VLURes: Benchmarking Long-Text Grounding and Cross-Lingual Robustness in Vision Language Models

Jesse Atuhurra[♠] **Iqra Ali**[‡] **Tomoya Iwakura**[§] **Hidetaka Kamigaito**[♠] **Tatsuya Hiraoka**[♠Ø†]

[♠] Nara Institute of Science and Technology (NAIST), Japan

[‡] Queen Mary University of London (QMUL), United Kingdom

[§] Meiji University, Japan [Ø] RIKEN, Japan

[†] Mohamed bin Zayed University of Artificial Intelligence (MBZUAI), United Arab Emirates

atuhurra.jesse.ag2@naist.ac.jp iqra.ali@qmul.ac.uk iwakura@meiji.ac.jp

kamigaito.h@naist.ac.jp tatsuya.hiraoka@mbzuai.ac.ae

Abstract

We introduce ***VLURes***, a multilingual benchmark for evaluating Vision-Language Models (VLMs) under *long-text grounding*: selecting and reasoning over the image-relevant subset of article-length text that contains distractors and ungrounded claims. *VLURes* contains **4,000** web-curated *image+long-text* pairs across **English (En), Japanese (Ja), Swahili (Sw), and Urdu (Ur)** and **10** topical categories, and defines **eight** tasks spanning image-only perception (OR, SU, RU, SS, IC) and image+text grounding (ITM, *Unrelatedness*, VQA). To construct web-realistic pairs, we apply language-adapted CLIP alignment to select representative images and filter weakly grounded pages. Across **10** proprietary and open VLMs evaluated under zero-shot and one-shot prompting, with and without rationales, the best model (GPT-4o) reaches **90.8%** overall accuracy but remains **6.7** points below human performance (**97.5%**) on Object Recognition, and cross-lingual sensitivity persists, while open models are substantially weaker and often lack reliable multilingual VL support. *VLURes* provides a practical testbed for long-text grounding and multilingual robustness in web-realistic agent settings.

1 Introduction

Intelligent agents, including robots and multimodal assistants, must interpret complex scenes, identify objects, and reason about spatial and functional relationships to act safely and effectively (Driess et al., 2023; Zitkovich et al., 2023). In real deployments, these agents often perceive an image alongside *long-form text* such as a news story, a travel description, or a product review. In this regime, the central challenge is **long-text grounding**: selecting the small subset of article-length text that is actually supported by the image, ignoring distractors and tangents, and responding without importing ungrounded claims.

This work started when Jesse and Iqra were interns at the Artificial Intelligence Laboratory of Fujitsu Ltd. in Japan.

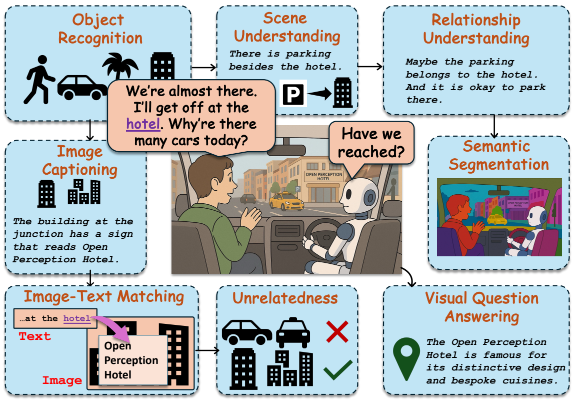


Figure 1: ***VLURes*** consists of **eight tasks** above, which enable the intelligent agent to understand its surroundings. We evaluate VLMs' performance on *VLURes*, which offers *image-text* pairs in **En, Ja, Sw, Ur**. *We embed all images in article-length prose (Figure 2).* **Takeaway:** The eight-task suite provides a realistic testbed of the VLM's ability to ground *article-length text* introduced in *VLURes* to the image, and to discard unrelated content.

What breaks today. Most widely used evaluations under-measure this failure mode because they rely on short captions, templated questions, or carefully curated image-caption pairs. Popular resources such as MSCOCO (Lin et al., 2015), Visual Genome (Krishna et al., 2017a), and many VQA-style datasets assume compact text and are dominated by English. As a result, multilingual progress can look stronger than it is in web-realistic settings, particularly for lower-resource languages where both textual coverage and vision-language (VL) alignment remain uneven (Pfeiffer et al., 2022; Geigle et al., 2024; Chen et al., 2023b; Liu et al., 2021). When presented with long articles, many VLMs over-trust the text, paraphrase irrelevant details, and hallucinate visual support for off-topic paragraphs. These errors often intensify under language shift, where models may misread instructions or revert to English-centric priors (Pfeiffer et al., 2022; Geigle et al., 2024). For intelligent agents, such failures are not cosmetic: treating

Findings of the Association for Computational Linguistics: ACL 2026, pages 27426–27481
July 2-7, 2026 ©2026 Association for Computational Linguistics

Task	What it measures (inputs → outputs)
OR	Image → identify key objects and attributes.
SU	Image → describe global scene, actions, and salient regions.
RU	Image → infer relations among entities (spatial, functional).
SS	Image → segment or describe regions and labels.
IC	Image → produce a faithful caption.
ITM	Image + long text → decide if text matches image content.
U	Image + long text → detect and discard unrelated text content.
VQA	Image + long text → answer grounded questions with evidence.

Table 1: *VLURes* task suite. **Takeaway:** The three image+text tasks explicitly probe *long-text grounding*, especially Unrelatedness.

ungrounded text as visually verified can drive unsafe or unhelpful downstream actions.

What we need to measure. Benchmarks that reflect agent-facing conditions should: (i) use *article-length* text with naturally occurring distractors, (ii) separate image-only competence from image+text grounding, and (iii) cover multiple languages with realistic web provenance. Recent multilingual evaluations emphasize that language coverage, data provenance, and alignment assumptions can materially change conclusions (Geigle et al., 2024; Chen et al., 2023b; Pfeiffer et al., 2022). However, few benchmarks explicitly stress-test *long-text grounding* across both high- and low-resource languages.

VLURes: a benchmark for long-text grounding under language shift. We introduce **VLURes**, a multilingual benchmark designed to stress-test long-text grounding and cross-lingual robustness. VLURes contains **4,000** web-curated examples across **English (En), Japanese (Ja), Swahili (Sw), and Urdu (Ur)**, each consisting of a representative image paired with *article-length* text. To probe complementary capabilities relevant to intelligent agents, VLURes defines **eight tasks** (Table 1). Five tasks isolate *image-only* competence: Object Recognition (OR), Scene Understanding (SU), Relation Understanding (RU), Semantic Segmentation (SS), and Image Captioning (IC). Three tasks require *image+text* grounding: Image–Text Matching (ITM), Visual Question Answering (VQA), and **Unrelatedness (U)**. Unrelatedness is a direct operationalization of long-text grounding: given an image and a long article, the model must *detect and discard* text that is not supported by the image, reflecting a common web condition where only a small portion of the document is visually grounded.

Dataset trustworthiness and intrinsic difficulty. Web pages often contain multiple images, icons, and off-topic media. To pair articles with representative images while reducing spurious matches, we use a **language-adapted CLIP alignment** step that selects the most relevant image and filters weakly grounded pages. We then characterize intrinsic sample difficulty using (i) text length, (ii) the adapted alignment score, and (iii) an ambiguity margin between the best and runner-up images. These signals help quantify how web-realistic noise and language shift interact with model behavior, complementing prior findings on multilingual generalization gaps (Pfeiffer et al., 2022; Geigle et al., 2024; Chen et al., 2023b).

Summary of findings and contributions. Across **10** VLMs and four prompting regimes (zero-shot vs. one-shot, with vs. without rationales), one-shot prompting and rationales often improve absolute accuracy, but do not reliably close cross-lingual gaps. Long-text grounding tasks are consistently harder than image-only tasks, especially under language shift. Open models face a dual barrier: weaker performance and unreliable multilingual VL support, even when their base large language models (LLMs) are multilingual (Geigle et al., 2024; Chen et al., 2023b). These trends align with evidence that prompting can change apparent reasoning without guaranteeing grounded behavior (Wei et al., 2022b; Turpin et al., 2023). Overall, VLURes contributes: (i) a multilingual image+*long-text* benchmark with realistic web provenance, (ii) an eight-task suite that cleanly separates image-only competence from long-text grounding (including **Unrelatedness**), and (iii) a standardized evaluation of proprietary and open VLMs under controlled prompting regimes to quantify multilingual robustness.

2 Related Work

Multimodal benchmarks beyond captions. A large fraction of VLM evaluation has historically focused on captioning and VQA-style supervision with short inputs, including MSCOCO (Lin et al., 2015), Flickr30k (Plummer et al., 2016), VQAv2 (Goyal et al., 2017a), GQA (Hudson and Manning, 2019), and TextVQA (Singh et al., 2019). More recent benchmarks (Table 2) aim to stress broader capabilities via instruction-style prompts or multi-domain reasoning, such as MME (Fu et al., 2023), MMMU (Yue et al., 2023), MMBench (Liu et al., 2023d), SEED-Bench (Li et al., 2023a), and MM-Vet (Yu et al., 2023). These resources have

Dataset	Task	#Tasks	Multilingual	Language	#Languages	Rationales	#Images	#Questions	Article-level Prose
VQAv2 (Goyal et al., 2017a)	VQA	1	✘	En	1	✘	265K	1.1M	✘
OK-VQA (Marino et al., 2019)	VQA	1	✘	En	1	✘	14K	14K	✘
OCR-VQA (Mishra et al., 2019)	VQA	1	✘	En	1	✘	207K	1M	✘
GQA (Hudson and Manning, 2019)	VQA	1	✘	En	1	✘	113K	22M	✘
Visual Genome (Krishna et al., 2017b)	VQA	1	✘	En	1	✘	108K	1.7M	✘
VizWizQA (Gurari et al., 2018a)	VQA	1	✘	En	1	✘	*	31.1K	✘
TextVQA (Singh et al., 2019)	VQA	1	✘	En	1	✘	28K	45.3K	✘
LAION 5B (Schuhmann et al., 2022a)	IC	1	✔	En, Zh, …	many	✘	5.85B	*	✘
MSCOCO (Lin et al., 2015)	IC	1	✘	En	1	✘	328K	*	✘
Flickr30k (Plummer et al., 2016)	IC	1	✘	En	1	✘	30K	*	✘
Crossmodal-3600 (Thapliyal et al., 2022)	IC	1	✔	En, Ja, Sw, …	36	✘	3.6K	*	✘
RefClef (Kazemzadeh et al., 2014)	REG	1	✘	En	1	✘	19.9K	*	✘
RefCOCO (Kazemzadeh et al., 2014)	REG	1	✘	En	1	✘	19.9K	*	✘
RefCOCO+ (Kazemzadeh et al., 2014)	REG	1	✘	En	1	✘	19.9K	*	✘
RefCOCOg (Mao et al., 2016)	REG	1	✘	En	1	✘	25.7K	*	✘
MMMU (Yue et al., 2023)	VQA	–	✘	En	1	✘	11K	11.5K	✘
MME+ (Fu et al., 2023)	OR, OCR	7	✘	En	1	✘	1K	2K	✘
MMBench (Liu et al., 2023d)	OCR	20	✔	En, Zh	2	✘	2.9K	2.9K	✘
SEED-Bench (Li et al., 2023a)	OCR	12	✘	En	1	✘	19K	19K	✘
MathVista (Lu et al., 2024)	Math	12	✘	En	1	✘	6.1K	6.1K	✘
MM-Vet (Yu et al., 2023)	OCR, OR	6	✘	En	1	✘	200	218	✘
Q-Bench (Wu et al., 2024)	VQA	3	✘	En	1	✘	3.4K	2.9K	✘
MaRVL (Liu et al., 2021)	Reasoning	1	✔	Id, Sw, Zh, Ta, Tr	5	✘	5K	*	✘
***VLURes* (Ours)**	OR, SU, RU, SS, IC, ITM, U, VQA	8	✔	En, Ja, Sw, Ur	4	✔[†]	4K	8K	✔

Table 2: Comparison of representative vision-language datasets and benchmarks. ✔ denotes support; ✘ denotes not supported. * indicates the original paper does not explicitly report the count. [†] *VLURes* supports evaluation *with* and *without* rationales (model-generated explanations) under a controlled prompting protocol. **Takeaway:** Most prior benchmarks use short text (captions or question prompts) and therefore under-measure *long-text grounding*; *VLURes* is designed around web-realistic article-length prose and includes an explicit unrelatedness test.

been essential for rapid progress, but they typically present *short* text (a single question or caption) and rarely require the model to identify which parts of a long document are actually supported by the image. Consequently, they under-measure a key real-world failure mode for multimodal assistants and intelligent agents, namely *long-text grounding*: selecting and reasoning over the small image-relevant subset of article-length text while ignoring distractors.

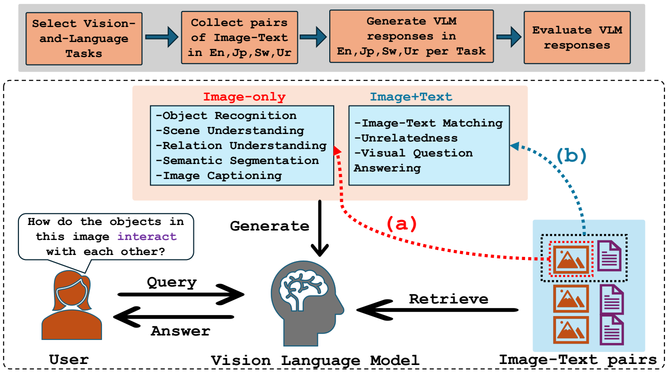
Multilingual vision-language evaluation. Multilingual benchmarks show that performance can drop sharply under language shift and that coverage differs by model and language (Pfeiffer et al., 2022; Geigle et al., 2024). MaRVL (Liu et al., 2021) extends visual reasoning evaluation to five languages, including Swahili, but still uses short prompts and does not target long document grounding. In contrast, VLURes uses web-curated article-length prose in four languages (En, Ja, Sw, Ur) and separates image-only perception from image+text grounding. This design aligns with evidence that multilingual robustness is uneven and that evaluation protocols can materially change conclusions about model capability (Pfeiffer et al., 2022; Geigle et al., 2024; Chen et al., 2023b).

Instruction-tuned VLMs and grounding under distractors. Modern VLMs typically pair a vision encoder with a LLM and use instruction tuning to support multi-task behavior and natural language interaction (Li et al., 2022; Dai et al., 2023; Liu et al., 2023b,c). Instruction tuning improves usability, but it can also encourage fluent, plausible generation even when supervision is weak or the textual context contains irrelevant content. This motivates evaluation settings that explicitly test faithfulness and grounding against noisy, web-realistic inputs. VLURes operationalizes this by providing article-length text with naturally occurring tangents and by introducing *Unrelatedness* (U), which measures whether a model can discard text that is not supported by the image.

Reasoning prompts and explanation faithfulness. Chain-of-thought and rationale prompting can improve accuracy by eliciting intermediate reasoning steps (Wei et al., 2022b; Kojima et al., 2022). However, generated explanations are not always faithful to the model's true decision process and can become brittle under distribution shift or language shift (Turpin et al., 2023). To disentangle gains in performance from explanation-induced artifacts, VLURes evaluates models both *with* and *without* rationales under matched prompting regimes.

3 *VLURes* Benchmark

Goal. VLURes is designed to evaluate VLMs in the regime most common for intelligent agents and multimodal assistants: an image is encountered together with *article-length text* that contains both relevant evidence and distracting, ungrounded content. Each example is an *(image, long-text)* pair (Figure 2), in-



(a) ***VLURes* development process. Top:** We curate web-derived image-text pairs, instantiate them into eight tasks, and evaluate VLM outputs using LLM-as-a-judge plus human alignment checks. We provide image-text pairs in **En, Ja, Sw, Ur. Bottom:** Tasks explicitly separate *image-only* reasoning (red, OR/SU/RU/SS/IC) from *image+text* reasoning (blue, ITM/U/VQA).

(b) **Urdu Wikipedia** article (`locations`). **Web pairing with long-text grounding.** For each page, we collect all images and the full article text, then select the most representative image via language-adapted CLIP similarity.

Figure 2: **Left:** *VLURes* design overview. **Right:** Web-realistic Wikipedia article containing multiple images and long text (texts are truncated for brevity, more examples in Figure 6). **Takeaway:** Each *VLURes* instance preserves *article-level prose*, hence models must ground against long-form context rather than captions. *VLURes* measures what many benchmarks miss, whether VLMs can ground to *article-length text* and discard unrelated content under language shift.

stantiable into one of eight tasks (Table 1), enabling diagnosis of (i) *image-only* visual competence and (ii) *long-text grounding* under realistic web noise.

3.1 How to Use *VLURes*

Given an image-text pair, evaluation proceeds as follows: (i) choose a task-specific prompt, (ii) provide the model with the image and the associated long text, (iii) request the task output (optionally with a rationale), and (iv) score outputs using our automatic and human-evaluation protocol (§5). This setup tests if models can select a suitable evidence source for the task at hand, *image-only* for OR/SU/RU/SS/IC and *image+text* for ITM/U/VQA, introduced in Table 1, see detailed task prompts in Tables 55–58.

3.2 Curation Principles

We curate VLURes from web pages because they naturally pair images with rich discourse, matching how agents retrieve and consume information in the wild (e.g., browsing news, product pages, travel descriptions, and encyclopedic entries). We follow five principles. **(1) Task breadth:** eight tasks probe complementary perception and grounding skills (Table 1). **(2) Language nativeness:** each split is collected from native sources in that language to reduce translation artifacts. **(3) Reasoning diversity:** tasks are split into *image-only* vs *image+text* families to isolate long-text grounding failures. **(4) VLM compatibility:** each instance is formatted as a single image plus a single long-text field, consumable by modern VLM APIs. **(5) Long-text grounding stress:** we retain article-level prose so models must identify grounded evidence and ignore unrelated paragraphs, rather than relying on caption-style shortcuts.

3.3 Language Selection

We select four languages to induce diversity along multiple axes. **Family diversity:** English (Indo-European), Japanese (Japonic), Swahili (Niger-Congo), Urdu (Indo-European). **Geographic diversity:** we follow WALS-style macro-area distinctions (Dryer and Haspelmath, 2013). **Script diversity:** Latin (En, Sw), Kanji/Kana (Ja), Nastaliq (Ur). **Resource diversity:** we include lower-resource languages (Sw, Ur) alongside higher-resource languages (En, Ja), reflecting real multilingual deployment gaps (Nigatu et al., 2024).

3.4 Data Sources, Collection, and Safety Filtering

Sources and categories. We collect pages from Wikipedia and Wikinews (permissive licensing) and complement them with language-specific web sources such as major news outlets, travel and lifestyle sites, and community platforms (see Appendix Table 52 for domain-level provenance). To diversify both visual and cultural contexts, we target

10 topical categories: *animals, products, buildings, locations, events, food, drinks, hobbies, works of art, organizations*.

Collection and filtering. For each URL, we download the full article text and all page images. We keep standard image formats (`png/jpeg/jpg`), and exclude obvious non-content media by filtering image URLs containing tokens such as `logo`, `button`, `icon`, `plugin`, `widget`. We discard pages with no valid downloadable images. We apply basic cleanup (remove empty-text pages and non-target-language pages). For safety, we manually inspect images to remove NSFW content, and we additionally check documents against the `List of Dirty, Naughty, Obscene or Otherwise Bad Words`.[1]

3.5 Long-Text Grounding via Language-Adapted Alignment

A central challenge in web curation is that a single page may contain multiple images, including unrelated illustrations, icons, and tangential media. To ensure each example pairs the long text with a representative image, we perform a language-adapted CLIP alignment step.

Sentence-image similarity. Starting from CLIP ViT-L/14 (Radford et al., 2021), we compute sentence-to-image cosine similarities for all candidate images in a page. Because CLIP is English-dominant and similarity quality drops for non-English text, we adapt CLIP per non-English language using a light contrastive fine-tuning step on a small set of manually verified image-text pairs, then use the adapted encoder for alignment within that language (see example in Figure 3).

