	-	-	-	**-**
Russian	33	7.51	1.12	2.05	2.89	**-2.159**
Serbian	30	7.34	1.02	2.02	2.83	**-2.298**
Vietnamese	29	-	-	-	-	**-**
English	26	6.85	1.04	1.44	2.25	**-2.703**

Table 9: 4 dimension's scores which determine GraphCom score. **Means** is calculated averaging all characters in graph inventory.

- **Graph Inventory (GI)**:
  - GI is the number of characters in the script's character set.
- **GraphCom Score**:
  - A weighted combination of the derived measures above.
  - Individual scores (PC, DC, CP, SF) are first transformed into z-scores so that they can be compared across writing systems.
  - The aggregated z-score gives a single, comprehensive assessment of the graphical complexity of each writing system.

D Detailed Per-Language Results

This section reports the per-language scores that lie behind the aggregated $S_{\text{avg.}}$ and $S_{\text{cv.}}$ values shown in Table 3. Tables 10, 11, 12 and 13 give the breakdown for MSOCR, MDUR, MIQA and MGUI, respectively.

D.1 Sanity-check task: $\text{MGUI}_{\text{text}}$

Because GPT-5 and GPT-5-mini both score unusually low on the main MGUI grounding task (Table 13), in contrast to their strong performance on every other benchmark, we ran a controlled companion task to verify that this is a genuine model property rather than an experimental artifact. We call this companion task $\text{MGUI}_{\text{text}}$.

Setup. The $\text{MGUI}_{\text{text}}$ task uses the exact same screenshots and the exact same natural-language instructions as the main MGUI grounding task. The only difference is the answer format: instead of emitting a click coordinate (x, y), the model is asked to output the visible *text* of the target UI element (*e.g.*, "One way", "Search flights"). A prediction is judged correct if its normalized text matches the target element's visible text. This isolates the recognition + understanding capability ("which element should be clicked") from the spatial grounding capability ("where exactly is that element").

Results. Table 14 reports the per-language results. GPT-5 reaches 94.04 averaged over 10 languages, second only to Gemini-3.1-Pro-Preview (97.15) and clearly above all open-source models in our suite. GPT-5-mini reaches 91.91. Across all 10 languages the two GPT-5 variants stay above 85 on every language, including the visually hard AR (88.00 / 87.94) and TH (93.50 / 85.00), demonstrating that their recognition pipeline copes with non-English on-screen text just fine.

Implication. Combined with the median click-to-bbox-edge distance of only ~ 33 px reported in Section 6.2, this confirms that GPT-5 cleanly identifies the correct UI element and emits well-formed coordinates aimed at roughly the right neighborhood; the failure is restricted to fine-grained pixel-level grounding. We surface this anomaly to forestall the natural reviewer concern that our pipeline mishandled GPT-5 outputs, and leave a deeper investigation of why a strong general VLM exhibits such a narrow precision-grounding gap to future work. The gap is also model-specific rather than vendor-specific: other commercial systems (e.g. Gemini-3.1-Pro-Preview) do not exhibit this asymmetry, suggesting it reflects a particular decoding-time bias of the GPT-5 family rather than a fundamental limitation of the screenshot-grounding paradigm.

Model	EN	ZH	HU	RU	SR	CS	AR	VI	TH	KO
Gemini-3.1-Pro-Preview	35.47	29.40	34.88	35.26	34.08	34.14	28.10	33.36	10.14	28.96
GPT-5	33.02	11.08	24.28	24.30	12.90	25.96	3.26	15.06	2.12	11.08
GPT-5-mini	33.30	10.70	16.30	23.74	7.92	16.96	1.74	13.32	1.90	9.90
Llama-4-Maverick	30.38	7.08	5.66	12.34	3.58	6.48	3.62	6.46	2.74	4.40
GLM-4.5v	31.88	24.68	5.10	7.82	0.12	6.24	0	0.92	0	0.26
Doubao-Seed-1.6	29.42	23.50	9.78	3.66	0.70	6.30	2.78	6.28	0	3.34
Qwen3-VL-235B-A22B-Thinking	32.29	30.39	12.69	21.96	11.55	18.36	3.98	13.86	0.52	14.29
Qwen3-VL-235B-A22B-Instruct	32.96	31.44	16.12	23.04	16.84	23.38	0.04	21.60	1.02	16.92
Qwen3-VL-32B-Thinking	28.76	26.09	8.81	14.95	7.20	9.43	0.72	8.12	0	9.64
Qwen3-VL-8B-Thinking	4.82	4.76	0.22	1.94	0.18	0.38	0.16	1.32	0	1.48
QGO-8B (ours)	20.86	16.58	5.32	12.02	5.16	8.10	1.36	6.88	0.10	5.36