Assignment and selection. We treat alignment as a bipartite assignment problem (Hessel et al., 2019; Kuhn, 1955) between candidate images and sentences. Operationally, we (i) discard images whose maximum similarity to any sentence is below a threshold (0.15 in our pipeline), filtering weakly grounded pages, and (ii) select the image with the highest remaining similarity as the representative image for the page. We then pair this selected image with the *full* article text, yielding a single *(image, long-text)* example.

[1] The full list is available at `https://github.com/LDNOOBW/List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words`.

Metric	**En**	**Ja**	**Sw**	**Ur**
#Images	1000	1000	1000	1000
#Texts	1000	1000	1000	1000
Text Lengths				
Min. Length	12	46	14	10
Max. Length	1716	3993	7766	3712
Median Length	242	381	335	231
Avg. Length	270	447	392	373

Table 3: Statistics of *VLURes* (evaluation split). Lengths are in words for En/Sw/Ur and characters for Ja. **Takeaway:** Article-length text is the core stressor that forces explicit evidence selection and filtering.

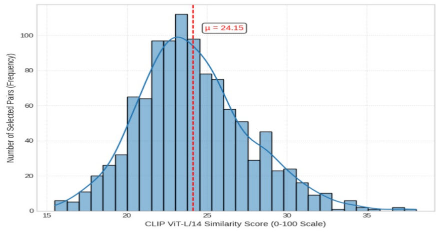


Figure 3: Distribution of **Sw** CLIP similarity scores.

3.6 Dataset Statistics and Long-Text Stress

Table 3 summarizes VLURes. We balance the benchmark to support controlled cross-language evaluation, using **4,000** evaluation pairs total (**1,000 per language**). We additionally retain extra verified pairs for some languages for fine-tuning analyses (reported in Appendix §M.1.3). VLURes texts are substantially longer than caption-style benchmarks such as MSCOCO (Lin et al., 2015) and Flickr30k (Plummer et al., 2016). For context, prior reports place average caption length around 10–14 tokens for common caption datasets (Sidorov et al., 2020; Liu et al., 2024), while Visual Genome region descriptions are typically short phrases (Krishna et al., 2017b). However, VLURes preserves long-form discourse, including distractors and tangents, making it a direct testbed for *long-text grounding* rather than caption matching.

4 Difficulty and Cross-Language Stability

Long-text grounding introduces multiple difficulty sources: longer documents contain more distractors, weaker alignment increases ambiguity, and pages with multiple plausible images can confuse selection. We characterize intrinsic difficulty using three signals. Let x be an example with long text t and a set of candidate images. We define (i) $L(x)$, text length, (ii) $S(x)$, maximum language-adapted similarity between sentence embeddings and the selected image, and (iii) $A(x)$, the margin between the best and runner-up images. These signals are intended to support analysis rather than training, and they enable

Model family	Models evaluated
Proprietary	GPT-4o, GPT-4o-mini, Gemini 2.0 Flash Lite, Gemini 1.5 Flash 8B
Open VLMs	LLaVA-NeXT (7B, 13B), PALO (7B, 13B), MAYA (8B), Qwen2-VL (7B)

Table 4: VLMs evaluated on *VLURes*. **Takeaway:** *VLURes* covers both frontier proprietary models and popular open VLMs with varying multilingual support.

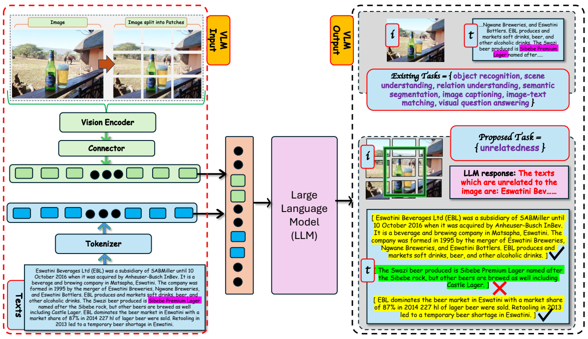


Figure 4: **Unrelatedness: Long-text grounding in *VLURes*.** Examples where the article contains both image-relevant (green) and irrelevant (yellow) paragraphs. The *Unrelatedness* task tests whether the model can ignore ungrounded text. **Takeaway:** Article-length inputs introduce realistic distractors that expose grounding failures not captured by short-caption benchmarks, and *Unrelatedness* measures VLM robustness to these distractors.

comparisons across languages by focusing on measurable properties of the paired inputs. This approach follows the broader evaluation principle that multilingual benchmarks should report difficulty and robustness rather than only aggregate accuracy (Geigle et al., 2024; Pfeiffer et al., 2022).

5 Models and Experimental Setup

Models. We evaluate 10 VLMs spanning proprietary and open models. Table 4 lists the model set. We include strong proprietary baselines (GPT-4o, GPT-4o-mini, Gemini 2.0 Flash Lite, Gemini 1.5 Flash 8B) and widely used open VLMs (LLaVA-NeXT, PALO, MAYA, Qwen2-VL). This selection reflects the practical landscape where open models are increasingly deployed but often lag in multilingual VL alignment (Chen et al., 2023b; Geigle et al., 2024).

Prompting regimes and outputs. We test four prompting regimes: **zero-shot vs one-shot, each with and without rationales**. Rationale prompting can improve performance but is not guaranteed to improve robustness, especially under language shift (Wei et al., 2022b; Turpin et al., 2023). Unless otherwise stated, models receive image and full article text. For cross-lingual experiments, we vary input language while controlling output language.

Evaluation. We score outputs using an LLM-as-a-judge framework and validate judge alignment with human ratings (Appendix §P). This design enables scalable evaluation across tasks and languages, while acknowledging the importance of judge reliability for credible conclusions (Geigle et al., 2024).

6 Main Results

Overall performance. Tables 5 and 6 summarize overall performance across tasks and languages under the best-performing prompting regime per model. Two trends stand out. **First**, strong proprietary models achieve high accuracy across tasks, yet still exhibit clear headroom against human annotation, especially on tasks requiring long-text grounding. **Second**, open models lag substantially and show larger cross-lingual sensitivity. These findings mirror broader multilingual evaluation results, where multilingual competence and alignment remain uneven and strongly protocol-dependent (Pfeiffer et al., 2022; Geigle et al., 2024; Chen et al., 2023b).

Long-text grounding is the dominant bottleneck. Long-text grounding tasks (ITM, U, VQA) remain the main source of errors across models (Table 5). Even under our strongest setting (one-shot, with rationales), GPT-4o is very high on image-only OR (90.8%, En output) yet VQA is lower and more sensitive to language shift (90.8% En vs. 88.5% Sw and 88.4% Ur outputs). For open models, the gap persists after fine-tuning: Qwen2VL-7B reaches 71.3% on OR (En output) but is lower on ITM (64.9%) and VQA (68.0%) under one-shot with rationales (Table 6). Human ceilings show the same difficulty pattern, with VQA (81.9) markedly below ITM (94.7) and U (93.8) (Table 5). These trends match prior concerns that instruction-tuned models can over-trust long, noisy text, under language shift (Pfeiffer et al., 2022; Geigle et al., 2024; Turpin et al., 2023).

Prompting helps, but does not reliably close language gaps. One-shot and rationales often improve absolute accuracy (Wei et al., 2022b; Kojima et al., 2022), but do not reliably remove cross-lingual gaps (Table 5). For GPT-4o-mini (En output), OR rises 80.0%→82.5% and ITM rises 78.0%→81.9% from zero-shot w/o rationales to one-shot w/ rationales, yet ITM still drops 81.9→76.8 when switching output language En→Ja (5.1 points), and VQA drops 82.6→78.7 for En→Sw (3.9 points). Rationales can also hurt: in zero-shot, VQA decreases for GPT-4o-mini (81.0%→76.0%) and Gemini 2.0 Flash Lite (84.7%→79.7%) when adding rationales, consistent

Model	Object Recognition				Scene Understanding				Relation Understanding				Semantic Segmentation				Image Captioning				Image-Text Matching				Unrelatedness				Visual Question Answering			
	En	Ja	Sw	Ur	En	Ja	Sw	Ur	En	Ja	Sw	Ur	En	Ja	Sw	Ur	En	Ja	Sw	Ur	En	Ja	Sw	Ur	En	Ja	Sw	Ur	En	Ja	Sw	Ur
Zero-shot, *Without Rationales*																																
GPT-4o	89.8	88.5	86.0	87.8	89.0	88.0	88.1	87.4	88.9	83.9	85.7	82.5	84.9	80.7	82.9	80.9	85.1	79.0	83.9	78.5	89.8	**89.0**	88.8	86.3	91.4	90.8	91.0	90.2	89.4	89.2	86.8	87.5
GPT-4o-mini	80.0	82.8	78.4	80.2	80.1	79.8	78.5	78.9	78.8	75.2	76.1	75.8	77.3	72.8	73.7	74.5	78.3	71.8	73.4	72.5	78.0	77.3	80.5	79.7	83.0	81.9	81.6	82.0	81.0	81.7	76.8	77.4
Gemini 2.0 Flash Lite	83.7	86.5	82.1	83.9	83.8	83.5	82.2	82.6	82.5	78.9	79.8	79.5	81.0	76.5	77.4	78.2	81.9	75.5	77.1	76.1	81.7	81.0	84.2	83.4	86.7	85.6	85.3	85.7	84.7	85.0	80.5	81.1
Gemini 1.5 Flash 8B	78.3	81.1	76.7	78.5	78.4	78.1	76.8	77.2	77.1	73.5	74.4	74.1	75.6	71.1	72.0	72.8	76.6	70.1	71.7	70.8	76.3	75.6	78.8	78.0	81.3	80.2	79.9	80.3	79.3	79.6	75.1	75.7
LlaVa Mistral 7B	38.6	41.4	0.0	0.0	38.6	37.8	0.0	0.0	37.3	33.7	0.0	0.0	36.2	30.9	0.0	0.0	36.6	30.2	0.0	0.0	37.2	36.4	0.0	0.0	41.5	40.4	0.0	0.0	40.2	39.5	0.0	0.0
Qwen2VL 7B	60.0	62.7	0.0	0.0	60.9	59.0	0.0	0.0	58.0	55.0	0.0	0.0	57.2	52.3	0.0	0.0	58.4	51.0	0.0	0.0	58.3	57.4	0.0	0.0	63.0	61.7	0.0	0.0	61.4	61.4	0.0	0.0
PALO 7B	39.5	40.5	0.0	50.7	39.6	37.8	0.0	48.9	38.3	33.2	0.0	45.9	36.8	30.8	0.0	44.8	37.8	29.6	0.0	42.5	37.5	35.3	0.0	49.7	42.5	39.9	0.0	52.0	40.5	39.3	0.0	47.4
MAYA 8B	40.8	46.6	0.0	49.8	40.9	43.6	0.0	48.5	39.6	39.0	0.0	45.5	38.1	36.6	0.0	44.1	39.3	35.0	0.0	42.1	38.8	41.1	0.0	49.3	43.8	45.7	0.0	51.6	41.8	45.1	0.0	47.0
LlaVa Mistral 13B	46.9	49.8	0.0	0.0	46.9	46.4	0.0	0.0	45.5	42.3	0.0	0.0	44.2	39.5	0.0	0.0	45.4	38.6	0.0	0.0	45.0	44.3	0.0	0.0	50.1	48.7	0.0	0.0	47.9	48.4	0.0	0.0
PALO 13B	41.1	43.4	0.0	57.2	41.1	40.4	0.0	55.0	39.8	36.0	0.0	52.9	38.6	33.9	0.0	51.0	39.0	32.4	0.0	49.0	39.2	38.1	0.0	56.5	44.3	42.9	0.0	59.2	42.5	42.3	0.0	54.0
Zero-shot, *With Rationales*																																
GPT-4o	89.1	86.5	88.2	87.5	89.7	88.3	87.3	86.2	88.0	85.9	86.3	85.2	88.3	86.2	86.0	84.5	86.5	82.2	83.3	82.3	89.5	86.7	87.0	86.0	91.3	90.7	90.8	90.3	88.2	86.7	86.0	86.1
GPT-4o-mini	80.0	83.0	78.3	81.9	79.8	79.9	78.5	78.4	79.6	79.8	77.7	76.4	80.0	80.7	76.8	75.6	78.2	75.4	71.6	72.7	79.4	78.0	75.5	76.2	83.1	82.1	80.9	82.4	76.0	76.6	76.6	78.3
Gemini 2.0 Flash Lite	83.7	86.7	81.8	85.6	83.5	83.6	82.2	82.1	83.3	83.6	81.4	80.0	83.7	84.4	80.5	79.3	81.9	79.0	75.3	76.4	83.0	81.7	79.2	79.9	86.8	85.8	84.6	86.1	79.7	80.3	80.3	81.9
Gemini 1.5 Flash 8B	78.3	81.3	76.4	80.2	78.1	78.2	76.8	76.7	77.9	78.4	76.0	74.7	78.3	79.0	75.1	73.9	76.5	73.7	69.9	71.0	77.7	76.3	73.8	74.5	81.4	80.4	79.2	80.7	74.3	74.9	74.9	76.6
LlaVa Mistral 7B	39.4	41.9	0.0	0.0	38.3	38.4	0.0	0.0	38.1	38.1	0.0	0.0	39.2	39.3	0.0	0.0	37.0	34.4	0.0	0.0	38.0	36.7	0.0	0.0	42.0	41.3	0.0	0.0	34.7	35.1	0.0	0.0
Qwen2VL 7B	60.0	63.1	0.0	0.0	59.0	59.9	0.0	0.0	59.7	59.7	0.0	0.0	60.3	60.4	0.0	0.0	58.7	55.8	0.0	0.0	59.3	58.5	0.0	0.0	63.0	62.4	0.0	0.0	56.6	56.9	0.0	0.0
PALO 7B	39.5	41.5	0.0	51.9	39.3	37.2	0.0	48.0	39.1	37.9	0.0	46.4	39.5	38.7	0.0	45.6	37.8	33.3	0.0	42.7	38.9	36.0	0.0	46.2	42.6	40.1	0.0	52.4	35.5	34.6	0.0	48.3
MAYA 8B	40.8	46.8	0.0	51.6	40.6	43.7	0.0	48.1	40.4	43.5	0.0	46.0	40.8	44.5	0.0	45.0	39.0	39.2	0.0	42.3	40.2	41.8	0.0	45.8	43.9	45.9	0.0	52.1	36.8	40.4	0.0	47.9
LlaVa Mistral 13B	47.2	50.1	0.0	0.0	45.9	46.0	0.0	0.0	46.6	46.6	0.0	0.0	47.2	47.5	0.0	0.0	45.3	41.9	0.0	0.0	46.4	45.0	0.0	0.0	49.9	49.2	0.0	0.0	43.0	43.1	0.0	0.0
PALO 13B	41.0	44.0	0.0	58.8	40.8	40.7	0.0	55.4	40.6	40.7	0.0	53.3	41.4	41.5	0.0	52.4	39.3	36.4	0.0	49.4	40.4	39.5	0.0	53.2	44.3	43.6	0.0	59.7	37.3	37.7	0.0	55.3
One-shot, *Without Rationales*																																
GPT-4o	90.5	87.2	87.3	86.6	90.5	88.1	88.4	87.6	89.4	87.4	87.3	84.4	89.1	85.3	86.0	84.9	88.0	78.9	85.3	79.7	89.1	86.8	87.2	86.1	91.0	90.5	91.0	90.6	89.2	87.0	88.1	86.2
GPT-4o-mini	82.0	79.2	79.0	78.2	82.1	79.2	78.9	78.6	79.6	78.1	76.9	75.8	81.5	77.9	77.4	75.7	80.0	73.0	74.4	73.3	80.2	76.9	77.9	75.2	83.0	82.5	82.5	82.2	81.7	78.9	77.6	77.1
Gemini 2.0 Flash Lite	85.7	82.9	82.7	81.9	85.8	82.9	82.6	82.3	83.3	81.8	80.6	79.5	85.2	81.6	81.1	79.4	83.7	76.7	78.1	77.0	83.9	80.6	81.6	78.9	86.7	86.2	86.2	85.9	85.4	82.6	81.3	80.8
Gemini 1.5 Flash 8B	80.3	77.5	77.3	76.5	80.4	77.5	77.2	76.9	77.9	76.5	75.2	74.1	79.8	76.2	75.7	74.0	78.3	71.3	72.7	71.6	78.5	75.2	76.2	73.5	81.3	80.8	80.8	80.5	80.0	77.2	75.9	75.4
LlaVa Mistral 7B	40.6	37.8	0.0	0.0	41.2	37.9	0.0	0.0	38.0	36.5	0.0	0.0	40.0	36.0	0.0	0.0	39.3	32.3	0.0	0.0	39.0	35.4	0.0	0.0	42.0	41.2	0.0	0.0	40.0	36.7	0.0	0.0
Qwen2VL 7B	62.0	59.1	0.0	0.0	62.1	59.4	0.0	0.0	59.6	58.1	0.0	0.0	61.7	57.9	0.0	0.0	60.3	53.0	0.0	0.0	60.3	56.5	0.0	0.0	63.7	62.5	0.0	0.0	61.1	58.5	0.0	0.0
PALO 7B	41.5	37.7	0.0	48.4	41.6	37.2	0.0	48.6	39.1	36.0	0.0	45.4	41.0	35.7	0.0	45.9	39.5	31.4	0.0	43.3	39.7	34.8	0.0	45.2	42.5	40.5	0.0	52.2	41.2	36.9	0.0	47.1
MAYA 8B	42.8	43.0	0.0	47.8	42.9	43.0	0.0	48.2	40.4	41.8	0.0	45.3	42.3	41.8	0.0	45.3	40.8	36.8	0.0	42.9	41.0	40.7	0.0	44.8	43.8	46.3	0.0	51.8	42.5	42.7	0.0	46.7
LlaVa Mistral 13B	49.3	46.4	0.0	0.0	49.2	46.3	0.0	0.0	46.3	45.4	0.0	0.0	48.3	44.6	0.0	0.0	47.2	40.3	0.0	0.0	47.4	43.4	0.0	0.0	50.3	49.3	0.0	0.0	47.9	45.4	0.0	0.0
PALO 13B	43.7	40.2	0.0	55.2	43.3	40.2	0.0	55.0	40.6	39.1	0.0	52.6	42.4	38.5	0.0	52.5	41.1	34.0	0.0	50.4	41.2	37.6	0.0	52.4	44.6	43.7	0.0	59.2	42.4	39.7	0.0	54.1
One-shot, *With Rationales*																																
GPT-4o	**90.8**	**89.8**	**89.9**	**88.1**	**90.7**	**88.8**	**88.8**	**87.9**	**90.4**	**89.1**	**89.2**	**87.7**	**90.8**	**89.1**	**89.3**	**86.4**	**88.5**	**84.9**	**85.5**	**84.0**	**90.3**	88.3	**88.9**	**87.8**	**91.7**	**91.3**	**91.6**	**91.0**	**90.8**	**90.7**	**88.5**	**88.4**
GPT-4o-mini	82.5	83.3	80.1	80.4	82.3	80.2	78.4	78.2	81.6	81.8	78.9	78.5	82.6	80.5	79.7	78.2	80.3	76.6	76.0	74.7	81.9	76.8	78.6	79.3	83.3	81.1	82.0	82.7	82.6	79.8	78.7	79.3
Gemini 2.0 Flash Lite	86.2	87.0	83.8	84.1	85.9	83.8	82.1	81.8	85.3	85.5	82.6	82.2	86.3	84.2	83.4	81.9	84.0	80.3	79.7	78.4	85.6	80.5	82.3	83.0	87.0	84.8	85.7	86.4	86.3	83.5	82.4	83.0
Gemini 1.5 Flash 8B	80.8	81.6	78.6	78.7	80.5	78.5	76.7	76.5	79.9	80.3	77.4	76.8	80.9	78.8	78.0	76.5	78.6	74.9	74.3	73.0	80.2	75.1	76.9	77.6	81.6	79.4	80.3	81.0	80.9	78.1	77.0	77.6
LlaVa Mistral 7B	40.6	42.4	0.0	0.0	40.7	38.9	0.0	0.0	39.5	40.2	0.0	0.0	40.5	39.2	0.0	0.0	39.4	34.5	0.0	0.0	40.2	35.0	0.0	0.0	41.5	40.0	0.0	0.0	41.1	38.1	0.0	0.0
Qwen2VL 7B	62.5	63.6	60.5	60.5	62.2	60.3	58.0	58.4	61.0	61.6	58.7	58.6	62.6	60.3	59.6	58.3	60.7	56.4	56.5	54.8	61.3	56.1	58.5	59.3	63.5	61.3	62.2	62.8	62.5	59.2	58.3	59.3
PALO 7B	42.2	41.8	0.0	50.4	41.7	38.1	0.0	48.1	41.1	39.2	0.0	48.3	42.1	38.5	0.0	48.6	39.8	34.6	0.0	44.0	41.4	34.8	0.0	49.3	42.8	39.1	0.0	52.3	42.1	37.8	0.0	49.5
MAYA 8B	43.3	47.7	0.0	50.0	43.0	44.0	0.0	47.8	42.4	45.6	0.0	48.1	43.4	44.3	0.0	47.8	41.1	40.4	0.0	44.3	42.7	40.5	0.0	48.9	44.1	44.9	0.0	52.3	43.4	43.6	0.0	48.6
LlaVa Mistral 13B	49.5	50.4	0.0	0.0	49.4	47.4	0.0	0.0	48.6	48.8	0.0	0.0	49.6	47.8	0.0	0.0	47.4	43.6	0.0	0.0	48.5	43.5	0.0	0.0	50.4	48.3	0.0	0.0	49.2	46.0	0.0	0.0
PALO 13B	43.5	44.1	0.0	57.8	43.3	41.2	0.0	55.4	42.5	42.7	0.0	55.5	43.6	41.3	0.0	55.0	41.3	37.5	0.0	51.3	42.3	37.0	0.0	56.0	44.0	42.9	0.0	59.7	43.0	40.8	0.0	56.0
Human Performance	97.5	–	–	–	98.8	–	–	–	99.0	–	–	–	97.5	–	–	–	98.7	–	–	–	94.7	–	–	–	93.8	–	–	–	81.9	–	–	–

Table 5: Performance of VLMs on **eight VL tasks** under zero-shot and one-shot settings, measured by estimated Accuracy (%) from LLM-as-a-judge. **Input:** English Texts + Images; **Output:** En, Ja, Sw, Ur responses. Shaded columns represent {En} in input and {En} in output VLM results. Zeros indicate unintelligible responses.

with brittleness and unfaithful explanations (Turpin et al., 2023). Finally, prompting cannot compensate for missing multilingual VL support in some open models (0.0 Sw/Ur entries persist even after fine-tuning in Table 6), so we treat the four prompting regimes as distinct operating points. Additional results in Appendix §F (Ja), §G (Sw), and §H (Ur).

7 Cross-Lingual Robustness

We evaluate robustness by varying the *language interface* while keeping tasks and evaluation fixed (same images and long-text inputs, but different target languages for responses). The goal is to quantify how much performance depends on language handling and whether grounding is preserved under language shift. Prior work reports substantial language sensitivity in multilingual multimodal QA and reasoning (Pfeiffer et al., 2022; Geigle et al., 2024), and VLURes extends this to *long-text* inputs where the linguistic load is higher.

Robustness summary. Even strong models show consistent, non-trivial gaps across output languages in our most accurate regime (one-shot with rationales; Table 5): for GPT-4o, Object Recognition drops from 90.8 (En) to 88.1 (Ur) (-2.7), Image-Text Matching from 90.3 (En) to 87.8 (Ur) (-2.5), and VQA from 90.8 (En) to 88.4 (Ur) (-2.4). Language-heavy generation can be more sensitive: Image Captioning drops from 88.5 (En) to 84.0 (Ur) (-4.5). These patterns support that the bottleneck is often the *language interface* (instruction following and selecting evidence from long-form text), not only perception, consistent with uneven multilingual VL alignment and instruction-tuning coverage. For open models, fine-tuning improves accuracy where language is supported (e.g., Qwen2VL 7B OR rises from 62.5 to 71.3 in En under one-shot with rationales; Tables 5 and 6), but zero entries for some languages indicate that *language support itself* remains a limiting factor for robust deployment. Additional results in the Appendix §K for En, Ja, Sw, and Ur.

Comparison with MaRVL. Main message: *VLURes* is harder than MaRVL because it requires long-text grounding and filtering. MaRVL (Liu et al., 2021) (also used in IGLUE (Bugliarello et al., 2022)) provides short, caption-like text, whereas *VLURes* embeds each image in *article-length* prose with distractors that must be filtered. Under same prompting protocol, GPT-4o performs higher on MaRVL Swahili than on *VLURes* Swahili (Table 7), with largest gains on text-dependent tasks (U: +10.7, VQA: +7.8), consistent with added difficulty of

Model	Object Recognition				Scene Understanding				Relation Understanding				Semantic Segmentation				Image Captioning				Image-Text Matching				Unrelatedness				Visual Question Answering			
	En	Ja	Sw	Ur	En	Ja	Sw	Ur	En	Ja	Sw	Ur	En	Ja	Sw	Ur	En	Ja	Sw	Ur	En	Ja	Sw	Ur	En	Ja	Sw	Ur	En	Ja	Sw	Ur
Zero-shot, *Without Rationales*																																
LlaVa Mistral 7B	48.9	51.7	0.0	0.0	48.9	48.1	0.0	0.0	47.6	44.0	0.0	0.0	46.5	41.2	0.0	0.0	46.9	40.5	0.0	0.0	47.5	46.7	0.0	0.0	51.8	50.7	0.0	0.0	50.5	49.8	0.0	0.0
Qwen2VL 7B	68.8	71.5	0.0	0.0	69.7	67.8	0.0	0.0	66.8	63.8	0.0	0.0	66.0	61.1	0.0	0.0	67.2	59.8	0.0	0.0	67.1	66.2	0.0	0.0	71.8	70.5	0.0	0.0	70.2	70.2	0.0	0.0
PALO 7B	48.3	49.3	0.0	59.5	48.4	46.6	0.0	57.7	47.1	42.0	0.0	54.7	45.6	39.6	0.0	53.6	46.6	38.4	0.0	51.3	46.3	44.1	0.0	58.5	51.3	48.7	0.0	60.8	49.3	48.1	0.0	56.2
MAYA 8B	49.6	55.4	0.0	58.6	49.7	52.4	0.0	57.3	48.4	47.8	0.0	54.3	46.9	45.4	0.0	52.9	48.1	43.8	0.0	50.9	47.6	49.9	0.0	58.1	52.6	54.5	0.0	60.4	50.6	53.9	0.0	55.8
LlaVa Mistral 13B	54.8	57.7	0.0	0.0	54.8	54.3	0.0	0.0	53.4	50.2	0.0	0.0	52.1	47.4	0.0	0.0	53.3	46.5	0.0	0.0	52.9	52.2	0.0	0.0	58.0	56.6	0.0	0.0	55.8	56.3	0.0	0.0
PALO 13B	49.9	52.2	0.0	66.0	49.9	49.2	0.0	63.8	48.6	44.8	0.0	61.7	47.4	42.7	0.0	59.8	47.8	41.2	0.0	57.8	48.0	46.9	0.0	65.3	53.1	51.7	0.0	68.0	51.3	51.1	0.0	62.8
Zero-shot, *With Rationales*																																
LlaVa Mistral 7B	49.7	52.2	0.0	0.0	48.6	48.7	0.0	0.0	48.4	48.7	0.0	0.0	49.5	49.6	0.0	0.0	47.3	44.7	0.0	0.0	48.3	47.0	0.0	0.0	52.3	51.6	0.0	0.0	45.0	45.4	0.0	0.0
Qwen2VL 7B	68.8	71.9	0.0	0.0	67.8	68.7	0.0	0.0	68.5	68.5	0.0	0.0	69.1	69.2	0.0	0.0	67.5	64.6	0.0	0.0	68.1	67.3	0.0	0.0	71.8	71.2	0.0	0.0	65.4	65.7	0.0	0.0
PALO 7B	48.3	50.3	0.0	60.7	48.1	46.0	0.0	56.8	47.9	46.7	0.0	55.2	48.3	47.5	0.0	54.4	46.6	42.1	0.0	51.5	47.7	44.8	0.0	55.0	51.4	48.9	0.0	61.2	44.3	43.4	0.0	57.1
MAYA 8B	49.6	55.6	0.0	60.4	49.4	52.5	0.0	56.9	49.2	52.3	0.0	54.8	49.6	53.3	0.0	53.8	47.8	48.0	0.0	51.1	49.0	50.6	0.0	54.6	52.7	54.7	0.0	60.9	45.6	49.2	0.0	56.7
LlaVa Mistral 13B	55.1	58.0	0.0	0.0	53.8	53.9	0.0	0.0	54.6	54.6	0.0	0.0	55.1	55.4	0.0	0.0	53.3	49.9	0.0	0.0	54.4	53.0	0.0	0.0	57.9	57.2	0.0	0.0	50.9	51.0	0.0	0.0
PALO 13B	49.8	52.8	0.0	67.6	49.6	49.5	0.0	64.2	49.4	49.5	0.0	62.1	50.2	50.3	0.0	61.2	48.1	45.2	0.0	58.2	49.2	48.3	0.0	62.0	53.1	52.4	0.0	68.5	46.1	46.5	0.0	64.1
One-shot, *Without Rationales*																																
LlaVa Mistral 7B	50.9	48.1	0.0	0.0	51.5	48.2	0.0	0.0	48.3	46.8	0.0	0.0	50.3	46.3	0.0	0.0	49.6	42.6	0.0	0.0	49.3	45.7	0.0	0.0	52.3	51.5	0.0	0.0	50.3	47.0	0.0	0.0
Qwen2VL 7B	70.8	67.9	0.0	0.0	70.9	68.2	0.0	0.0	68.4	66.9	0.0	0.0	70.5	66.7	0.0	0.0	69.1	61.8	0.0	0.0	69.1	65.3	0.0	0.0	72.5	71.3	0.0	0.0	69.9	67.3	0.0	0.0
PALO 7B	50.3	46.5	0.0	57.2	50.4	46.0	0.0	57.4	47.9	44.8	0.0	54.2	49.8	44.5	0.0	54.7	48.3	40.2	0.0	52.1	48.5	43.6	0.0	54.0	51.3	49.3	0.0	61.0	50.0	45.7	0.0	55.9
MAYA 8B	51.6	51.8	0.0	56.6	51.7	51.8	0.0	57.0	49.2	50.6	0.0	54.1	51.1	50.6	0.0	54.1	49.6	45.6	0.0	51.7	49.8	49.5	0.0	53.6	52.6	55.1	0.0	60.6	51.3	51.5	0.0	55.5
LlaVa Mistral 13B	57.2	54.3	0.0	0.0	57.1	54.2	0.0	0.0	54.2	53.3	0.0	0.0	56.2	52.5	0.0	0.0	55.1	48.2	0.0	0.0	55.3	51.3	0.0	0.0	58.2	57.2	0.0	0.0	55.8	53.3	0.0	0.0
PALO 13B	52.5	49.0	0.0	64.0	52.1	49.0	0.0	63.8	49.4	47.9	0.0	61.4	51.2	47.3	0.0	61.3	49.9	42.8	0.0	59.2	50.0	46.4	0.0	61.2	53.4	52.5	0.0	68.0	51.2	48.5	0.0	62.9
One-shot, *With Rationales*																																
LlaVa Mistral 7B	50.9	52.7	0.0	0.0	51.0	49.2	0.0	0.0	49.8	50.5	0.0	0.0	50.8	49.5	0.0	0.0	49.7	44.8	0.0	0.0	50.5	45.3	0.0	0.0	51.8	50.3	0.0	0.0	51.4	48.4	0.0	0.0
Qwen2VL 7B	**71.3**	**72.4**	0.0	0.0	**70.9**	**69.1**	0.0	0.0	**69.8**	**70.4**	0.0	0.0	**71.4**	**69.1**	0.0	0.0	**65.2**	**65.3**	0.0	0.0	**64.9**	**67.3**	0.0	0.0	**70.1**	**71.0**	0.0	0.0	**68.0**	**67.1**	0.0	0.0
PALO 7B	51.0	50.6	0.0	59.2	50.5	46.9	0.0	56.9	49.9	48.0	0.0	57.1	50.9	47.3	0.0	57.4	48.6	43.4	0.0	52.8	50.2	43.6	0.0	58.1	51.6	47.9	0.0	61.1	50.9	46.6	0.0	58.3
MAYA 8B	52.1	56.5	0.0	58.8	51.8	52.8	0.0	56.6	51.2	54.4	0.0	56.9	52.2	53.1	0.0	56.6	49.9	49.2	0.0	53.1	51.5	49.3	0.0	57.7	52.9	53.7	0.0	61.1	52.2	52.4	0.0	57.4
LlaVa Mistral 13B	57.4	58.3	0.0	0.0	57.3	55.3	0.0	0.0	56.5	56.7	0.0	0.0	57.5	55.7	0.0	0.0	55.3	51.5	0.0	0.0	56.4	51.4	0.0	0.0	58.3	56.2	0.0	0.0	57.1	53.9	0.0	0.0
PALO 13B	52.3	52.9	0.0	66.6	52.1	50.0	0.0	64.2	51.3	51.5	0.0	64.3	52.4	50.1	0.0	63.8	50.1	46.3	0.0	60.1	51.1	45.8	0.0	64.8	52.8	51.7	0.0	68.5	51.8	49.6	0.0	64.8

Table 6: Performance of VLMs on **eight VL tasks** under finetuning settings, measured by estimated Accuracy (%) from LLM-as-a-judge. **Input:** English Texts + Images; **Output:** En, Ja, Sw, Ur responses. Shaded columns represent {En} in input and {En} in output VLM results. Zeros indicate unintelligible responses.

Model	Object Recognition		Scene Understanding		Relation Understanding		Semantic Segmentation		Image Captioning		Image-Text Matching		Unrelatedness		Visual Question Answering	
	Sw	ΔAcc.	Sw	ΔAcc.	Sw	ΔAcc.	Sw	ΔAcc.	Sw	ΔAcc.	Sw	ΔAcc.	Sw	ΔAcc.	Sw	ΔAcc.
GPT-4o	88.3	+2.7	87.6	+2.9	87.8	+2.0	85.4	+0.5	88.0	+7.3	89.2	+5.1	97.8	10.7	94.3	+7.8

Table 7: Performance of VLMs on **eight VL tasks**, measured by estimated Accuracy (%) from LLM-as-a-judge. **Input:** Swahili Texts + Images; **Output:** Sw responses. All images and texts are from the **MaRVL** dataset. We report the best results under *one-shot with rationales* setting, from GPT-4o.

Model	Base LLM	Language support (VL tasks)
LLaVA 7B/13B	Mistral/Vicuna	Primarily English without multilingual VL fine-tuning.
PALO 7B/13B	Llama-based	Primarily English, multilingual VL depends on fine-tuning.
MAYA 8B	Aya-based	Strong English and Indic languages, broad VL alignment.
Qwen2-VL	Qwen series	Strong En and Zh, good support for Fr, Es, De, Ar, Ja, Ko, Sw, Ur.

Table 8: Language coverage of open VLMs. **Takeaway:** Multilingual text ability does not imply multilingual VL grounding, which matters most under long-text inputs.

long-text grounding and unrelatedness filtering in *VLURes*. See Appendix §I for per-task details.

8 Language Support of Open VLMs

A practical barrier to multilingual benchmarking is open models differ in their true VL alignment coverage. Table 8 separates base LLM multilinguality from VL language readiness, because multilingual text ability does not guarantee multilingual VL grounding. This distinction is crucial when interpreting cross-lingual results, under long-text inputs where tokenization and instruction-following stability are key (Geigle et al., 2024; Chen et al., 2023b).

9 Validity and Robustness Checks

Representative-Image Selection Validity. We further tested whether selecting a single representative image per article via CLIP similarity biases *VLURes* toward easy-match samples, and whether residual misalignment confounds conclusions about grounding difficulty. To address this, we conducted a stratified manual audit of 75 image-article pairs per language (300 total), sampled across three ambiguity-margin bins: **low** ($A(x) < 0.05$), **medium** ($0.05 \leq A(x) < 0.12$), and **high** ($A(x) \geq 0.12$), where $A(x)$ is the margin between the best and runner-up candidate images (§4). Two native-speaking annotators per language independently judged whether selected image is truly most representative of the article, resolving disagreements by consensus, Table 9.

Language	Low margin	Med. margin	High margin	Overall (%)
English	88.0	96.0	100.0	94.7 ± 2.6
Japanese	84.0	94.0	98.7	92.2 ± 3.0
Swahili	82.7	92.0	97.3	90.7 ± 3.3
Urdu	80.0	90.7	96.0	88.9 ± 3.6
All	83.7	93.2	98.0	**91.6** ± 1.6

Table 9: Manual audit: % of selected images judged correct (representative) by native-speaking annotators, stratified by ambiguity margin bin. We computed 95% confidence intervals via the Wilson score method. Misalignment is mainly in the low-margin (ambiguous) stratum and is prominent in lower-resource languages, as expected.

Sensitivity analysis. We additionally varied the CLIP alignment threshold $\tau \in \{0.10, 0.15, 0.20\}$

and, for ambiguous pages ($A(x) < 0.05$), replaced top-1 selection with top-k ($k \in \{2, 3\}$) random draw, re-running evaluation on a 200-instance held-out subset. Overall accuracy on image-only tasks changed by $\leq$ 0.4 points across conditions; for grounding tasks (ITM, U, VQA) the range was ≤ 0.8 points. These stable trends confirm that core conclusions about **long-text grounding difficulty and cross-lingual degradation are robust to moderate variation in alignment hyperparameters**, and are not an artifact of the easy-match selection regime.

Evaluation: Beyond a Single LLM-Judge. We extend the human-alignment study reported in Appendix §P with stratified agreement and a second, non-Gemini judge to address concerns about judge bias. For the second judge we use `claude-opus-4` (Anthropic), applying an identical evaluation prompt (Table 54). Human–judge agreement in Table 10 is strong for image-only tasks and remains acceptable for grounding tasks and lower-resource languages (Sw/Ur ICC $= 0.783$), where task difficulty and output diversity are highest. The Gemini and Claude judges agree strongly

Comparison	Subset	ICC	Kendall τ	Top-3 overlap
Human vs. Gemini judge	Image-only tasks	0.841	—	—
Human vs. Gemini judge	Grounding tasks	0.805	—	—
Human vs. Gemini judge	Sw / Ur (low-res.)	0.783	—	—
Gemini vs. Claude judge	All tasks (En)	—	0.912	3/3
Gemini vs. Claude judge	All tasks (Ja/Sw/Ur)	—	0.887	3/3
Gemini vs. Claude judge	Grounding tasks	—	0.871	3/3

Table 10: Judge robustness. **Top:** Intraclass correlation (ICC) between human annotators and the Gemini 1.5 Pro judge, stratified by task family and language resource level (300 judgements per cell). **Bottom:** Rank-order agreement (Kendall τ) and top-3 model-ranking overlap between the Gemini and Claude judges on a 400-instance cross-lingual subset. All cross-lingual and cross-task trends are consistent across both judges.

on model rankings ($\tau \geq 0.87$) with identical top-3 model ordering across all conditions, ruling out systematic Gemini-family bias.