Table 10: MSOCR detailed results.

Model	Traditional setting									
	EN	ZH	HU	RU	SR	CS	AR	VI	TH	KO
Gemini-3.1-Pro-Preview	81.49	80.87	79.83	80.92	78.96	81.04	81.10	81.16	76.39	80.69
GPT-5	74.28	72.72	73.82	73.01	72.43	73.64	71.79	72.49	72.72	72.02
GPT-5-mini	72.02	69.65	69.65	69.65	69.88	69.48	68.90	70.40	68.27	69.88
Llama-4-Maverick	60.87	57.46	58.15	58.21	56.76	58.55	55.20	57.57	56.01	56.18
GLM-4.5v	59.69	57.94	56.22	57.79	54.24	54.53	48.78	53.90	52.24	54.94
Doubao-Seed-1.6	65.41	63.16	62.41	57.72	53.79	60.61	58.07	59.11	59.05	56.54
Qwen3-VL-235B-A22B-Thinking	68.14	64.92	66.33	62.54	62.70	65.58	65.17	65.64	64.54	64.83
Qwen3-VL-235B-A22B-Instruct	65.17	64.57	61.03	62.44	60.50	60.31	61.13	57.97	58.84	58.71
Qwen3-VL-32B-Thinking	66.10	64.40	64.27	64.54	61.67	63.13	63.90	64.51	64.42	63.49
Qwen3-VL-8B-Thinking	40.12	38.27	40.35	33.06	38.67	41.62	36.18	40.58	38.09	38.61
QGO-8B (ours)	51.04	42.89	47.86	37.80	46.24	51.04	47.57	48.84	46.94	47.92
Model	Vision setting									
	EN	ZH	HU	RU	SR	CS	AR	VI	TH	KO
Gemini-3.1-Pro-Preview	80.75	79.71	79.29	78.79	77.11	79.71	77.05	79.54	73.90	78.66
GPT-5	72.89	68.21	69.71	69.88	70.40	69.36	63.24	71.27	66.01	69.31
GPT-5-mini	72.02	68.61	68.79	70.12	69.19	69.77	63.91	69.88	69.25	69.77
Llama-4-Maverick	55.38	48.90	49.13	52.08	46.82	47.46	40.92	50.29	45.38	46.59
GLM-4.5v	61.75	58.44	52.87	54.58	46.43	55.95	22.17	50.44	25.89	41.06
Doubao-Seed-1.6	65.26	63.47	60.87	55.61	50.23	62.20	41.45	57.98	34.10	52.89
Qwen3-VL-235B-A22B-Thinking	66.18	63.25	64.04	63.21	59.20	64.62	56.09	62.79	53.52	62.35
Qwen3-VL-235B-A22B-Instruct	64.45	62.09	55.55	59.48	57.57	59.13	50.46	53.06	49.42	53.70
Qwen3-VL-32B-Thinking	66.94	64.44	64.57	62.97	60.43	64.57	52.13	63.35	47.43	57.81
Qwen3-VL-8B-Thinking	44.22	41.16	38.96	25.55	38.15	38.73	24.34	40.23	21.45	36.01
QGO-8B (ours)	49.42	46.07	46.01	23.24	43.53	46.36	33.24	45.49	34.10	40.98

Table 11: MDUR detailed results.