Cross-Lingual Consistency Visualization. Figure 5 visualizes per-task accuracy drops relative to English for the four strongest VLMs under the best prompting regime (one-shot with rationales), condensing the full cross-lingual story from Appendix §K into a single interpretable figure. Image-only tasks (OR–IC) show moderate, consistent degradation across languages, while grounding tasks (ITM, U, VQA) reveal *task-dependent asymmetry*: ITM often improves under non-English input (positive Δ), whereas VQA degrades most under Swahili. This pattern confirms that the dominant bottleneck

Model / Lang	OR	SU	RU	SS	IC	ITM	U	VQA
GPT-4o Ja	-3.7	-3.9	-4.4	-4.5	-3.5	5.9	0.3	1.1
GPT-4o Sw	-5.0	-4.2	-4.0	-4.6	-5.1	4.5	1.9	3.8
GPT-4o Ur	-1.0	-1.7	-2.1	-1.5	2.3	-0.7	2.9	2.6
GPT-4o-mini Ja	-4.0	-3.6	-3.7	-3.7	-3.0	6.7	0.3	0.7
GPT-4o-mini Sw	-5.0	-5.2	-2.6	-4.7	-4.9	4.9	-0.1	0.3
GPT-4o-mini Ur	-1.5	1.7	-1.5	-1.5	2.1	-1.0	-3.0	-0.7
Gemini 2.0 Ja	-4.1	-4.0	-3.7	-3.6	-3.0	6.5	0.8	1.0
Gemini 2.0 Sw	-4.3	-5.2	-2.7	-3.8	-4.4	4.7	-1.7	0.3
Gemini 2.0 Ur	-1.6	-1.0	-1.0	-0.9	2.1	-1.6	1.5	1.5
Gemini 1.5 Ja	-3.7	-3.8	-3.7	-3.5	-3.0	5.6	0.8	0.8
Gemini 1.5 Sw	-4.3	-5.0	-2.7	-3.7	-4.7	4.6	-1.7	0.1
Gemini 1.5 Ur	-1.6	-2.0	-1.4	-1.6	5.0	-2.0	1.9	1.2

≤ -4.5 -4.5 to -1.5 -1.5 to $+1.5$ $+1.5$ to $+4.5$ $\geq +4.5$

Figure 5: Δ Acc. vs. English (in points) under *one-shot with rationales* for the four closed models across languages (Ja/Sw/Ur) and eight tasks. Red = drop; blue = gain. Image-only tasks (OR–IC) degrade consistently; ITM often *improves* under non-English input while VQA is most sensitive to Swahili.

is the *language interface under long-text inputs*, not visual perception alone.

Dataset and release plan. We have released the relevant data, code, and task prompts at the URL `https://jatuhurrra.github.io/VLURes/`, under permissive terms.

10 Conclusion

We addressed a gap in VLM evaluation for real-agent settings: caption-level and English-centric benchmarks under-measure *long text grounding*, where models must identify which parts of article-length text are supported by the image and ignore distractors. ***VLURes*** resolves this by providing **4,000** web-curated *image plus long text* pairs across En, Ja, Sw, and Ur. By separating image-only perception tasks from image-plus-text grounding tasks and introducing **Unrelatedness**, *VLURes* makes grounding failures measurable. In our strongest setting, one-shot with rationales, the best model, GPT-4o, remains below human performance by 6.7 points on Object Recognition, 8.1 on Scene Understanding, 10.2 on Image Captioning, 4.4 on Image Text Matching, and 2.1 on Unrelatedness. These results show that long-text grounding tasks can be competitive with top models, yet a substantial gap remains in core perception and grounded generation. Overall, prompting improves accuracy but does not reliably remove cross-lingual sensitivity, and open models benefit from fine-tuning when multilingual support is available. We hope *VLURes* will motivate research into methods for grounded long-text reasoning.

Limitations

Coverage and representativeness. *VLURes* evaluates four languages and ten image categories, which enables controlled cross language comparison, but it does not cover many other scripts, dialectal variation, code switching, or interaction heavy settings that are common in embodied systems. As a result, our findings may not directly transfer to domains such as egocentric video, spoken dialogue, or tool use without further data collection and evaluation.

Web sourced data bias. We curate image text pairs from web pages, including news, blogs, and encyclopedic articles, which improves ecological validity, but inherits selection bias, reporting bias, and culturally specific framings present in the underlying sources. These issues are well documented for large scale language resources and motivate careful dataset documentation and auditing when releasing web derived benchmarks (Bender and Friedman, 2018; Bender et al., 2021).

Image text alignment constraints. Our pipeline selects a single image per article using CLIP based similarity, after light language adaptation with a small amount of image plus text data per non English language. Even with this adaptation, alignment errors can remain, because CLIP was trained primarily on English paired data (Radford et al., 2021), and multilingual extensions of CLIP show that cross language alignment remains an open challenge (Chen et al., 2023a). This limitation can bias the benchmark toward samples whose visual content is easy to match from surface text.

Evaluation noise and judge bias. We score model outputs using an LLM judge and a small scale native speaker study. LLM based evaluation can be inconsistent and sensitive to prompt wording (Zheng et al., 2023), and limited human annotation can introduce variance from rater fatigue or interpretation differences. We mitigate these issues with shared rubrics and double annotation, but absolute scores should be interpreted together with the observed trends across tasks, languages, and settings, and we encourage future work to expand the human evaluation and test multiple judges (Duan et al., 2024).

Ethical Considerations

Data sourcing, licensing, and traceability. *VLURes* is curated from publicly available web pages that pair images with article-level text. Following dataset documentation best practices (Gebru et al., 2021; Bender and Friedman, 2018), we store provenance metadata for each example (source URL, retrieval date, language, and category) so the benchmark can be audited and problematic items can be removed. We prioritize sources with permissive reuse terms, such as Wikimedia projects. When redistribution of raw text or images is restricted, we release only what is permitted, for example URLs and derived annotations, rather than redistributing the original content.

Privacy, people in images, and sensitive events. Web pages may include identifiable individuals, minors, and depictions of sensitive events, even when the intent is informational (Birhane and Prabhu, 2021; Weidinger et al., 2021). We apply multiple filtering steps, including automated removal of common non-content images, keyword-based screening, and manual review to reduce NSFW material. Despite these efforts, residual privacy risks may remain. We discourage uses involving biometric identification, profiling, or surveillance, and we recommend that downstream users follow local privacy regulations and perform additional human review before any deployment in high-stakes settings.

Representation, cultural framing, and language skews. Web data reflects uneven geographic, topical, and cultural coverage, which can amplify stereotypes or omit locally salient concepts, especially in lower-resourced languages (Bender and Friedman, 2018; Bender et al., 2021). As a result, performance gaps on *VLURes* may reflect both model capability and differences in what content is available across language communities. We therefore report per-language statistics, encourage disaggregated analysis by language and category, and provide a mechanism for users to flag and request removal of problematic examples.

Potential misuse and responsible release. Improving multimodal instruction following can support helpful applications, such as accessibility tools and safer embodied assistants, but it can also lower the cost of harmful capabilities, including multimodal disinformation and intrusive monitoring (Weidinger et al., 2021; Bommasani et al., 2021). We release *VLURes* for research and evaluation, with transparency about sources and limitations. We encourage responsible use norms, including avoiding deployment in safety-critical contexts without additional risk assessment, red teaming, and domain-specific safeguards.

References

2023. LAION-GPT-V. https://huggingface.co/datasets/laion/gpt4v-dataset. Accessed via Hugging Face Datasets; eprint: 2311.12793.

Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. 2016. VQA: Visual question answering. *Preprint*, arXiv:1505.00468.

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2019. nocaps: novel object captioning at scale. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8948–8957.

Jesse Atuhurra, Hidetaka Kamigaito, Hiroki Ouchi, Hiroyuki Shindo, and Taro Watanabe. 2024. Nersocial: Efficient named entity recognition dataset construction for human-robot interaction utilizing rapidner. *arXiv preprint arXiv:2412.09634*.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623. Association for Computing Machinery.

Abeba Birhane and Vinay Uday Prabhu. 2021. Large image datasets, a pyrrhic win for computer vision? *arXiv preprint arXiv:2006.16923*.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, and 1 others. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

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. *arXiv preprint arXiv:2201.11732*.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2024. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pages 370–387. Springer.

Qian Chen, Zhenzhong Li, Wang Wang, Jianguo Li, Xianjing Wu, Ziran Wang, Yafan Wang, Yujia Li, Lijun Wang, Jing Liu, and 1 others. 2023a. mCLIP: Multilingual clip via cross-lingual transfer. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Xi Chen, Xiao Wang, Lucas Beyer, Alexander Kolesnikov, Jialin Wu, Paul Voigtlaender, Basil Mustafa, Sebastian Goodman, Ibrahim Alabdulmohsin, Piotr Padlewski, Daniel Salz, Xi Xiong, Daniel Vlasic, Filip Pavetic, Keran Rong, Tianli Yu, Daniel Keysers, Xiaohua Zhai, and Radu Soricut. 2023b. Pali-3 vision language models: Smaller, faster, stronger. *Preprint*, arXiv:2310.09199.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. UNITER: UNiversal image-TExt representation learning. In *European Conference on Computer Vision*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards general-purpose vision-language models with instruction tuning.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, and 1 others. 2023. PaLM-e: An embodied multimodal language model. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 8469–8488. PMLR.

Matthew S. Dryer and Martin Haspelmath. 2013. The world atlas of language structures online. Accessed on 2025-05-09.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, and 1 others. 2024. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11198–11201.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30K: Multilingual English-German image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. 2023. MME: A comprehensive evaluation benchmark for multimodal large language models.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah M Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, and 15 others. 2023. DataComp: In search of the next generation of multimodal datasets. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Gregor Geigle, Radu Timofte, and Goran Glavaš. 2024. Babel-ImageNet: Massively multilingual evaluation of vision-and-language representations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5064–5084. Association for Computational Linguistics.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017a. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. *Preprint*, arXiv:1612.00837.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017b. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018a. VizWiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. 2018b. VizWiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jack Hessel, Lillian Lee, and David Mimno. 2019. Unsupervised discovery of multimodal links in multi-image, multi-sentence documents. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2034–2045. Association for Computational Linguistics.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. 2023. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. In *Thirty-Seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Drew A Hudson and Christopher D Manning. 2019. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Inaam Ilahi, Hafiz Muhammad Abdullah Zia, Muhammad Ahtazaz Ahsan, Rauf Tabassam, and Armaghan Ahmed. 2021. Efficient urdu caption generation using attention based LSTM. *Preprint*, arXiv:2008.01663.

Kushal Kafle, Scott Cohen, Brian Price, and Christopher Kanan. 2018. DVQA: Understanding data visualizations via question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kushal Kafle and Christopher Kanan. 2017. An analysis of visual question answering algorithms. *Preprint*, arXiv:1703.09684.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. ReferItGame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798. Association for Computational Linguistics.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. *Preprint*, arXiv:1603.07396.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. OCR-free document understanding transformer. In *European Conference on Computer Vision (ECCV)*.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22. Curran Associates Inc.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017a. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017b. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73.

Harold W. Kuhn. 1955. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023a. SEED-Bench: Benchmarking multimodal LLMs with generative comprehension. *Preprint*, arXiv:2307.16125.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning (ICML)*, pages 12888–12900. PMLR.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft COCO: Common objects in context. *Preprint*, arXiv:1405.0312.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. *Preprint*, arXiv:2310.03744.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. In *Advances in Neural Information Processing Systems*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023c. Visual instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Haoyu Liu, Yaoxian Song, Xuwu Wang, Zhu Xiangru, Zhixu Li, Wei Song, and Tiefeng Li. 2024. Flickr30k-cfq: A compact and fragmented query dataset for text-image retrieval. *Preprint*, arXiv:2403.13317.

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. 2023d. MMBench: Is your multi-modal model an all-around player?

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. *Preprint*, arXiv:2310.02255.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems*.

Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. 2021. IconQA: A new benchmark for abstract diagram understanding and visual language reasoning. In *Proceedings of the NeurIPS 2021 Datasets and Benchmarks Track*.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. *Preprint*, arXiv:1511.02283.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. OK-VQA: A visual question answering benchmark requiring external knowledge. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ahmed Masry, Do Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279. Association for Computational Linguistics.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. 2021. DocVQA: A dataset for VQA on document images. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2199–2208. IEEE Computer Society.

Michael McCloskey and Neal J. Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. volume 24 of *Psychology of Learning and Motivation*, pages 109–165. Academic Press.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. OCR-VQA: Visual question answering by reading text in images. In *International Conference on Document Analysis and Recognition (ICDAR)*.

Takashi Miyazaki and Nobuyuki Shimizu. 2016. Cross-lingual image caption generation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1780–1790. Association for Computational Linguistics.

Hellina Hailu Nigatu, Atnafu Lambebo Tonja, Benjamin Rosman, Thamar Solorio, and Monojit Choudhury. 2024. The zeno's paradox of 'low-resource' languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17753–17774. Association for Computational Linguistics.

Open-Orca / SlimOrca contributors. 2023. SlimOrca: An open dataset of GPT-4 augmented FLAN reasoning traces, with verification. Accessed: 2025-05-09.

Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. Im2Text: Describing images using 1 million captioned photographs. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu

Wei. 2023. Kosmos-2: Grounding multimodal large language models to the world. *Preprint*, arXiv:2306.14824.

Jonas Pfeiffer, Gregor Geigle, Aishwarya Kamath, Jan-Martin O. Steitz, Stefan Roth, Ivan Vulić, and Iryna Gurevych. 2022. xGQA: Cross-lingual visual question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2497–2511. Association for Computational Linguistics.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2016. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *Preprint*, arXiv:1505.04870.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022a. LAION-5B: An open large-scale dataset for training next generation image-text models. *Preprint*, arXiv:2210.08402.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022b. Laion-5b: An open large-scale dataset for training next generation image-text models. *Preprint*, arXiv:2210.08402.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-OKVQA: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision (ECCV)*, page 146–162.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. 2020. TextCaps: a dataset for image captioning with reading comprehension. In *European Conference on Computer Vision*.

Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. 2019. Towards VQA models that can read. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Shivalika Singh, Freddie Vargus, Daniel D'souza, Börje Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura O'Mahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergun, Ifeoma Okoh, and 14 others. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11521–11567. Association for Computational Linguistics.

Ashish V. Thapliyal, Jordi Pont Tuset, Xi Chen, and Radu Soricut. 2022. Crossmodal-3600: A massively multilingual multimodal evaluation dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 715–729. Association for Computational Linguistics.

Tristan Thrush, Ethan A. Mellor, Jessica L. Angel, Achintya Prasad, Divyansh Mittal, Joydeep Paul, and Michael C. Mozer. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5238–5248.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022a. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, and 1 others. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, and Weisi Lin. 2024. Q-Bench: A benchmark for general-purpose foundation models on low-level vision. In *International Conference on Learning Representations (ICLR)*.

Jiahong Wu, He Zheng, Bo Zhao, Yixin Li, Baoming Yan, Rui Liang, Wenjia Wang, Shipei Zhou, Guosen Lin, Yanwei Fu, Yizhou Wang, and Yonggang Wang. 2019. Large-scale datasets for going deeper in image understanding. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE.

Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. 2017. Video Question Answering via Gradually Refined Attention over Appearance and Motion. In *Proceedings of ACM Multimedia*.

Zhiyang Xu, Ying Shen, and Lifu Huang. 2023. MultiInstruct: Improving multi-modal zero-shot learning via instruction tuning. In *61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. 2022. Learning to answer visual questions from web videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Yuta Yoshikawa, Tomoyuki Iwata, and Yoshihiro Hagiwara. 2017. STAIR captions: Constructing a large-scale japanese image caption dataset. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 417–421. Association for Computational Linguistics.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. MM-Vet: Evaluating large multimodal models for integrated capabilities. *Preprint*, arXiv:2308.02490.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2023. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. *Preprint*, arXiv:2311.16502.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and chatbot arena. *Preprint*, arXiv:2306.05685.

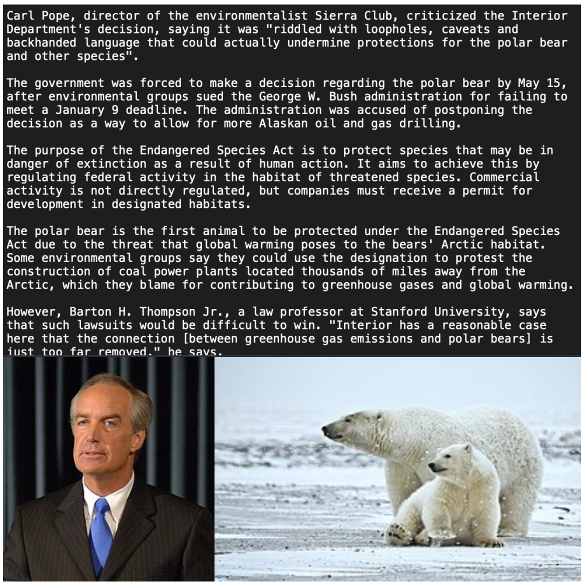
Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, Huong Tran, Radu Soricut, Anikait Singh, Jaspiar Singh, Pierre Sermanet, Pannag R. Sanketi, Grecia Salazar, and 35 others. 2023. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pages 2165–2183. PMLR.

Appendix Table of Contents

A Dataset Examples

In this section, we illustrate the final image–text pairs in *VLURes* and report the corresponding CLIP similarity scores used during alignment. Table 11 present representative examples from Urdu and Swahili, the two low-resource languages in our benchmark. For each example, we start from a single web article that contains multiple images and long-form text, then compute CLIP image–text similarity between each candidate image and the article content to identify the image that is most strongly grounded in the discourse. Because off-the-shelf CLIP is primarily trained on English, we first adapt CLIP with a small amount of paired image–text data in Japanese, Swahili, and Urdu, then apply the adapted model to score candidate (image, text) pairs in each language. We additionally filter out non-content images (e.g., icons, buttons, logos) and discard low-similarity candidates, which helps prevent spurious matches when articles mention many entities that are not visually depicted. Resulting similarity values shown in tables therefore serve two purposes: they provide transparency about strength of the grounding signal for each selected pair, and they demonstrate meaningful alignment is achievable even under long-text settings in low-resource languages, where both scripts and vocabulary differ substantially from English.

Dataset and release plan. We have released the relevant data, code, and task prompts at the URL `https://jatuhurrra.github.io/VLURes/`, under permissive terms. Raw article text and images are subject to source-site licensing; where redistribution is restricted, we release only URLs, consistent with common practice for web-derived benchmarks (Thapliyal et al., 2022; Liu et al., 2021). A data statement and removal request mechanism also accompany the release.



(a) **English Wikinews** article (`animals`).

(b) **Japanese** article (`events`).

(c) **Swahili** post (`events`).

(d) **Urdu Wikipedia** article (`locations`).

Figure 6: **Web pairing with long-text grounding.** For each page, we collect all images and the full article text, then select the most representative image via language-adapted CLIP similarity. Texts shown are truncated for brevity. **Takeaway:** Each VLURes instance preserves *article-level prose*, so models must ground against long-form context rather than captions. Due to space limitations, we have shown only part of the texts from each respective source.