Model	Traditional setting									
	EN	ZH	HU	RU	SR	CS	AR	VI	TH	KO
Gemini-3.1-Pro-Preview	52.37	64.30	65.16	62.40	60.79	58.11	65.68	64.52	46.90	62.45
GPT-5	50.44	61.65	53.51	52.29	52.89	57.12	54.36	55.21	52.64	58.05
GPT-5-mini	64.01	65.44	61.09	61.89	61.22	64.75	63.46	60.03	61.44	66.51
Llama-4-Maverick	49.54	47.14	41.58	51.83	46.23	45.85	43.70	49.52	41.92	43.42
GLM-4.5v	53.59	56.31	49.16	51.77	51.78	51.68	45.83	50.01	46.08	50.35
Doubao-Seed-1.6	62.32	60.36	51.99	56.05	52.74	54.13	51.37	54.65	51.08	55.91
Qwen3-VL-235B-A22B-Thinking	64.47	67.35	59.67	61.91	60.26	62.64	61.16	59.91	58.41	59.78
Qwen3-VL-235B-A22B-Instruct	72.54	73.33	63.25	64.19	62.44	63.33	55.93	42.81	59.92	61.88
Qwen3-VL-32B-Thinking	60.77	60.92	50.77	50.95	53.93	56.77	54.89	52.11	53.50	55.21
Qwen3-VL-8B-Thinking	59.46	61.13	47.64	53.31	50.28	52.75	52.39	54.95	53.17	51.23
QGO-8B (ours)	59.14	62.10	50.34	54.25	52.05	53.64	53.47	57.70	55.77	53.91
Model	Vision setting									
	EN	ZH	HU	RU	SR	CS	AR	VI	TH	KO
Gemini-3.1-Pro-Preview	50.79	53.77	57.75	55.89	52.29	51.00	52.95	56.63	43.12	56.96
GPT-5	52.59	60.30	54.76	54.09	55.31	57.20	52.69	57.51	55.25	59.78
GPT-5-mini	65.14	63.13	58.89	60.08	58.09	65.05	60.38	58.24	61.32	65.15
Llama-4-Maverick	48.54	43.87	43.65	49.84	42.94	42.39	39.66	48.61	39.90	41.09
GLM-4.5v	54.85	57.21	47.93	52.03	48.90	51.79	22.11	47.51	21.46	29.88
Doubao-Seed-1.6	59.79	55.98	48.68	53.13	47.06	52.53	36.81	51.72	19.47	45.23
Qwen3-VL-235B-A22B-Thinking	62.06	66.57	50.52	57.97	53.87	56.17	47.45	52.87	40.93	55.15
Qwen3-VL-235B-A22B-Instruct	71.98	72.83	61.25	66.06	61.47	61.54	43.72	46.39	48.46	61.54
Qwen3-VL-32B-Thinking	57.22	57.17	49.68	51.20	51.84	53.37	46.95	50.49	27.71	49.00
Qwen3-VL-8B-Thinking	60.68	62.55	46.61	51.92	47.81	49.38	38.29	52.74	25.31	41.63
QGO-8B (ours)	61.97	66.03	49.68	54.86	50.70	52.14	42.88	54.68	33.98	43.65

Table 12: MIQA detailed results.

Model	EN	ZH	HU	RU	SR	CS	AR	VI	TH	KO
Gemini-3.1-Pro-Preview	90.00	86.50	87.00	82.00	89.00	84.50	89.00	88.00	88.50	86.50
GPT-5	17.50	6.53	8.00	10.00	10.50	11.06	10.50	12.50	10.00	9.05
GPT-5-mini	17.09	11.06	16.50	18.09	16.08	17.26	16.58	13.50	10.05	12.06
Llama-4-Maverick	67.50	30.00	62.00	39.50	61.00	64.50	45.00	63.00	59.00	56.50
GLM-4.5v	87.50	84.50	78.50	82.00	78.50	85.00	46.50	81.50	50.00	70.00
Doubao-Seed-1.6	88.00	88.50	85.00	87.00	79.50	87.00	68.50	87.00	61.00	83.00
Qwen3-VL-235B-A22B-Thinking	90.00	89.50	90.50	89.50	90.00	91.50	84.50	90.00	85.50	90.50
Qwen3-VL-235B-A22B-Instruct	88.50	90.00	81.50	83.50	80.00	87.50	70.50	82.50	71.50	85.50
Qwen3-VL-32B-Thinking	88.50	89.50	86.50	88.50	88.00	88.50	80.50	87.50	76.50	89.50
Qwen3-VL-8B-Thinking	87.50	82.50	83.00	85.00	81.00	84.00	63.50	82.50	54.00	80.00
QGO-8B (ours)	87.50	86.50	81.50	86.00	82.50	82.50	66.50	83.00	64.50	79.50

Table 13: MGUI detailed per-language results (`vision` setting). Each cell reports the click-grounding accuracy: a prediction (x, y) is correct iff it falls inside the target element's bbox.

Model	EN	ZH	HU	RU	SR	CS	AR	VI	TH	KO
Gemini-3.1-Pro-Preview	98.00	96.00	97.00	98.00	97.50	97.00	94.50	97.00	98.50	98.00
GPT-5	94.50	94.97	96.00	95.48	93.00	95.94	88.00	95.00	93.50	93.97
GPT-5-mini	93.97	93.47	94.00	93.94	92.50	93.85	87.94	94.00	85.00	90.45
Llama-4-Maverick	88.50	89.50	84.00	88.00	78.50	83.50	64.00	72.50	66.00	62.50
GLM-4.5v	92.00	92.50	75.00	89.50	74.00	91.50	27.00	68.50	35.00	67.50
Doubao-Seed-1.6	94.00	93.00	77.50	85.50	56.50	86.50	60.00	86.00	41.50	77.50
Qwen3-VL-235B-A22B-Thinking	94.50	96.00	92.00	94.50	94.00	94.00	79.50	89.50	74.00	90.00
Qwen3-VL-235B-A22B-Instruct	93.00	93.50	85.50	93.00	90.50	92.50	79.00	87.00	68.00	87.50
Qwen3-VL-32B-Thinking	91.50	93.50	90.00	93.50	87.50	93.00	69.00	92.00	57.50	82.00
Qwen3-VL-8B-Thinking	94.00	93.50	84.50	89.50	84.00	88.50	55.00	86.00	38.50	68.50

Table 14: $\text{MGUI}_{\text{text}}$ per-language results (`vision` setting). The companion task probes *element identification* on the same screenshots and instructions used by the main MGUI grounding task: instead of emitting click coordinates, the model is asked to output the visible text of the target UI element. A prediction is judged correct if its normalized text matches the target element's visible text. This isolates "what to click" (recognition + understanding) from "where to click" (spatial grounding) and serves as a sanity-check for the unusually low MGUI grounding scores of GPT-5 / GPT-5-mini reported in Table 13.

Model	Direct Judge	Translate to EN before Judge	Δ
GPT-5-mini	61.15	58.74	-2.41
Qwen3-VL-235B-A22B-Instruct	58.14	56.52	-1.62
GPT-5	56.32	54.35	-1.97
Qwen3-VL-235B-A22B-Thinking	53.50	52.00	-1.50
Gemini-3.1-Pro-Preview	53.37	51.59	-1.79
QGO-8B	49.85	46.82	-3.03
Qwen3-VL-32B-Thinking	48.60	47.46	-1.14
Qwen3-VL-8B-Thinking	46.25	43.70	-2.55
Doubao-Seed-1.6	45.62	44.69	-0.93
Llama-4-Maverick	43.55	42.99	-0.56
GLM-4.5v	42.09	41.20	-0.89

Table 15: The scores represent the average of 9 non-English language evaluations, with a maximum possible score of 100.

E Reliability of LLM-as-Judge on MIQA

E.1 Translate-to-English consistency

While our benchmark is designed to evaluate cross-lingual capability disparities in LVLMs, one might worry that using an LLM-as-judge introduces bias from the inherently uneven cross-lingual performance of the judge itself. To verify that this is not the case, we implemented a **translate-to-EN-before-judge** strategy: for the MIQA vision task, we translated LVLM responses into English before evaluation. The results for both judging strategies are summarized in Table 15.