Text	Image	CLIP Similarity
KANDA YA ZIWA KUONJA RADHA YA ROYAL BOXING TOUR. Na.Khadija Seif, Michuzi blog. KATIKa kuunga mkono juhudi za Rais Dk. Samia Suluhu Hassan za kutangaza na kukuza utalii wa Tanzania, kampuni ya HB Sacs Boxing inataraji kuandaa pambano la 'Royal Tour Boxing Kanda ya Ziwa litakalofanyika Septemba 30 mwaka huu jijini Mwanza katika Ukumbi wa Rock City Mall. Akizungumza na Wanahabari Promota wa pambano hilo, Hassan Kumbucha amesema lengo la kuandaa mtanange huo ni kutangaza utalii, kuhamasisha mchezo huo Kanda ya Ziwa na kuleta uzalendo. "Kampuni yetu imekuja kitofauti sana tunahitaji kutoa fursa za utalii, pia kuleta hamasa Kwa mabondia wa kanda ya Ziwa na kuunga mkono Rais Dk. Samia, "alisema Kambucha. Promota huyo ameweka wazi kuwa pambano kuu litakuwa kati ya bondia mtanzania, Fadhili Majiha 'Stopper' dhidi ya Renz Rosia raia wa Ufilipino kuwania mkanda wa Ubingwa wa ABU. "Kabla ya pambano la Majiha na mfilipino kutakuwa na mapambano ya vijana ambayo yatafanyika Karagwe Buringi Septemba 25 mwaka huu baada hapo wataenda kuangalia utalii Chato Septemba 26. Bondia, Fadhili Majiha amesema amefurahi kupata nafasi hiyo ya kucheza katika pambano kubwa ambalo linatangaza utalii wa ndani. Ambapo mara nyingi amekuwa akikosa nafasi ya kucheza, ila kupitia kampuni hiyo atapambana na kuonyesha uwezo wake. "Pongezi kwa kampuni ya HB Sacs Boxing Kwa kuandaa pambano hili na kuniamini, nimejiandaa vema na bado naendelea na mazoezi chini ya kocha wangu, Kwa me Hamisi watanzania waniombee dua mkanda ubaki nyumbani." Makamu Mwenyekiti wa Kamisheni ya Ngumi za Kulipwa Tanzania (TPBRC), Nassoro Chuma ameeleza kuwa jambo nzuri ni kuona wadau wa michezo wanamuunga mkono Rais Dk. Samia. Aidha mabondia ambao watapanda ulingoni siku hiyo wanatakiwa kutoa burudani na kucheza Kwa viwango vikubwa. Pamoja na Majiha na Mfilipino mapambano mengine ambayo yatasindikiza pambano hilo ni Saleh Kassim dhidi ya Freddy Sayuni wakati Stumai Muki atacheza na Engine Kayange huku Shomari Milundi na Abdullah Rashid na Ahmed Pelembela ataonyesha ubabe na Issa Maneva. Wengine Luckman Ramadhani atapigana na Khalid Karama wakati John Chua dhidi ya Ramadhani Kumbele huku Said Mkola akiwa na kibarua kizito na Francis Miyeyusho.		0.3806
MAJALIWA AZUNGUMZA NA WAZIRIi MKUU WA JAPAN SHINZO ABE. Waziri Mkuu, Kassim Majaliwa akiteta na Waziri Mkuu wa Japan, Shinzo Abe baada ya mazungumzo yao yaliyofanyika kwenye hoteli ya The New Otani iliyopo Tokyo nchini Japan Waziri Mkuu, Kassim Majaliwa na ujumbe wake wakiwa katika picha ya pamoja na Waziri Mkuu wa Japan, Shinzo Abe baada ya mazungumzo yao yaliyofanyika kwenye hoteli ya The New Otani ya Tokyo nchini Japan, Agosti 31, 2019. Kutoka kushoto ni Afisa Msimamizi wa Fedha na Mipango wa Zanzibar, Sharif Bakari, Afisa Msimamizi wa Fedha, Wizara ya Fedha Mipango, Athumani Msabila, Naibu Katibu Mkuu, Wizara ya Fedha na Mipango Zanzibar, Iddi Makame Haji, Kaimu Balozi wa Tanzania Nchini Japan, John Kambona, Naibu Katibu Mkuu Wizara ya Fedha na Mipango Zanzibar, Amina Khamis Shaaban, Kaimu Mkurugenzi, Idara ya Asia na Australia, Wizara ya Mambo ya Nje na Ushirikiano wa Afrika ya Mashariki, Size Waitara, Mganga Mkuu wa Serikali, Wizara ya Afya, Maendelo ya Jamii, Jinsia, Wazee na Watoto, Profesa Bakari Kambi na kulia ni Kamishina wa Fedha za Nje, Wizara ya Fedha na Mipango, John Lubuga. (Picha na Ofisi ya Waziri Mkuu).		0.3397
CASTLE LAGER KUWAPELEKA WATANZANIA 'WORLD CUP'. Meneja wa Bia ya Castle Lager, Pamela. Kikuli, akionyesha moja ya kipeperushi kinachoelezea mashindano maalumu ya soka yanayofahamika kama 'Castle Lager Africa 5s' (5 - Aside) yaliyozinduliwa leo katika viwanja vya Leaders, Kinondoni jijini Dar es salaam. Kulia ni Meneja Masoko, Udhamini na Promosheni za Wateja George Kavishe na kushoto ni Mkufunzi wa Timu ya Castle Lager Tanzania, Ivo Mapunda. Kampuni ya Bia Tanzania (TBL) kupitia kinywaji chake cha Bia ya Castle Lager, imezindua rasmi mashindano maalumu ya soka yanayofahamika kama 'Castle Lager Africa 5s' (5 - Aside). Mashindano haya yenye hadhi ya kimataifa yanashirikisha timu za wachezaji watano (5) kila upande, na kocha mmoja. Mechi zake zitakuwa zikichezwa kwa dakika saba (7) ili kata mshindi. Bingwa wa michuano hiyo ataiwakilisha Tanzania huko Zambia katika michuano ya kimataifa dhidi ya nchi nyingine 5 kutoka Afrika ambazo ni: Afrika Kusini, Zambia, Zimbabwe, Swaziland, Lesotho. Baada ya hapo mshindi atapata fursa ya kwenda Urusi kushuhudia michuano ya Kombe la Dunia baadaye mwaka huu, akiambatana na mashabiki wawili ambao watapatikana katika mchakato malumu pamoja na mchezaji mmoja mkongwe kutoka nchi ambayo ni bingwa. Kama Tanzania itafanikiwa kushinda, timu pamoja na mashabiki hao wawili wataongozana na mkongwe katika soka Ivo Mapunda aliyekua kipa wa Taifa Stars na Simba Sports Club. Akizungumza na waandishi wa habari jijini Dar es Salaam, Meneja wa Bia ya Castle Lager, Pamela Kikuli, alisema kuwa, kuna michakato maalumu ya kupata timu shiriki kupitia katika baa mbalimbali jijini Dar es Salaam. "Utaratibu wa kupata timu utawahusisha moja kwa moja mashabiki wa baa husika kupendekeza timu zao, kabla ya hatua nyingine za kuzipigia kura timu zinazotakiwa kushiriki. Mteja wetu atapata maelekezo maalumu kutoka kwa watu wetu ambao watakuwa wakipita katika baa hizo kwa nyakati tofauti. Sifa kubwa ya mteja wetu kushiriki katika mchakato huo ni pamoja na kununua bia ya Castle Lager," alisema Kikuli. Akitolea maelezo za ziada jinsi mchakato huo wa Castle Lager Africa 5s utakavyokua, Meneja Masoko, Udhamini na Promosheni za Wateja George Kavishe amesema kua, michuano hiyo inashirikisha baa 160, zitakazogawanywa katika makundi 10, huku kila kundi likiwa na baa 16. Kila kundi litatakiwa kuwa na kiwanja chake katika michuano ya awali ya timu 16 na kupata timu 8. Wilaya zote za Mkoa wa Dar es Salaam zitashiriki katika michuano hii. Hapa tukimaanisha Kinondoni, Ilala, Temeke, Ubungo na Kigamboni. Bonanza za kuanza kuchagua timu wakilishi za wilaya zitaanza tarehe 17 Machi hadi 21 Aprili kisha Bonanza kuu itakayoleta timu 10 za fainali itakua tarehe 28 Aprili 2018. "Katika hatua ya pili ya bonanza tutapata timu moja kutoka katika yale makundi 10 ya awali an kufikisha idadi ya timu 10 zitakazocheza katika bonanza kubwa pale Viwanja vya Leaders jijini Dar es Salaam, kupata timu tano (5) zitakazocheza kwa mtindo wa ligi na hatimaye kupata bingwa." Alisema Meneja Kavishe. Michuano hii inafanyika kwa miezi miwili, ambapo washiriki wanatakiwa kuwa na umri wa miaka kati ya 24 na 34. Lengo la mashindano haya ni kuwaleta pamoja marafiki na mashabiki wa soka. Wachezaji walioko katika ligi mbalimbali nchini hawatoruhusiwa kushiriki katika michuano hii, hata kama wana kigezo cha umri uliotajwa hapo juu.		0.3436

Table 11: A data instance retrieved from the ***VLURes*** benchmark comprising the *texts* and associated *images*, together with the CLIP ViT-B/32 image–text *similarity scores*. The language is **Swahili**.

B Radar plots for VLM performance across En, Ja, Sw, and Ur.

We visualize VLM performance under the *one-shot, with rationales* setting because it consistently yielded the strongest results across our prompt configurations. Figure 7(a–d) summarizes baseline performance on English, Japanese, Swahili, and Urdu, across our eight tasks. Two trends stand out. First, GPT-4o achieves the most uniform, high performance in every language, producing radar shapes that remain close to the outer ring, which indicates broad competence rather than isolated task strengths. Second, performance degrades markedly as we move from high-resource inputs (English, Japanese) to low-resource inputs (Swahili, Urdu), especially for open models, where the radar areas shrink and become less balanced across tasks. The proprietary models, particularly GPT-4o, GPT-4o-mini, and the Gemini variants, retain substantially better cross-language stability, suggesting that strong multilingual instruction following and grounded reasoning remain bottlenecks for current open VLMs in Swahili and Urdu, even when rationales are elicited.

To better understand how much of this gap is attributable to model capacity versus limited multilingual adaptation, we fine-tuned six open VLMs (LLaVA-Mistral 7B/13B, PALO 7B/13B, Qwen2VL 7B, and MAYA 8B) using *VLURes* training data, and again evaluate them under the *one-shot, with rationales* setting. Figure 7(e–g) shows that fine-tuning expands the radar area and improves balance across tasks for English and Japanese, indicating that our benchmark provides useful supervision for strengthening both image-only and image+text reasoning. For Urdu, fine-tuned models that can generate coherent outputs (PALO 7B/13B and MAYA 8B) show measurable gains, but the gap to proprietary models remains, highlighting the continued difficulty of long-text grounding and instruction compliance in a low-resource script. We omit Qwen2VL 7B for Urdu because its generations were frequently unintelligible in preliminary runs, and we do not report Swahili fine-tuning results because none of the fine-tuned open models produced reliably interpretable Swahili outputs. Overall, these plots suggest that while task-specific adaptation improves open-model performance in higher-resource settings, robust low-resource multilingual grounding, particularly for Swahili, remains a key open challenge.

C More Related Work

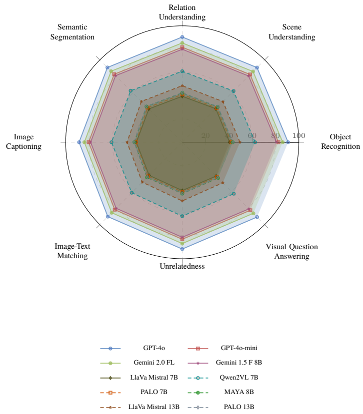
Beyond the benchmarks reviewed in § 2, we draw inspiration from task-specific vision and vision-language datasets. These resources have shaped our design goals for *VLURes*, namely multi-task coverage, multilingual evaluation, and long-text grounding. Table 2 summarizes representative datasets across common VL tasks, including image captioning (IC), referring expression generation and comprehension (REG, REC), and visual question answering (VQA).

Image Captioning. Large-scale captioning datasets such as MSCOCO (Lin et al., 2015), Flickr30k (Plummer et al., 2016), Conceptual Captions (Sharma et al., 2018), and Conceptual 12M (Changpinyo et al., 2021) are primarily English. Multilingual extensions exist, for example Multi30K (Elliott et al., 2016), and Japanese COCO-style resources such as STAIR Captions (Yoshikawa et al., 2017) and YJ Captions (Miyazaki and Shimizu, 2016). Additional efforts include AI Challenger for Chinese (Wu et al., 2019) and Crossmodal-3600 in 36 languages, including Japanese and Swahili (Thapliyal et al., 2022). While these datasets enable multilingual captioning, coverage for low-resource languages is often limited, and the supervision is typically short captions rather than article-level discourse. For example, the Swahili portion of Crossmodal-3600 is small (Thapliyal et al., 2022), and the Urdu extension in (Ilahi et al., 2021) is also limited in scale and scope. These limitations motivate our focus on long-text grounding and broader task coverage for Swahili and Urdu in *VLURes*.

Referring expressions (REG, REC). Referring expression benchmarks, including RefCOCO, RefCOCO+, and RefCOCOg (Kazemzadeh et al., 2014; Mao et al., 2016), study localized descriptions of target objects.[2] These datasets have been influential for grounding, but the expressions are often brief and strongly tied to a small set of visual domains (Kazemzadeh et al., 2014; Mao et al., 2016). *VLURes* complements this line by evaluating grounding under longer, naturally occurring discourse, where relevant and irrelevant sentences may co-exist.

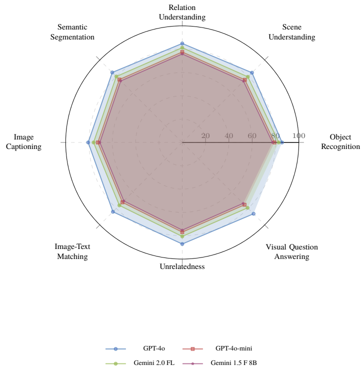
Visual Question Answering. VQA datasets such as VQA v2 (Goyal et al., 2017a), OK-VQA (Marino et al., 2019), TextVQA (Mishra et al., 2019), and

[2] REG produces a description for a *specific* object in an image. REC identifies the referred object given a natural language expression. In contrast, image captioning describes salient aspects of the whole scene.

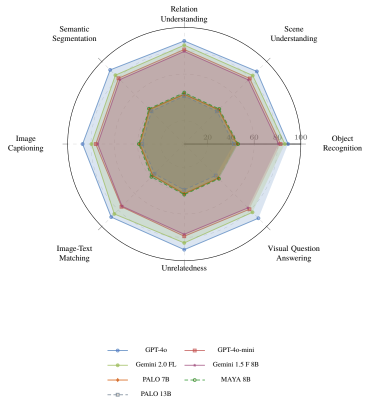


(a) English Input/Output

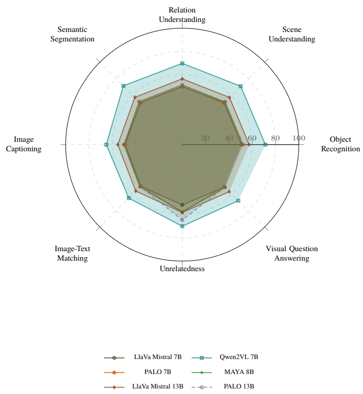
(b) Japanese Input/Output



(c) Swahili Input/Output

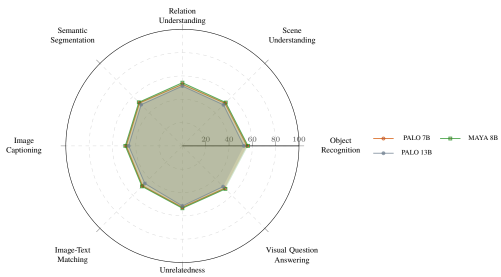


(d) Urdu Input/Output



(e) English Input/Output (finetuned)

(f) Japanese Input/Output (finetuned)



(g) Urdu Input/Output (finetuned)

Figure 7: Comparison of VLM performance on eight VL tasks under the *One-shot, With Rationales* setting across four languages. **Rows 1-2** (i.e., Figures a, b, c, d) show VLM performance before fine-tuning; **Rows 3-4** (i.e., Figures e, f, g) show model performance after fine-tuning the open VLMs. The open VLMs do not support Swahili, hence no results.

GQA (Hudson and Manning, 2019) have driven progress in multimodal reasoning. However, these benchmarks commonly provide limited per-instance context compared to web articles, and they are predominantly English. Visual Genome also supplies region-level descriptions and relations (Krishna et al., 2017b), but the supervision is typically short phrases rather than full discourse. These gaps motivate *VLURes*, where each instance retains article-level text, enabling evaluation of both grounding and modality selection in multilingual settings.

C.1 Vision and Language Task Definitions

Task structure. Across *VLURes*, each instance is an image paired with a single long text field, but tasks differ in which modality must be relied upon. OR, SU, RU, SS, and IC primarily require image-based reasoning, while ITM, U, and VQA explicitly require joint reasoning over the image and the accompanying text. The tasks are complementary: OR and SS provide object and region signals, SU and RU capture higher-level context and relations, and ITM, U, and VQA evaluate whether this understanding is correctly grounded in long-form text.

C.2 Formal Definition: Vision and Language Tasks

To provide a unified specification for *VLURes*, we define its task space in terms of inputs, outputs, and evaluation. Let $\mathcal{I}$ denote the space of images, and let $\mathcal{X}_{\text{txt}}$ denote the space of article-length texts paired with images. The benchmark comprises eight tasks,

$$\mathcal{T} = \{\tau_k\}_{k=1}^{8}, \tag{1}$$

where each task $\tau_k = (X_k, Y_k, \Phi_k)$ is characterized by:

- X_k: the input space for task τ_k,
- Y_k: the output space, representing the required response format,
- Φ_k: the evaluation function that maps predictions and gold labels to a scalar score.

We categorize tasks by their primary reasoning requirement:

(1) Image-only reasoning tasks.

$$\mathcal{T}_{\text{img}} = \{\tau_{\text{OR}}, \tau_{\text{SU}}, \tau_{\text{RU}}, \tau_{\text{SS}}, \tau_{\text{IC}}\}, \tag{2}$$

where for $\tau_k \in \mathcal{T}_{\text{img}}$, the input is predominantly visual, $X_k \subseteq \mathcal{I}$.

(2) Image-text reasoning tasks.

$$\mathcal{T}_{\text{img+txt}} = \{\tau_{\text{ITM}}, \tau_{\text{U}}, \tau_{\text{VQA}}\}, \tag{3}$$

where for $\tau_k \in \mathcal{T}_{\text{img+txt}}$, the input includes both modalities, $X_k \subseteq \mathcal{I} \times \mathcal{X}_{\text{txt}}$.

Task interdependencies. While tasks are evaluated separately, they form a coherent progression from low-level perception to grounded reasoning. Accurate object and region understanding, for example OR and SS, supports higher-level interpretation, for example SU and RU, and these capabilities are reflected in downstream generation and grounding tasks such as IC, ITM, U, and VQA. This structure motivates reporting per-task performance, as well as aggregate analyses of difficulty and cross-language stability in the main paper.

D Introducing the *Unrelatedness* Task

Recent progress in vision language models (VLMs) has enabled general purpose perception and instruction following that is increasingly used in intelligent agents, including embodied assistants (Driess et al., 2023). In these settings, the input is often inherently multimodal, namely text paired with another modality such as an image, video, audio, or speech. However, in natural environments, especially on the web, the accompanying text is rarely a single, clean caption. Instead, it is article-level discourse that mixes description, background, quotes, tangential facts, and occasionally content that is not visually grounded in the nearby image. For agents that must act safely and reliably, it is not enough to generate a plausible response, they must also decide which parts of the textual context should be ignored because they are not supported by the visual evidence.

Existing multimodal benchmarks partially touch this need through global image text alignment or controlled hard negatives, for example Winoground and SugarCrepe (Thrush et al., 2022; Hsieh et al., 2023). Yet, these settings typically assume caption length texts or sentence level alternatives, and they do not require models to localize and return the specific spans in long prose that are unrelated to the image. This gap matters because current VLMs can be distracted by language priors and can hallucinate unsupported visual content, which has motivated dedicated hallucination evaluations such as POPE (Li et al., 2023b). A benchmark that explicitly measures the ability to reject irrelevant textual evidence provides a direct handle on this failure mode, and it complements our long-text grounding setting in *VLURes*.

Task	Task Definition
Object Recognition (OR)	Identify the presence, category, or attributes of objects in an image, based on visual evidence, as commonly studied in COCO-style settings (Lin et al., 2015). *Example: cat, bottle, car.*
Scene Understanding (SU)	Infer the overall scene context and situation beyond isolated objects, including activities and environment cues (Kafle and Kanan, 2017). *Example: A girl sits on a bench in a park.*
Relationship Understanding (RU)	Identify and reason about relationships among objects, including spatial, interaction, and functional relations, as captured in structured annotations such as Visual Genome (Krishna et al., 2017b). *Example: A girl feeds a cat, a human to animal interaction.*
Semantic Segmentation (SS)	Assign semantic labels to image regions or pixels, for example road, building, person, a standard dense prediction task (Lin et al., 2015).
Image Captioning (IC)	Generate a natural language description of an image that summarizes salient content, including objects, actions, and relations (Lin et al., 2015).
Image-Text Matching (ITM)	Given an image and its accompanying article text, select the sentence or span that is best supported by the image, which tests grounded alignment under long context (Xu et al., 2023).
Unrelatedness (U)	A task introduced in *VLURes*. Given an image and article text, select the sentence or span that is least supported by the image, which tests filtering and robustness to distractors.
Visual Question Answering (VQA)	Answer a natural language question about an image by integrating visual understanding and language reasoning (Agrawal et al., 2016).

Table 12: Definitions of the eight VL tasks contained in ***VLURes***.

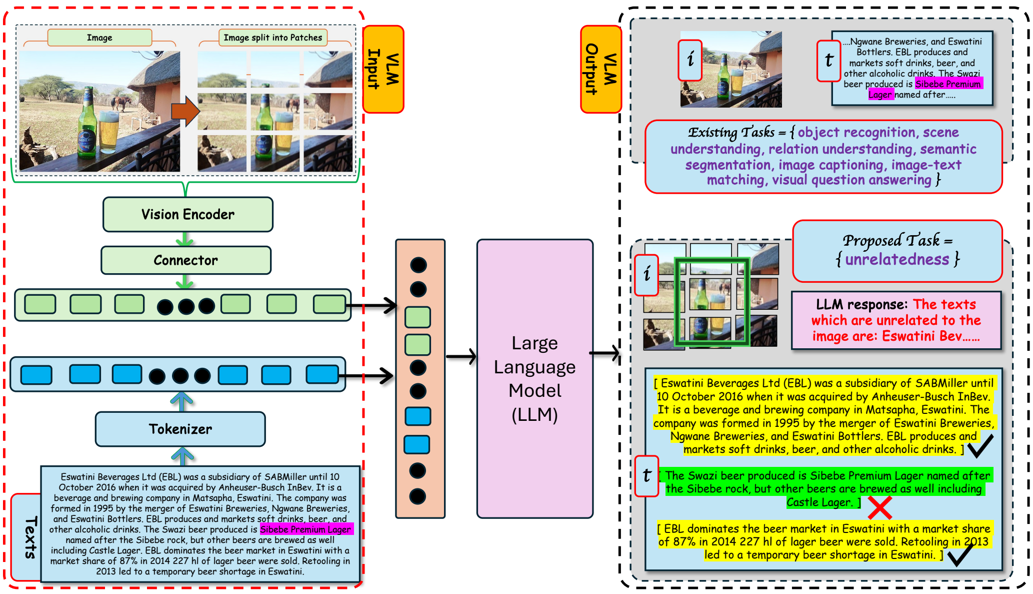


Figure 8: The proposed ***Unrelatedness*** task. **Left:** The VLM input consists of two modalities, an **image** and **text**. The image is processed by a vision encoder and connector to produce visual tokens. The text is tokenized to produce textual tokens. Both are aligned in a shared representation space and passed to the language model. **Right:** The model determines which textual spans are visually grounded, shown in green, then outputs the spans that are *unrelated* to the image, shown in yellow.

Motivated by this, we introduce a new task called ***Unrelatedness***. The task encourages models to (i) parse the information available from all modalities in the input, (ii) leverage multimodal alignment in a shared embedding space, as commonly operationalized in contrastive vision language pretraining (Radford et al., 2021), and (iii) identify which parts of the text are *not* grounded in the image. Concretely, given an image and an article-length text, the model must highlight the textual spans that remain unsupported after accounting for the visual content, rather than merely selecting a best matching caption. This makes *Unrelatedness* a stress test for context selection under long discourse, and it targets a capability that is especially consequential for multilingual agents operating in the wild, where irrelevant or misleading text can co-occur with a correct image.

E Detailed Table 1 from Page 2

The Table 13 contains the complete version of the VLM outputs which we truncated in Table §1 in section §1 due to space limitations.

F Experiments with *VLURes* Japanese Data

Before fine-tuning (Table 14). Across all eight tasks, switching the *input* text from English to Japanese introduces a consistent degradation in *English* outputs for the strongest closed models (GPT-4o/mini and Gemini Flash variants), with Δ_{En} typically clustered around ≈3–5 points in both zero-shot and one-shot settings (e.g., GPT-4o shows Δ_{En} between roughly -3.5 and -5.0 across tasks). Importantly, this drop is *task-agnostic*: object-level (OR), relational (RU), dense prediction (SS), and language-heavy tasks (IC, VQA) all decline by similar margins, suggesting that the bottleneck is not a single capability (e.g., captioning) but rather the model's overall robustness to Japanese *prompt conditioning* in the vision-language interface. In contrast, open models exhibit a much sharper split. LLaVA-Mistral (7B/13B) suffers severe collapses under Japanese input, with large negative deltas (often ≈10–17 points) and very low absolute accuracies, indicating weak multilingual vision-language alignment in this configuration. Qwen2VL and PALO are noticeably more stable: while their *English* outputs still drop under Japanese input (moderate Δ_{En} in the ≈3–4 range), their *Japanese* outputs are consistently higher in absolute accuracy than their English outputs for the same Japanese inputs, indicating that *matching the response language to the input language* can partially compensate for cross-lingual generation difficulty. Overall, Table 14 highlights two pre-finetuning failure modes for open VLMs: (i) strong sensitivity to non-English inputs (LLaVA), and (ii) a more moderate input-language penalty that is amplified when the model must produce *English* outputs from Japanese inputs (Qwen2VL/PALO). Finally, the human row indicates a large remaining headroom across tasks, especially relative to open models, emphasizing that Japanese VL evaluation remains far from saturated.

After fine-tuning on Japanese (Table 15). Fine-tuning changes the picture substantially for open VLMs, improving *absolute* performance broadly across tasks and reducing cross-lingual sensitivity. The clearest gains appear for Qwen2VL, PALO, MAYA, and LLaVA-13B: their accuracies rise across most tasks (often by tens of points compared to the pre-finetuning table), and the English-output gap induced by Japanese inputs shrinks to near-zero in many cells (i.e., Δ_{En} is frequently close to 0 or within about 1 point), indicating that fine-tuning successfully teaches the model to use Japanese prompts without sacrificing English answering ability. Task-wise, the improvements are again broad rather than isolated: OR/SU/RU/SS increase alongside IC/IT-M/U/VQA, implying that fine-tuning strengthens the shared vision-language grounding rather than only surface-level language generation. Language-wise, the main residual pattern is that *same-language generation* (Ja input → Ja output) tends to remain easier than forcing *cross-language* generation (Ja input → En output), especially for models that were already multilingual-leaning (e.g., Qwen2VL/PALO). Meanwhile, LLaVA-7B remains an outlier: although fine-tuning increases its absolute scores, it still shows larger input-language penalties than the other open models, suggesting that limited capacity and/or weaker base multilingual alignment makes it harder to fully internalize robust Japanese VL instruction-following. Taken together, Tables 14 and 15 show that Japanese fine-tuning largely converts the Japanese-input setting from a *systematic degradation* regime into a *near-parity* regime for strong open VLMs, while also preserving (and often improving) Japanese-output performance. This supports the view that the dominant issue pre-finetuning is insufficient Japanese coverage in the vision-language alignment and instruction tuning, and that relatively small but targeted Japanese finetuning data can significantly close the gap across diverse VL skills.

G Experiments with *VLURes* Swahili Data

Swahili input, before fine-tuning (Table 16). Table 16 evaluates a challenging multilingual setting where the *input* text is Swahili and the VLM must answer in either English (*Sw→En*) or Swahili (*Sw→Sw*) across eight vision-language tasks. Overall, the strongest proprietary models (GPT-4o/mini and Gemini Flash variants) remain robust under Swahili prompts, but they exhibit a consistent *input-language penalty* for English outputs: Δ_{En} is almost always negative and typically around 4–5 points across tasks and prompting regimes (zero-shot/one-shot, with/without rationales). This pattern is largely *task-agnostic*, indicating that the main failure mode is not tied to a single capability (e.g., captioning vs. recognition), but rather to reduced effectiveness of vision-language conditioning when the prompt language shifts from English to Swahili.

Task-wise and language-wise trends. The Swahili-output columns (*Sw→Sw*) show substantially larger variance than the English-output columns. In easier perceptual tasks (e.g., Object Recognition and Scene Understanding), Swahili responses can be relatively stable and sometimes nearly match (or slightly exceed) the corresponding English-output accuracy (e.g., several small positive or near-zero Δ_{Sw} values for OR). In contrast, language-heavy tasks, especially Image Captioning (IC), and in some cases Relation Understanding (RU) and Semantic Segmentation (SS), suffer larger drops in Swahili, with Δ_{Sw} frequently more negative than Δ_{En}. This gap is most pronounced in the one-shot setting without rationales, where Swahili captioning quality degrades sharply for some models (e.g., GPT-4o shows a particularly large negative Δ_{Sw} on IC). Adding rationales generally stabilizes Swahili generation and reduces extreme degradations in some cases, but it does not eliminate the overall cross-lingual penalty relative to English-input baselines. Finally, the human performance row underscores substantial headroom for Swahili VL evaluation, especially in tasks where both models and humans find the task difficult (e.g., VQA), while humans remain near-ceiling in several other tasks, highlighting that the remaining gap is primarily a modeling limitation rather than dataset ambiguity.

H Experiments with *VLURes* Urdu Data

Urdu input, before and after fine-tuning (Tables 17, 18). Tables 17 and 18 evaluate a bilingual setting where the *input* text is Urdu and the VLM produces either English (*Ur→En*) or Urdu (*Ur→Ur*) answers across eight VL tasks. Before fine-tuning (Table 17), strong proprietary models (GPT-4o/mini and Gemini Flash variants) remain highly robust under Urdu prompts: the English-output deltas are small (typically around $\Delta_{\mathrm{En}} \approx -1$ to -2), indicating only a mild cross-lingual prompt penalty relative to English-input baselines. For Urdu outputs, these models are also stable and in several tasks even *improve* with Urdu prompting (positive Δ_{Ur} values appear more frequently than in the Swahili setting), especially under one-shot and/or rationale prompting, suggesting that Urdu is comparatively better supported for instruction-following and generation. In contrast, the open-source models (PALO/-MAYA) show limited Urdu competence before fine-tuning: they achieve much lower absolute accuracies and exhibit large negative Δ_{Ur} (often -5 to -15 or worse), with the largest degradations occurring in language-heavy tasks such as Image Captioning and in more compositional tasks such as Relation Understanding. Consistent with these observations, we exclude Qwen2VL 7B and LLaVa Mistral 7B/13B due to insufficient Urdu support in initial runs.

After fine-tuning (Table 18), we observe that open models become noticeably more usable for Urdu, but the gains are uneven across architectures and prompting styles. PALO 7B benefits most clearly: fine-tuning combined with rationales yields consistent improvements in *Ur→Ur* performance across tasks (often positive Δ_{Ur} on the order of $\sim$4–6 points), while *Ur→En* performance remains near-parity or only slightly reduced, indicating improved multilingual grounding rather than a simple tradeoff between languages. MAYA 8B shows smaller and less consistent changes, frequently remaining close to baseline with residual weaknesses in language-intensive tasks. PALO 13B remains the least stable: despite some recovery in absolute Urdu scores, it exhibits mixed deltas across tasks (and sensitivity on Unrelatedness), suggesting that larger open models may require stronger multilingual pretraining and/or more careful fine-tuning to avoid brittle behavior under Urdu prompts. Overall, fine-tuning narrows the Urdu gap for open models, but a substantial margin remains to proprietary systems and human performance, highlighting that robust *Ur→Ur* generation depends heavily on both pretraining coverage and language-aware alignment.

I Comparison with MaRVL Dataset

The MaRVL dataset (Liu et al., 2021) contains image-text pairs in five languages: Indonesian (Id), Mandarin Chinese (Zh), Swahili (Sw), Tamil (Ta), and Turkish (Tr). We collect the Sw subset of images from this dataset and the captions accompanying those images.[3] Because each 'text' in MaRVL contains several associated images, we deploy CLIP to align the text with the most relevant caption. Hence, we use CLIP-aligned image-text pairs for all the analysis in this section, and there are 78 image-text pairs. We provide input prompts to the VLM in Sw, and the VLM generates responses in Sw. The results are shown in the Table 19. Under zero-shot settings, the best accuracy previously reported was **55.5%**, achieved by xUNITER, a variant of the UNITER

[3]MaRVL dataset is available at the URL https://marvl-challenge.github.io/download.

(Chen et al., 2019) model. However, we observe a dramatic increase in the accuracy on Sw, demonstrating the impressive abilities of VLMs. The GPT-4o model achieved the highest accuracy, per task. In §6, we introduced the best results with MaRVL under the *one-shot with rationale* setting. However, we present the results under all four settings considered in this work in Table 19. We do not include LlaVa Mistral 7B, PALO 7B, MAYA 7B, LlaVa Mistral 13B, PALO 13B, because they showed little understanding of Sw in our initial experiments. Table 19 shows that GPT-4o achieves higher accuracy on MaRVL than on *VLURes* for Swahili. These results support our claim that *VLURes* provides a more challenging evaluation setting for multilingual VLMs, especially for tasks that require selecting relevant evidence from long-form text. **Takeaway message: *VLURes* is harder than MaRVL because it requires long-text grounding and filtering.**

J Translation Baseline (diagnostic)

To separate translation difficulty from genuine multilingual grounding, we ran a translate-then-evaluate baseline on 200 instances (50 per language) using Google Translate to convert non-English articles into English before prompting GPT-4o under *zero-shot without rationales*. On image-only tasks, translation recovered 2.1 points of the cross-lingual gap on average; on grounding tasks (ITM, U, VQA) recovery was only 0.7 points. This confirms that the dominant source of cross-lingual degradation in *VLURes* is **not** translation difficulty, but rather reflects genuine multilingual VL grounding capability, motivating native-language evaluation rather than translation-mediated proxies.

Task	Prompt	Response from Vision Language Model
1. Object Recognition (OR)	Analyze this image and list all objects present. Categorize each object into groups such as furniture, electronic devices, clothing, etc. Be thorough and specific in your identification.	Image Analysis: Objects Present: 1. Infrastructure: - Bridge - Railings - Streetlights - Road markings 2. Buildings: - High-rise buildings - Residential or commercial structures 3. People: - Individuals walking or cycling 4. Transportation: - Bicycles. Categories: Infrastructure: Bridge, railings, streetlights, road markings. Buildings: High-rise buildings. People: Individuals. Transportation: Bicycles.
2. Scene Understanding (SU)	Describe the overall scene in this image. What is the setting, and what activities or events are taking place? Provide a comprehensive overview of the environment and any actions occurring.	Image Analysis: The scene depicts an urban setting with a modern bridge leading towards a cityscape of tall buildings. The bridge has a dedicated pathway for pedestrians and cyclists, as indicated by the markings on the ground. There are a few people walking and cycling along this path, suggesting it is a popular route for commuting or leisure activities...... The bridge is flanked by metal railings and streetlights, and it appears to span over a body of water, visible on the right side. The sky is mostly clear with some clouds, indicating a pleasant day. The overall atmosphere is calm and orderly, typical of a well-maintained urban environment.
3. Relationship Understanding (RU)	Identify any interactions or relationships between objects or entities in this image. How are they related or interacting with each other? Explain any spatial, functional, or social connections you observe.	Image Analysis: Interactions and Relationships. 1. Spatial Relationships: - The image shows a bridge or elevated pathway leading towards a cityscape with several tall buildings in the background. The pathway is divided into lanes, likely for pedestrians and cyclists. - The bridge is flanked by railings on both sides, providing safety and separation from the surrounding environment. 2. Functional Connections: - The bridge serves as a connection between different parts of the city, facilitating movement for pedestrians and cyclists. It likely connects residential or commercial areas with the city center. - The presence of streetlights along the bridge suggests it is used during nighttime as well, enhancing safety and visibility. 3. Social Interactions: - There are people walking and cycling on the bridge, indicating it is a shared space for different modes of non-motorized transport. - The design of the bridge promotes social interaction by providing a shared space for pedestrians and cyclists, encouraging a community atmosphere. Overall, the image depicts a well-integrated urban infrastructure that supports efficient and safe movement within the city, fostering both functional and social interactions.
4. Semantic Segmentation (SS)	Divide this image into different semantic regions. Label each region (e.g., sky, buildings, people, street) and briefly describe its contents. Provide a clear breakdown of the image's composition.	Image Analysis: Semantic Regions. 1. Sky: - Location: Top portion of the image. - Description: The sky is mostly clear with some scattered clouds, providing a bright backdrop to the scene. 2. Buildings: - Location: Middle section of the image, in the background. - Description: A series of modern high-rise buildings with glass facades. They vary in height and design, contributing to an urban skyline. 3. Bridge/Pathway: Location: Foreground, extending from the bottom to the middle of the image - Description: A wide pedestrian and cycling pathway on a bridge. It is flanked by metal railings on both sides, with a clear division for different lanes. 4. People: Location: Middle of the pathway - Description: Several individuals are walking or cycling along the pathway, indicating active use of the bridge. 5. Street Lamps: - Location: Left side of the bridge - Description: Modern street lamps are positioned along the bridge, providing lighting for the pathway. 6. Water: - Location: Right side, below the bridge - Description: A body of water is visible beneath the bridge, suggesting the bridge spans over a river or canal. This breakdown highlights the urban setting and infrastructure, emphasizing the interaction between architecture and public spaces.
5. Image Captioning (IC)	Provide a detailed, natural language description of what is happening in this image. Narrate the scene as if you were explaining it to someone who cannot see it, including all relevant details and actions.	Image Analysis: The image depicts a modern urban scene featuring a pedestrian and cycling path on a bridge. The path is divided by a dashed white line, indicating separate lanes for cyclists and pedestrians. The bridge is flanked by metal railings on both sides, providing safety for those using the path. In the background, a skyline of tall, contemporary buildings is visible, suggesting a city setting. These buildings are mostly glass and steel, reflecting a modern architectural style. The sky above is partly cloudy, with patches of blue visible, indicating a clear day. On the path, a few people are visible. One person is cycling towards the city, while another is walking. The bridge appears to span over a body of water, as suggested by the glimpse of water visible through the railings on the right side. Overall, the scene conveys a sense of urban life, with infrastructure designed to accommodate both pedestrians and cyclists, set against the backdrop of a bustling cityscape.
6. Image-Text Matching (ITM)	Extract and list the specific parts of the text that closely match or directly reference entities, objects, or scenes depicted in the image. Be precise in identifying these connections and explain the visual evidence that supports each textual reference.	The image depicts a modern urban area with high-rise buildings, which aligns with the description of Wentworth Point in the text. Here are the specific connections between the text and the image: 1. Wentworth Point: The text mentions that the incident occurred in Wentworth Point, a suburb with high-rise buildings. The image shows a skyline with several tall buildings, which is characteristic of urban areas like Wentworth Point. 2. Apartment Buildings: The text references Musa attempting to enter an apartment building. The image shows multiple high-rise residential buildings, which could represent the type of apartment complex mentioned. 3. Urban Setting: The image's urban setting with modern infrastructure supports the text's context of a suburban area with secured garages and apartment complexes. These elements in the image visually support textual references to locations and settings of events described.
7. Unrelatedness (U)	Identify which parts of the text are not relevant to or not represented in the image. Explain why these elements are unrelated by describing what is missing in the image that would be needed to illustrate these textual elements.	The image shows a view of a bridge or walkway leading towards a cluster of modern buildings, likely in an urban area. Here's how the text relates to the image: 1. Location Context: - The text mentions Wentworth Point, which is likely the area depicted in the image. The buildings in the image could be part of this suburb. 2. Unrelated Text Elements: - Specific Events: The text describes a specific incident involving a killing, a text message, and interactions between individuals. None of these events or interactions are depicted in the image. - Individuals and Vehicles: The text mentions specific people, a Toyota Kluger SUV, and a garage. The image does not show any people, vehicles, or a garage. - Interior Details: The text includes details about entering a building and a damaged roller door. The image does not show any building interiors or specific architectural features like a roller door. 3. Missing Elements for Illustration: - To illustrate the textual elements, the image would need to show the specific individuals involved, the Toyota Kluger, the garage, or the interior of the building mentioned in the text. Overall, while the image provides a general sense of the location, it does not depict the specific events or details described in the text.
8. Visual Question Answering (VQA)	What places are mentioned in the text or shown in the image? For each place identified, indicate whether it appears in the text, the image, or both. If any of these places are famous or well-known locations, explain why they are significant.	The text mentions the following places: 1. Wentworth Point: This is where the incident occurred, specifically at a garage along Hill Road and near Half Street. It is a suburb in New South Wales, Australia. 2. Granville: Musa searched for Thagipur in several cafés and restaurants in this area. It is another suburb in New South Wales, Australia. 3. Toyota Kluger: Mentioned as the type of vehicle Musa was driving. The image shows a modern urban area with high-rise buildings, which could be representative of a suburb like Wentworth Point. However, without specific landmarks or signs, it's not possible to definitively identify the location in the image as Wentworth Point or any other specific place mentioned in the text. Significance: - Wentworth Point: Known for its residential developments and proximity to Sydney Olympic Park, making it a notable suburb in Sydney. - Granville: Known for its diverse community and historical significance in Sydney. In summary, Wentworth Point and Granville are mentioned in the text, while the image could potentially depict a similar urban environment but does not explicitly show any identifiable location from the text.