The translate-to-EN strategy produces results **nearly identical to those obtained by judging the original (non-English) responses directly**. This indicates that our structured prompt design, which incorporates multiple evaluation dimensions and detailed scoring criteria, effectively mitigates linguistic bias in the LLM-as-judge assessment process. We therefore retain direct LLM-as-judge in all subsequent experiments.

E.2 Cross-validation with Claude-Opus-4.7

To address concerns about single-judge reliability, we re-evaluated MIQA-vision with an independent second judge: **Claude-Opus-4.7**, replacing our default DeepSeek-v4-flash judge while keeping every other element of the protocol (the same six scoring dimensions, the same human-written references, the same prompt template) fixed. The second judge was applied to the three models for which full re-evaluations were available: GPT-5 (commercial leader), Qwen3-VL-8B-Thinking (our open-source baseline), and QGO-8B (our trained model). Table 16 reports the per-language comparison.

Across all 10 languages and 3 models, the two judges agree closely: mean $|\Delta|$ is 1.24, 2.27, and 2.01 points respectively. Crucially, the relative ordering of models is preserved on every language and the cross-lingual pattern (large drop on AR/TH, strong on EN/ZH) reproduces under the alternate judge. This confirms that the cross-lingual disparities we report are properties of the evaluated LVLMs rather than artifacts of our judge choice.

Model	Judge	AR	CS	EN	HU	KO	RU	SR	TH	VI	ZH	Avg.
GPT-5	DeepSeek-v4-flash	52.69	57.20	52.59	54.76	59.78	54.09	55.31	55.25	57.51	60.30	55.95
	Claude-Opus-4.7	54.20	60.30	54.80	55.80	61.10	54.30	55.70	56.70	58.20	60.80	57.19
	Δ	+1.51	+3.10	+2.21	+1.04	+1.32	+0.21	+0.39	+1.45	+0.69	+0.50	+1.24
Qwen3-VL-8B-Thinking	DeepSeek-v4-flash	38.29	49.38	60.68	46.61	41.63	51.92	47.81	25.31	52.74	62.55	47.69
	Claude-Opus-4.7	36.00	47.20	59.20	42.40	38.40	51.10	43.50	25.00	51.10	60.30	45.42
	Δ	−2.29	−2.18	−1.48	−4.21	−3.23	−0.82	−4.31	−0.31	−1.64	−2.25	−2.27
QGO-8B (ours)	DeepSeek-v4-flash	42.88	52.14	61.97	49.68	43.65	54.86	50.70	33.98	54.68	66.03	51.06
	Claude-Opus-4.7	44.80	53.90	64.20	51.20	46.90	57.60	51.80	37.60	57.90	64.80	53.07
	Δ	+1.92	+1.76	+2.23	+1.52	+3.25	+2.74	+1.10	+3.62	+3.22	−1.23	+2.01

Table 16: MIQA-vision per-language scores under our default DeepSeek-v4-flash judge versus a Claude-Opus-4.7 second judge. Both judges score on a 0–100 scale (six dimensions averaged, then $\times 10$). Δ = Claude − DeepSeek. Mean $|\Delta|$ across languages is 1.24, 2.27, 2.01 for the three models respectively, showing no systematic per-language bias that would alter our cross-lingual conclusions.

F External-Benchmark Per-Language Results

§6.3 reports aggregated results of QGO-8B on three external multilingual suites; this appendix lists the per-language breakdowns.

CC-OCR (Liu et al., 2024c) is a 10-language full-text OCR set (150 samples per language). We report Normalized Edit Distance (NED) and character-level F1. Table 17 shows that QGO-8B improves on every one of the 10 languages, with a +5.7 trained-language average and a +4.0 average on six languages (FR, DE, IT, JA, PT, ES) that were *never* seen during RL, direct evidence of cross-lingual transfer.

MTVQA (Tang et al., 2024a) is a multilingual text-centric VQA benchmark. Five of its nine languages overlap with our training set. Table 18 reports per-language accuracy. Trained-language scores improve modestly (+0.5 on average); the smaller margin is consistent with MTVQA's reasoning-heavy formulation, where OCR is necessary but far from sufficient. Unseen-language averages move by -0.4, well within run-to-run variance.

OCRBench (Liu et al., 2024b) is a single-score (out of 1000) English/Chinese OCR benchmark whose 1000 samples are released only with a 10-sub-task taxonomy (no per-language split is shipped with the dataset). As Table 19 shows, QGO-8B matches the baseline overall (791 vs 789) and is within ± 6 points on every sub-task, confirming no regression of base OCR capability.

Lang	Baseline	QGO-8B	Δ
Trained languages (overlap with our 10)			
AR	36.8	47.8	**+11.0**
KR	57.0	61.9	+4.9
RU	73.8	77.5	+3.7
VI	53.1	56.3	+3.2
Unseen languages			
FR	77.0	80.0	+3.0
DE	69.5	74.6	+5.1
IT	67.7	73.2	+5.4
JA	46.7	51.6	+4.9
PT	76.0	77.9	+1.9
ES	67.6	71.5	+3.9
Trained avg	55.2	60.9	+5.7
Unseen avg	67.4	71.5	+4.0
Overall	62.5	67.2	**+4.7**
Char-F1	80.2	87.0	+6.8

Table 17: Per-language CC-OCR (Liu et al., 2024c) results. NED (%); higher is better. All 10 languages improve; the largest gain is on Arabic (+11.0). Six unseen languages (FR/DE/IT/JA/PT/ES) gain +4.0 on average, evidencing cross-lingual transfer beyond the training set.

Lang	Baseline	QGO-8B	Δ
Trained languages			
AR	20.1	20.5	+0.4
KR	42.3	42.5	+0.2
TH	28.6	29.4	+0.9
VI	44.8	46.2	+1.4
RU	17.6	17.2	−0.4
Unseen languages			
DE	38.6	39.2	+0.6
FR	48.3	47.9	−0.5
IT	43.8	43.2	−0.6
JA	26.1	24.8	−1.3
Trained avg	30.7	31.1	+0.5
Unseen avg	39.2	38.8	−0.4
Overall	34.5	34.5	+0.1

Table 18: Per-language MTVQA (Tang et al., 2024a) results. Accuracy (%); higher is better. The trained-language gain (+0.5) is positive but modest, reflecting that MTVQA's text-VQA task involves a substantial reasoning component beyond raw OCR; unseen-language scores are essentially preserved.

Sub-task	Baseline	QGO-8B	Δ
Regular Text Recog.	98.0	98.0	0.0
Irregular Text Recog.	94.0	96.0	+2.0
Artistic Text Recog.	94.0	92.0	−2.0
Handwriting Recog.	72.0	66.0	−6.0
Digit String Recog.	70.0	72.0	+2.0
Non-Semantic Text Recog.	92.0	94.0	+2.0
Scene Text-centric VQA	85.0	85.5	+0.5
Doc-oriented VQA	92.5	92.0	−0.5
Key Information Extraction	87.0	87.0	0.0
HMER†	0.0	0.0	0.0
Total (out of 1000)	789	791	+2

Table 19: Per-sub-task OCRBench (Liu et al., 2024b) results. OCRBench is an English/Chinese OCR benchmark whose 1000 samples are released only with a fixed 10-task taxonomy (no per-language split is provided by the dataset), so we report per-sub-task accuracy as the finest available breakdown. QGO-8B matches the baseline overall, confirming no degradation of base OCR ability despite our RL training using none of OCRBench's data distribution. †HMER = Handwritten Mathematical Expression Recognition; both baseline and QGO-8B score 0 because the task requires LaTeX-formatted math output that neither model is trained to emit.