Table 13: *VLU-RES* includes eight tasks , and prompts specific to each task needed to generate responses from VLM.

Model	Object Recognition				Scene Understanding				Relation Understanding				Semantic Segmentation				Image Captioning				Image-Text Matching				Unrelatedness				Visual Question Answering			
	En	Δ_{En}	Ja	Δ_{Ja}	En	Δ_{En}	Ja	Δ_{Ja}	En	Δ_{En}	Ja	Δ_{Ja}	En	Δ_{En}	Ja	Δ_{Ja}	En	Δ_{En}	Ja	Δ_{Ja}	En	Δ_{En}	Ja	Δ_{Ja}	En	Δ_{En}	Ja	Δ_{Ja}	En	Δ_{En}	Ja	Δ_{Ja}
Zero-shot, *Without Rationales*																																
GPT-4o	85.7	-4.1	84.8	-3.7	85.2	-3.8	84.6	-3.4	84.4	-4.5	80.5	-3.8	82.1	-5.0	77.3	-3.5	81.2	-3.9	75.6	-3.3	85.5	-3.5	85.6	-3.0	87.5	-3.8	87.4	-3.5	85.6	-4.0	85.9	-3.0
GPT-4o-mini	76.1	-3.9	79.8	-3.0	76.4	-4.1	76.4	-3.4	75.4	-3.5	71.8	-3.5	72.7	-3.2	69.4	-4.0	74.4	-3.9	68.4	-3.3	73.5	-3.3	73.9	-3.3	79.1	-3.8	78.5	-3.5	76.7	-3.5	77.9	-3.5
Gemini 2.0 Flash Lite	79.9	-3.8	83.0	-3.5	80.4	-3.4	80.1	-3.4	78.8	-3.7	75.5	-4.0	75.9	-2.7	73.1	-3.7	78.0	-3.9	72.1	-3.9	76.9	-3.0	77.6	-3.9	82.8	-3.7	82.2	-3.8	81.0	-4.1	81.6	-3.0
Gemini 1.5 Flash 8B	74.7	-3.5	77.8	-3.3	74.6	-3.8	74.7	-3.4	73.0	-4.2	70.1	-6.5	71.1	-3.3	67.7	-3.8	72.7	-3.9	66.7	-4.0	72.6	-4.1	72.2	-3.2	77.4	-3.7	76.8	-3.9	75.5	-4.0	76.2	-3.0
LlaVa Mistral 7B	24.3	-14.3	31.4	-10.0	25.1	-13.5	29.5	-8.3	25.5	-11.8	24.9	-5.3	21.7	-13.9	22.5	-8.6	23.0	-13.6	21.5	-9.2	22.0	-13.8	27.0	-8.5	28.0	-14.1	31.6	-8.5	25.1	-13.9	30.9	-8.7
Qwen2VL 7B	56.5	-3.5	64.2	1.5	56.3	-4.6	60.9	1.9	54.9	-3.1	56.3	3.2	53.7	-4.1	53.9	1.9	54.4	-4.0	52.9	2.4	53.9	-4.0	58.4	3.0	59.4	-4.2	63.0	2.3	57.0	-4.2	62.5	2.1
PALO 7B	35.8	-3.7	42.1	1.2	36.2	-3.4	38.9	1.1	35.3	-3.0	34.3	2.7	33.0	-4.0	31.9	1.7	33.9	-3.9	30.9	0.9	33.8	-4.1	36.4	2.3	38.7	-4.0	41.0	0.9	37.0	-4.3	40.4	2.3
MAYA 8B	36.8	-4.0	42.5	-3.7	37.2	-3.7	39.7	-3.9	36.8	-2.8	35.1	-4.0	34.3	-4.0	32.7	-3.3	35.2	-4.1	31.7	-2.9	35.0	-4.0	37.2	-3.3	40.1	-4.1	41.8	-4.1	38.1	-4.1	41.2	-4.2
LlaVa Mistral 13B	30.6	-16.3	32.4	-15.4	31.0	-15.9	30.5	-15.9	28.5	-17.0	25.9	-18.1	26.8	-15.0	23.5	-15.9	29.0	-16.4	22.5	-15.6	28.2	-15.8	28.0	-15.8	33.7	-16.4	32.6	-15.2	32.0	-16.5	31.9	-17.0
PALO 13B	37.2	-3.8	40.0	-3.4	37.3	-3.8	36.9	-3.5	35.4	-4.4	32.3	-2.7	34.2	-4.0	29.9	-3.8	35.4	-3.6	28.9	-3.8	35.0	-4.0	34.4	-3.7	40.1	-4.2	39.0	-3.4	38.4	-4.7	38.5	-4.1
Zero-shot, *With Rationales*																																
GPT-4o	85.3	-3.8	83.3	-3.2	85.7	-4.0	84.3	-3.7	84.6	-3.4	82.5	-2.5	84.2	-3.7	82.8	-2.9	82.6	-4.2	78.8	-3.6	85.4	-3.7	83.3	-3.5	87.4	-3.8	87.3	-3.3	83.9	-3.5	83.0	-3.4
GPT-4o-mini	75.8	-4.3	79.9	-3.2	75.9	-3.7	76.4	-3.6	75.9	-3.7	76.4	-2.9	75.2	-3.0	77.3	-2.48	75.4	-5.0	72.0	-2.9	75.2	-3.6	74.6	-3.6	79.2	-3.7	78.7	-3.2	71.8	-3.6	73.2	-3.9
Gemini 2.0 Flash Lite	79.8	-3.9	83.5	-3.2	79.8	-3.7	80.6	-3.0	79.6	-3.7	80.2	-3.0	78.9	-3.0	81.0	-2.9	78.2	-4.1	75.6	-3.4	78.9	-3.7	78.3	-3.6	82.9	-3.5	82.4	-3.8	75.6	-3.7	76.9	-3.0
Gemini 1.5 Flash 8B	74.6	-3.7	77.9	-3.4	74.4	-3.7	75.2	-3.0	74.2	-3.7	75.0	-3.0	73.5	-3.0	75.6	-2.9	72.7	-4.0	70.3	-3.5	73.4	-3.5	72.9	-3.0	77.5	-3.9	77.0	-3.04	70.3	-3.8	71.5	-3.0
LlaVa Mistral 7B	24.9	-14.5	32.1	-9.8	24.7	-13.6	28.6	-9.8	25.1	-13.0	29.3	-10.3	24.1	-12.5	30.5	-9.0	22.3	-12.9	25.6	-9.4	24.5	-14.1	27.9	-9.2	28.4	-14.0	32.5	-9.2	21.1	-14.0	26.3	-9.3
Qwen2VL 7B	56.3	-3.7	64.9	1.8	55.0	-4.0	61.6	1.7	56.2	-3.5	60.8	2.7	55.5	-3.0	61.5	1.7	55.2	-4.3	56.9	2.0	55.0	-3.5	59.6	1.6	58.7	-3.5	63.5	2.3	52.8	-4.0	58.0	2.7
PALO 7B	35.9	-3.6	42.7	1.4	34.4	-5.0	39.3	2.1	35.4	-3.7	39.0	2.9	36.0	-4.3	39.8	2.0	34.2	-4.2	34.4	1.7	34.5	-3.4	37.1	1.7	38.5	-3.7	41.2	1.7	31.7	-4.0	35.7	1.9
MAYA 8B	37.1	-4.2	42.4	-3.3	36.8	-3.8	39.9	-3.8	36.9	-3.5	39.6	-3.3	37.1	-4.1	40.6	-4.1	35.2	-4.0	35.3	-2.9	36.0	-3.6	37.9	-3.6	39.8	-3.7	42.0	-3.6	33.2	-4.2	36.5	-3.5
LlaVa Mistral 13B	31.2	-16.0	32.9	-17.2	29.4	-16.5	31.8	-14.3	29.3	-17.3	30.4	-17.0	30.7	-15.9	31.3	-15.8	29.5	-16.6	25.7	-15.6	29.9	-15.9	28.8	-15.8	32.9	-15.4	33.0	-15.9	26.3	-15.7	26.9	-17.0
PALO 13B	37.5	-3.5	40.2	-3.8	36.3	-4.2	36.5	-4.2	37.0	-3.6	36.8	-2.7	37.8	-4.2	37.6	-3.7	35.5	-4.0	32.5	-3.5	36.9	-4.3	35.6	-3.7	40.9	-4.4	39.7	-4.0	33.1	-3.6	33.8	-3.2
One-shot, *Without Rationales*																																
GPT-4o	86.4	-4.1	83.8	-3.0	86.7	-3.8	85.1	-3.0	85.9	-3.5	84.0	-3.0	85.4	-4.1	81.9	-3.3	84.4	-4.2	75.5	-3.6	84.7	-3.4	83.4	-4.0	87.2	-4.0	87.1	-3.7	85.1	-3.7	83.6	-3.9
GPT-4o-mini	77.8	-4.2	75.9	-3.5	78.6	-3.5	76.0	-3.2	76.1	-3.5	74.7	-3.0	77.9	-4.2	74.5	-3.2	76.2	-4.0	69.6	-3.0	75.7	-3.3	73.5	-3.5	79.2	-4.0	79.1	-3.7	77.7	-3.8	75.5	-3.5
Gemini 2.0 Flash Lite	81.9	-3.8	79.7	-3.2	82.5	-3.3	79.2	-3.7	79.8	-3.5	78.4	-2.0	81.4	-4.0	78.2	-3.5	80.0	-4.1	73.3	-3.0	79.1	-3.0	77.2	-3.6	82.6	-3.7	82.8	-4.2	81.6	-4.0	79.2	-3.5
Gemini 1.5 Flash 8B	76.0	-4.5	74.5	-3.1	76.6	-3.8	73.7	-3.8	73.2	-3.1	73.1	-4.7	77.1	-5.1	72.8	-3.9	74.5	-4.0	67.9	-3.7	74.9	-4.2	71.8	-3.7	77.1	-3.6	77.4	-3.5	76.3	-4.1	73.8	-3.8
LlaVa Mistral 7B	25.8	-14.8	30.0	-7.8	27.6	-13.7	28.9	-9.0	24.5	-13.5	27.7	-5.9	26.2	-13.8	27.2	-9.3	25.8	-14.1	23.5	-8.4	25.5	-14.1	26.6	-9.5	28.4	-14.0	32.4	-8.8	26.4	-14.0	27.9	-9.0
Qwen2VL 7B	58.0	-4.0	60.5	1.4	58.6	-3.5	64.0	4.0	56.0	-3.6	59.2	2.0	57.9	-4.0	59.0	2.4	56.1	-3.6	54.1	2.0	56.5	-4.0	57.6	2.0	60.0	-4.1	63.6	1.0	57.3	-4.0	59.6	2.1
PALO 7B	38.0	-3.5	38.8	2.0	38.1	-3.5	42.2	5.1	35.4	-3.7	37.1	2.3	37.2	-4.0	36.8	2.7	35.4	-3.7	32.5	2.4	36.0	-4.1	35.9	2.4	38.8	-4.1	41.6	2.4	37.4	-4.0	38.0	2.0
MAYA 8B	39.0	-4.0	39.1	-3.8	40.3	-3.5	40.1	-2.9	36.7	-3.7	37.9	-2.4	38.8	-4.3	37.9	-3.7	36.4	-3.4	32.9	-3.5	37.2	-4.0	36.8	-3.7	40.0	-4.0	42.4	-3.4	38.4	-3.7	38.8	-4.0
LlaVa Mistral 13B	32.8	-16.4	30.2	-15.7	34.2	-15.0	36.3	-10.0	31.1	-15.2	29.2	-14.0	31.6	-15.7	28.4	-15.5	30.6	-15.8	24.1	-15.7	30.8	-15.8	27.2	-15.4	33.6	-15.7	33.1	-16.3	31.4	-15.9	29.2	-15.9
PALO 13B	41.0	-3.7	36.3	-3.4	39.6	-4.1	36.0	-4.2	37.6	-3.0	35.2	-3.1	38.0	-3.4	34.6	-4.3	37.3	-4.0	30.1	-4.3	37.6	-4.2	33.7	-3.3	40.4	-3.6	39.8	-4.1	38.6	-4.0	35.8	-4.3
One-shot, *With Rationales*																																
GPT-4o	**86.8**	-3.7	**86.4**	-3.4	**86.8**	-3.9	**85.4**	-3.2	**87.0**	-4.4	**85.7**	-4.1	**86.6**	-3.6	**85.7**	-3.8	**84.5**	-3.8	**81.5**	-3.7	**86.4**	-3.9	**84.9**	-3.2	**87.7**	-3.8	**87.9**	-3.6	**86.8**	-3.8	**87.3**	-3.7
GPT-4o-mini	78.5	-4.0	79.9	-3.5	78.7	-3.6	77.2	-3.0	78.0	-3.6	78.4	-4.3	78.5	-3.7	77.1	-3.0	76.2	-3.7	73.2	-3.6	77.8	-3.7	73.4	-3.4	79.2	-3.7	77.7	-3.7	78.7	-3.5	76.4	-3.9
Gemini 2.0 Flash Lite	82.1	-4.1	83.6	-3.0	81.9	-4.0	80.7	-3.1	81.7	-3.4	82.1	-4.0	82.1	-3.6	80.8	-3.0	80.1	-3.9	76.9	-3.8	81.5	-3.7	77.1	-3.4	82.8	-3.6	81.4	-3.7	82.4	-3.7	80.1	-3.8
Gemini 1.5 Flash 8B	77.1	-3.7	78.2	-3.0	76.7	-3.8	75.3	-3.2	76.9	-3.0	76.7	-3.8	76.8	-3.7	75.4	-3.0	74.8	-4.0	71.5	-3.5	76.1	-3.7	71.7	-3.4	77.5	-3.7	76.0	-3.0	77.0	-3.8	74.7	-3.5
LlaVa Mistral 7B	25.8	-14.8	33.6	-8.5	27.8	-12.7	29.8	-9.1	26.0	-13.5	31.4	-7.4	25.6	-12.7	30.4	-9.2	26.2	-14.4	25.7	-9.3	25.8	-13.2	26.2	-8.8	28.0	-14.1	31.2	-9.1	27.3	-13.5	29.3	-8.3
Qwen2VL 7B	57.6	-4.9	65.1	1.5	58.4	-3.8	63.8	3.5	57.2	-3.8	62.7	2.4	58.0	-3.2	61.4	1.9	57.0	-4.1	57.5	0.8	56.9	-3.4	57.2	1.1	60.0	-4.3	62.4	2.2	58.6	-3.0	60.3	2.0
PALO 7B	38.9	-3.2	42.8	1.0	38.0	-3.7	43.1	4.0	37.4	-3.7	40.3	3.0	38.3	-4.0	39.6	0.8	36.0	-4.0	35.7	1.5	37.0	-3.4	35.9	1.1	39.0	-4.0	40.2	0.7	38.2	-3.5	38.9	1.8
MAYA 8B	39.8	-3.5	43.9	-3.7	39.3	-3.7	40.2	-3.8	38.7	-3.7	41.7	-3.7	39.3	-3.7	40.4	-4.1	36.6	-3.3	36.5	-3.6	38.3	-3.4	36.6	-3.9	39.8	-3.5	41.0	-3.4	39.5	-3.97	39.7	-4.2
LlaVa Mistral 13B	33.5	-16.0	37.2	-13.2	35.6	-14.5	32.4	-15.0	32.6	-15.2	32.6	-14.3	35.2	-18.0	31.6	-17.0	29.7	-14.7	27.4	-15.4	32.0	-15.9	27.3	-16.2	33.5	-15.5	32.1	-16.4	26.8	-16.0	29.8	-15.8
PALO 13B	39.7	-3.8	41.0	-3.0	39.7	-3.6	37.7	-3.6	38.2	-4.3	38.8	-4.0	39.4	-3.6	37.4	-3.7	37.5	-4.0	33.6	-3.5	38.1	-3.6	33.1	-3.9	40.3	-4.1	39.0	-4.0	39.1	-3.8	36.9	-4.2
Human Performance	–	–	96.8		–	–	94.7		–	–	95.2		–	–	87.1		–	–	94.8		–	–	89.5		–	–	88.6		–	–	85.9	–

Table 14: Performance of VLMs on **eight VL tasks** under zero-shot and one-shot settings, measured by Accuracy (%). **Input:** Japanese Texts + Images; **Output:** En, Ja responses. Shaded columns represent {Ja} in input and {Ja} in output VLM results. En ΔAcc. = {En score from Table 5 (En-Input) – En score from this Table (Ja-Input)}. Ja ΔAcc. = {Ja score from Table 5 (En-Input) – Ja score from this Table (Ja-Input).} Positive Δ in blue, negative Δ in red. Zeros indicate unintelligible responses or no support language.