G Per-language PCC Analysis

To probe whether OCR capability is linked to downstream task performance *across languages*, we compute Pearson Correlation Coefficients (PCCs) between three pairs of per-language score vectors (length 10, one entry per language) for each of the 10 evaluated LVLMs: *MSOCR vs. MDUR$_{vision}$*, *MSOCR vs. MIQA$_{vision}$*, and *MSOCR vs. MGUI*. Given two such per-language score vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{10}$, PCC is defined as

$$\mathrm{PCC}(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^{10}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{10}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{10}(y_i - \bar{y})^2}}, \tag{6}$$

where $\bar{x}$ and $\bar{y}$ denote the language-wise means; $|\mathrm{PCC}| > 0.5$ is conventionally read as a strong linear correlation. The first pair probes the link between multi-scale OCR and reasoning VQA on visually rendered text; the second probes the link to open-ended generation; the third probes the link to coordinate-based grounding on multilingual screenshots. The full per-model heatmap visualizing these correlations is shown in the main text (Fig. 4b): MSOCR vs. MDUR-vision is consistently strong across all 10 evaluated models; MIQA-vision and MGUI also exhibit predominantly strong positive correlations, with a few model-specific exceptions (e.g. GPT-5 on MIQA, Gemini-3.1-Pro on MGUI). The pattern supports our claim that OCR is a key cross-lingual factor.

H Training and Inference Configuration

This appendix details the inference setup we apply to all 10 evaluated LVLMs and the training recipe used to obtain QGO-8B from the Qwen3-VL-8B-Thinking base model.

Inference setup. For all 10 evaluated models on every PM^{4}Bench task, we use **greedy decoding** (temperature 0) and apply each model's official default chat template. This deterministic configuration eliminates run-to-run sampling variance, so any cross-language or cross-model score gap is attributable to the model rather than to decoding noise. It also matches how production multimodal agents are typically deployed for tool-use and grounding workloads, where reproducibility is required. Per-task user prompts (one figure per task) and the LLM-as-judge prompt for MIQA are listed in appendix B.

Training framework and optimizer. The OCR-centric GRPO post-training of QGO-8B is implemented on top of VeRL (Sheng et al., 2024), a flexible RLHF framework that supports group-relative advantage computation. We optimize the base Qwen3-VL-8B-Thinking with AdamW at learning rate 1×10^{-6} and a global batch size of 256

prompts. Each optimization step generates 8 rollouts per prompt to provide the in-group advantage baseline required by GRPO; rewards are computed per rollout and the group-relative advantage is used to update the policy.

Convergence and compute cost. Empirically, the training reward stabilizes well within the first 200 optimization steps (see Fig. 18), at which point we stop training. The full post-training run completes in approximately 9 hours on 8 NVIDIA H200 GPUs. Crucially, our reward is computed entirely from synthetic OCR data without any human annotation or task-specific VQA supervision, and the base model is small (8B parameters), which makes our recipe a low-friction add-on that can be re-run after each base-model release.

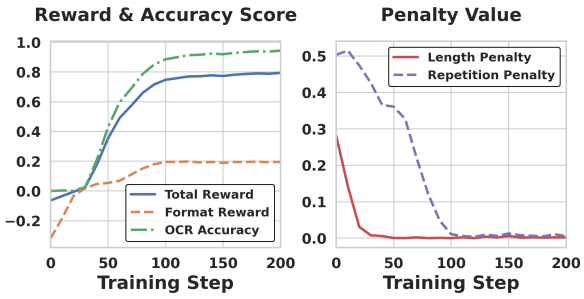


Figure 18: Training dynamics of OCR-centric GRPO. **Left:** The evolution of total reward, R_{fmt}, and R_{acc} over training steps on validation split, calculated every 10 steps. **Right:** The reduction in P_{len} and P_{rep}.

Reward hyperparameters. During training we empirically set $L_{\min} = 1000$, $L_{\max} = 10000$, $\lambda = 1200$, $\tau = 0.6$, $\alpha_1 = 0.2$, $\alpha_2 = 0.8$, $\alpha_3 = 0.4$, $\omega_1 = 0.8$, and $\omega_2 = 0.2$. The format reward is intentionally weighted lower than the accuracy reward ($\omega_1/\omega_2 = 4$): satisfying the format and length constraint should yield only a sustaining signal, while the OCR accuracy term drives the optimization trajectory. This weighting preserves the model's underlying chain-of-thought reasoning behavior, which is critical for the non-OCR transfer results reported in the main text.