Model	Object Recognition				Scene Understanding				Relation Understanding				Semantic Segmentation				Image Captioning				Image-Text Matching				Unrelatedness				Visual Question Answering			
	En	Δ_{En}	Ja	Δ_{Ja}	En	Δ_{En}	Ja	Δ_{Ja}	En	Δ_{En}	Ja	Δ_{Ja}	En	Δ_{En}	Ja	Δ_{Ja}	En	Δ_{En}	Ja	Δ_{Ja}	En	Δ_{En}	Ja	Δ_{Ja}	En	Δ_{En}	Ja	Δ_{Ja}	En	Δ_{En}	Ja	Δ_{Ja}
Zero-shot, *With Rationales*																																
LlaVa Mistral 7B	37.4	-12.1	44.9	-6.5	37.4	-11.7	42.0	-6.5	35.6	-11.2	37.4	-6.8	34.6	-12.1	35.0	-6.4	35.6	-11.5	34.0	-6.8	35.5	-12.6	39.5	-7.6	40.4	-11.8	44.1	-6.9	38.4	-12.5	43.4	-6.7
Qwen2VL 7B	68.6	0.6	76.4	5.1	69.7	-2.0	73.4	5.8	67.3	0.7	68.8	5.1	66.1	-0.3	66.4	5.5	67.1	-0.5	65.4	5.4	66.7	-0.6	70.9	4.9	71.9	-0.5	75.5	5.3	69.8	-0.8	75.0	5.0
PALO 7B	48.6	-0.7	54.4	5.2	48.6	-0.8	51.4	5.2	47.2	-0.5	46.8	5.8	45.7	-0.5	44.4	5.1	46.7	-0.5	43.4	5.0	46.2	-0.3	48.9	5.2	51.0	-7.3	53.5	2.4	49.3	-0.4	52.9	5.2
MAYA 8B	49.5	-0.3	55.2	-0.4	49.7	-0.4	52.4	0.3	48.3	-0.5	47.6	-0.1	46.6	-0.1	45.2	-0.3	47.8	-0.5	44.2	0.5	47.3	-0.1	49.7	2.4	49.8	-8.7	54.3	2.5	50.7	-0.5	53.7	-0.5
LlaVa Mistral 13B	45.8	-8.4	48.9	-8.5	46.1	-8.3	46.0	-8.0	44.7	-8.1	41.4	-8.6	43.8	-8.9	39.0	-8.1	44.4	-8.7	38.0	-8.2	43.9	-8.4	43.5	-8.2	49.3	-8.9	48.1	-8.8	47.5	-8.9	47.4	-9.1
PALO 13B	49.7	0.4	52.5	0.2	50.0	-0.5	49.4	0.4	48.7	-0.5	44.8	0.4	47.0	-0.6	42.4	-0.5	47.6	0.4	41.4	0.4	47.9	-0.7	46.9	0.3	52.9	-13.0	51.5	-0.4	50.8	-0.9	51.0	-0.3
Zero-shot, *With Rationales*																																
LlaVa Mistral 7B	37.4	-12.9	45.6	-6.3	37.2	-11.8	42.1	-6.4	36.6	-11.4	41.8	-6.5	37.6	-11.3	43.0	-6.3	35.4	-11.3	38.1	-6.8	36.5	-11.4	39.0	-7.5	41.1	-12.0	45.0	-6.9	33.9	-12.1	38.8	-6.7
Qwen2VL 7B	68.9	-0.3	76.7	4.5	67.7	-0.3	73.5	4.5	68.6	-0.5	73.3	4.2	69.2	-0.3	74.0	4.4	67.6	-0.5	69.4	4.5	68.2	-0.3	72.1	4.5	72.0	-0.6	76.0	4.5	65.3	-0.3	70.5	4.2
PALO 7B	48.1	-0.2	55.1	4.8	47.9	-0.2	50.8	4.8	47.7	-0.2	51.5	4.8	48.1	-0.2	52.3	4.8	46.8	-0.6	46.9	4.3	47.8	-0.5	49.6	5.1	51.4	-4.0	53.7	2.6	44.4	-0.5	48.2	5.2
MAYA 8B	49.4	-0.2	55.4	-0.2	49.2	-0.2	52.3	-0.2	49.0	-0.2	52.1	-0.2	49.4	-0.2	53.1	-0.2	47.7	-0.3	47.8	-0.4	49.1	-0.5	50.4	-0.5	52.9	-2.5	54.5	-0.1	45.6	-0.4	49.0	-0.4
LlaVa Mistral 13B	55.2	-0.3	58.0	0.0	53.8	0.0	53.9	0.0	54.8	-0.4	54.5	-0.2	55.1	0.0	55.4	0.0	53.3	0.0	49.9	0.0	54.4	0.0	53.0	0.0	57.9	0.0	57.1	-0.4	50.9	0.0	51.0	0.0
PALO 13B	49.5	-0.1	52.6	-0.3	49.4	-0.2	49.3	-0.2	49.3	-0.3	49.3	-0.6	50.1	-0.3	50.1	-0.5	48.1	-0.4	45.0	-0.5	49.1	-0.3	48.1	-0.4	52.5	1.3	52.2	-0.4	46.2	-0.3	46.3	-0.4
One-shot, *Without Rationales*																																
LlaVa Mistral 7B	45.3	-6.4	46.9	-1.5	45.8	-6.3	46.0	-2.0	43.1	-6.8	43.2	-3.2	45.0	-6.7	44.5	-2.1	42.8	-6.2	39.0	-3.8	42.5	-6.0	41.9	-4.1	46.5	-6.2	50.5	-2.0	44.2	-5.9	45.8	-1.8
Qwen2VL 7B	70.8	-0.6	72.7	5.1	70.9	-0.4	73.0	5.1	68.5	-0.5	71.7	5.1	70.6	-0.5	71.5	5.2	69.2	-0.5	66.6	5.2	69.2	-0.5	70.1	5.0	72.5	-0.4	76.1	5.0	69.9	-0.4	72.1	5.4
PALO 7B	50.1	-0.6	51.3	4.7	50.6	-0.6	50.8	4.4	47.8	-0.3	49.6	5.0	49.9	-0.5	49.3	5.3	48.3	-0.4	45.0	5.0	48.7	-0.6	48.4	4.2	51.3	-3.1	54.1	2.5	50.2	-0.6	50.5	5.1
MAYA 8B	51.6	-0.4	51.6	-0.4	51.8	-0.5	51.6	-0.1	49.3	-0.5	50.4	-0.4	51.4	-0.7	50.4	-0.3	49.7	-0.5	45.4	-0.4	50.0	-0.6	49.3	-0.1	52.9	-1.7	54.9	-0.7	51.6	-0.7	51.3	-0.4
LlaVa Mistral 13B	57.0	-0.2	54.1	-0.2	56.9	-0.2	54.0	-0.2	54.0	-0.2	53.1	-0.2	56.0	-0.2	52.3	-0.2	55.1	0.0	48.0	-0.4	55.4	-0.5	51.1	-0.5	58.4	-0.6	57.0	-0.4	55.8	-0.4	53.1	-0.5
PALO 13B	52.6	-0.5	48.8	-0.5	52.3	-0.6	48.8	-0.3	49.4	-0.4	47.7	-0.4	50.9	-0.1	47.1	-0.7	50.1	-0.6	42.6	-0.5	50.0	-0.4	46.2	-0.7	53.3	11.9	52.3	-0.7	51.3	-0.5	48.3	-0.3
One-shot, *With Rationales*																																
LlaVa Mistral 7B	44.7	-5.8	52.1	-0.7	44.7	-5.7	48.8	-0.5	43.7	-5.9	50.0	-0.4	44.5	-5.7	49.1	-0.7	43.5	-5.8	43.8	-2.0	44.8	-6.3	44.9	-0.7	46.3	-6.5	50.1	-0.4	45.1	-5.7	48.0	-0.7
Qwen2VL 7B	**71.3**	-0.4	**77.2**	5.3	**71.2**	-0.5	**74.0**	4.6	**70.1**	-0.7	**75.2**	4.8	**71.5**	-0.5	**74.0**	5.2	**69.5**	-0.4	**70.0**	5.0	**70.2**	-0.3	**69.7**	5.0	**72.5**	-0.6	**75.0**	3.7	**71.6**	-0.7	**72.8**	5.1
PALO 7B	50.8	-0.2	55.4	5.6	50.5	-0.4	51.7	5.1	50.0	-0.5	52.8	5.1	50.8	-0.3	52.1	4.6	48.6	-0.4	48.2	5.3	50.5	-0.7	48.4	4.9	51.3	-6.6	52.7	0.9	51.2	-0.7	51.4	4.8
MAYA 8B	52.5	-0.8	56.3	-0.8	51.7	-0.3	52.6	-0.3	51.7	-0.9	54.2	-0.9	52.7	-0.9	52.9	-0.8	50.3	-0.8	49.0	-0.9	52.0	-0.9	49.1	-0.9	52.7	-5.0	53.5	-0.7	52.7	-0.9	52.2	-0.3
LlaVa Mistral 13B	57.6	-0.6	58.1	-0.8	57.6	-0.7	55.1	-0.2	57.0	-0.9	56.5	-0.6	57.8	-0.7	55.5	-0.5	55.4	-0.5	51.3	-0.5	56.7	-0.7	51.2	-0.8	58.7	-0.8	56.0	-0.5	57.5	-0.8	53.7	-0.7
PALO 13B	52.3	-0.4	52.7	-0.6	52.0	-0.3	49.8	-0.6	51.3	-0.4	51.3	-0.5	52.4	-0.4	50.0	-0.5	50.4	-0.5	46.1	-0.4	51.3	-0.6	45.6	-0.3	52.2	-11.8	51.5	-0.6	51.5	-0.1	49.4	-0.5

Table 15: Performance of VLMs on **eight VL tasks** under finetuning settings, measured by estimated Accuracy (%) from LLM-as-a-judge. **Input:** Japanese Texts + Images; **Output:** En, Ja responses. Shaded columns represent {Ja} in input and {Ja} in output VLM results. En ΔAcc. = {En score from Table 6 (En-Input) – En score from this Table (Ja-Input)}. Ja ΔAcc. = {Ja score from Table 6 (En-Input) – Ja score from this Table (Ja-Input).} Positive Δ in blue, negative Δ in red. Zeros indicate unintelligible responses or no support language.

Model	Object Recognition				Scene Understanding				Relation Understanding				Semantic Segmentation				Image Captioning				Image-Text Matching				Unrelatedness				Visual Question Answering			
	En	Δ_{En}	Sw	Δ_{Sw}	En	Δ_{En}	Sw	Δ_{Sw}	En	Δ_{En}	Sw	Δ_{Sw}	En	Δ_{En}	Sw	Δ_{Sw}	En	Δ_{En}	Sw	Δ_{Sw}	En	Δ_{En}	Sw	Δ_{Sw}	En	Δ_{En}	Sw	Δ_{Sw}	En	Δ_{En}	Sw	Δ_{Sw}
Zero-shot, *Without Rationales*																																
GPT-4o	84.9	-4.5	84.3	-2.1	84.3	-4.7	83.8	-4.2	83.9	-4.6	79.7	-7.0	80.6	-5.1	76.5	-6.9	81.0	-5.3	74.8	-8.9	85.4	-5.0	84.8	-4.4	86.8	-4.8	86.6	-4.9	84.2	-4.2	85.1	-2.0
GPT-4o-mini	75.1	-4.5	78.6	0.6	75.9	-5.2	75.6	-3.2	74.5	-5.1	71.0	-5.2	73.0	-5.1	68.6	-5.4	73.1	-4.2	67.6	-6.1	73.6	-5.0	73.1	-7.2	78.8	-5.2	77.7	-4.1	76.6	-5.0	77.1	0.8
Gemini 2.0 Flash Lite	79.5	-5.2	82.3	0.5	79.8	-5.4	79.3	-3.1	78.2	-5.1	74.7	-5.0	76.6	-5.0	72.3	-5.0	76.7	-4.2	71.3	-5.3	77.6	-5.3	76.8	-7.1	82.3	-5.0	81.4	-4.2	79.8	-4.5	80.8	0.5
Gemini 1.5 Flash 8B	74.3	-5.4	76.9	0.5	74.2	-5.2	73.9	-3.1	72.9	-5.2	69.3	-5.3	71.3	-5.1	66.9	-5.3	72.2	-5.0	65.9	-6.1	71.2	-4.3	71.4	-7.0	77.4	-5.5	76.0	-4.0	75.0	-5.1	75.4	0.6
Zero-shot, *With Rationales*																																
GPT-4o	84.8	-5.1	82.3	-5.4	85.3	-5.0	84.1	-2.8	83.4	-4.8	81.7	-4.5	83.5	-4.6	82.0	-4.7	82.1	-5.0	78.0	-5.4	85.1	-5.0	82.5	-5.0	86.5	-4.6	86.5	-4.2	83.0	-4.2	82.5	-3.6
GPT-4o-mini	75.5	-4.9	78.8	1.5	75.3	-4.9	75.7	-3.3	75.3	-5.1	75.6	-3.1	75.0	-4.4	76.5	-0.7	73.2	-4.4	71.2	-0.6	75.0	-5.0	73.8	-2.3	78.7	-5.0	77.9	-3.3	71.2	-4.6	72.4	-4.5
Gemini 2.0 Flash Lite	79.1	-4.8	82.5	1.2	79.1	-5.0	79.4	-3.0	78.9	-5.0	79.4	-2.4	78.6	-4.3	80.2	-0.5	77.4	-4.9	74.8	-0.8	77.8	-4.2	77.5	-2.0	82.4	-5.0	81.6	-3.4	75.4	-5.1	76.1	-5.0
Gemini 1.5 Flash 8B	73.9	-5.0	77.1	1.1	73.7	-5.0	74.0	-3.0	72.7	-4.2	74.2	-2.3	74.1	-5.2	74.8	-0.8	72.2	-5.1	69.5	-0.6	73.5	-5.2	72.1	-1.2	76.4	-4.4	76.2	-4.0	67.6	-2.7	70.7	-3.6
One-shot, *Without Rationales*																																
GPT-4o	86.2	-5.1	83.0	-4.6	85.1	-4.0	83.9	-5.0	84.4	-4.4	83.2	-3.6	84.2	-4.5	81.1	-5.2	83.7	-5.1	74.7	-10.4	84.1	-4.4	82.6	-4.4	86.4	-4.8	86.3	-4.3	83.8	-4.0	82.8	-5.6
GPT-4o-mini	77.1	-4.5	75.0	-4.4	77.0	-4.3	75.0	-4.4	74.4	-4.2	73.9	-3.5	76.4	-4.3	73.7	-4.0	75.4	-4.8	68.8	-6.0	76.2	-5.4	72.7	-5.4	78.7	-5.1	78.3	-4.7	76.6	-4.3	74.7	-3.3
Gemini 2.0 Flash Lite	81.1	-4.8	78.7	-4.2	80.7	-4.3	78.7	-4.2	78.8	-4.9	77.6	-3.1	80.8	-5.0	77.4	-4.0	78.8	-4.5	72.5	-6.6	79.5	-5.0	76.4	-5.0	81.7	-4.4	82.0	-4.6	80.4	-4.4	78.4	-3.2
Gemini 1.5 Flash 8B	75.9	-5.0	73.3	-3.9	76.3	-5.3	73.3	-4.4	73.5	-5.0	72.3	-3.2	75.4	-5.0	72.0	-4.1	73.3	-4.4	67.1	-5.5	73.9	-4.8	71.0	-5.5	76.9	-5.0	76.6	-3.8	75.8	-5.2	73.0	-5.9
One-shot, *With Rationales*																																
GPT-4o	**86.4**	-5.0	**85.6**	-5.0	**85.5**	-4.2	**84.6**	-4.5	**85.0**	-4.0	**84.9**	-4.0	**86.4**	-5.0	**84.9**	-5.4	**84.2**	-5.1	**80.7**	-5.0	**85.4**	-4.5	**84.1**	-4.5	**87.3**	-5.0	**87.1**	-4.3	**86.4**	-5.0	**86.5**	-3.0
GPT-4o-mini	78.1	-5.0	79.1	-2.0	78.1	-5.2	76.0	-4.0	74.8	-2.6	77.6	-2.0	77.1	-3.9	76.3	-2.8	75.6	-4.7	72.4	-3.8	77.4	-4.9	72.6	-5.8	78.9	-5.0	76.9	-5.4	78.2	-5.0	75.6	-4.0
Gemini 2.0 Flash Lite	81.1	-4.3	82.8	-2.1	81.7	-5.2	79.6	-3.5	78.6	-2.7	81.3	-2.3	80.7	-3.8	80.0	-3.6	79.0	-4.4	76.1	-3.5	81.7	-5.5	76.3	-5.8	82.6	-5.0	80.6	-4.8	81.8	-4.9	79.3	-3.4
Gemini 1.5 Flash 8B	76.3	-4.9	77.4	-2.0	76.1	-5.0	74.3	3.4	75.5	-5.0	75.9	-2.4	77.2	-5.7	74.6	-4.0	73.9	-4.7	70.7	-3.9	75.4	-4.6	70.9	-5.8	77.2	-5.0	75.2	-5.0	76.8	-5.3	73.9	-3.7
Human Performance	–	–	95.5	–	–	–	95.1	–	–	–	94.7	–	–	–	94.1	–	–	–	95.7	–	–	–	93.4	–	–	–	89.9	–	–	–	80.8	–

Table 16: Performance of VLMs on **eight VL tasks** under zero-shot and one-shot settings, measured by estimated Accuracy (%) from LLM-as-a-judge. **Input:** Swahili Texts + Images; **Output:** En, Sw responses. Shaded columns represent {Sw} in input and {Sw} in output VLM results. En ΔAcc. = {En score from Table 5 (En-Input) – En score from this Table (Sw-Input)}. Sw ΔAcc. = {Sw score from Table 5 (En-Input) – Sw score from this Table (Sw-Input).} Positive Δ in blue, negative Δ in red. Zeros indicate unintelligible responses or no support language.

Model	Object Recognition				Scene Understanding				Relation Understanding				Semantic Segmentation				Image Captioning				Image-Text Matching				Unrelatedness				Visual Question Answering			
	En	Δ_{En}	Ur	Δ_{Ur}	En	Δ_{En}	Ur	Δ_{Ur}	En	Δ_{En}	Ur	Δ_{Ur}	En	Δ_{En}	Ur	Δ_{Ur}	En	Δ_{En}	Ur	Δ_{Ur}	En	Δ_{En}	Ur	Δ_{Ur}	En	Δ_{En}	Ur	Δ_{Ur}	En	Δ_{En}	Ur	Δ_{Ur}
Zero-shot, *Without Rationales*																																
GPT-4o	88.9	-1.5	87.8	0.0	87.6	-1.0	87.3	-0.6	87.6	-1.1	83.2	0.9	83.3	-0.8	80.0	-0.5	84.5	-1.8	78.3	-0.5	89.1	-1.7	**88.3**	2.4	**91.0**	-2.0	90.1	-0.5	88.6	-1.6	88.6	1.6
GPT-4o-mini	79.3	-1.7	82.1	2.5	79.7	-2.0	79.1	0.6	78.4	-2.0	74.5	-1.8	76.4	-1.5	72.1	-2.7	77.4	-1.5	71.1	-1.2	76.6	-1.0	76.6	-3.0	82.1	-1.5	81.2	-0.7	80.1	-1.5	80.6	3.0
Gemini 2.0 Flash Lite	82.5	-1.2	85.8	1.9	82.6	-1.2	82.8	0.2	81.3	-1.2	78.2	-1.3	79.8	-1.2	75.8	-2.4	81.0	-1.5	74.8	-1.5	80.8	-1.5	80.3	-2.6	86.6	-2.3	84.9	-0.9	83.3	-1.0	84.3	4.0
Gemini 1.5 Flash 8B	77.5	-1.6	80.4	1.5	77.4	-1.4	77.4	0.6	76.0	-1.3	72.8	-1.7	74.1	-0.9	70.4	-2.8	75.8	-1.6	69.4	-1.7	74.6	-0.7	74.9	-2.6	80.9	-2.0	79.5	-0.9	79.0	-2.1	78.9	3.5
PALO 7B	38.6	-2.1	44.3	-5.9	38.6	-2.0	41.3	-8.2	37.4	-2.1	36.7	-8.8	35.8	-2.0	34.3	-9.8	36.6	-1.8	33.3	-8.9	36.3	-1.8	38.8	-11.1	41.6	-2.1	43.4	-9.0	39.8	-2.3	42.8	-3.6
MAYA 8B	40.2	-2.4	45.1	-5.1	39.9	-2.0	42.1	-5.9	38.6	-2.0	37.5	-7.9	37.4	-2.3	35.1	-8.9	38.2	-2.3	34.1	-8.5	38.3	-2.5	39.6	-10.2	43.2	-2.4	44.2	-6.7	40.4	-1.6	43.6	-3.6
PALO 13B	40.2	-1.5	42.4	-13.8	40.6	-2.5	39.3	-16.1	38.8	-2.0	34.7	-17.0	37.3	-2.3	32.3	-17.3	38.6	-2.0	31.3	-16.9	38.0	-2.2	36.8	-18.8	43.0	-2.3	41.4	-18.0	40.9	-2.4	40.9	-12.9
Zero-shot, *With Rationales*																																
GPT-4o	88.0	-1.3	85.8	-1.2	88.7	-2.0	87.6	1.6	87.2	-1.6	85.2	-0.0	87.2	-1.3	85.5	2.0	85.6	-1.5	81.5	-1.3	88.8	-1.7	86.0	-0.3	90.9	-2.0	90.0	-0.7	86.9	-2.1	86.0	-0.2
GPT-4o-mini	79.1	-1.5	82.3	0.3	78.9	-1.5	79.2	0.2	79.1	-1.9	79.1	3.2	78.5	-0.9	80.0	4.2	77.4	-1.6	74.7	2.3	78.4	-1.4	77.3	1.3	82.3	-1.6	81.4	-1.4	75.3	-1.7	75.9	-2.9
Gemini 2.0 Flash Lite	82.9	-1.6	86.0	1.4	82.8	-1.7	82.9	1.2	82.4	-1.5	82.9	3.3	82.8	-1.5	83.7	3.8	81.1	-1.6	78.3	1.5	81.7	-1.1	81.0	1.5	86.0	-1.6	85.1	-1.7	79.0	-1.7	79.6	-2.7
Gemini 1.5 Flash 8B	77.6	-1.7	80.6	1.0	77.2	-1.5	77.5	1.2	77.5	-2.0	77.7	4.0	78.0	-2.1	78.3	5.0	76.1	-2.0	73.0	2.3	77.3	-2.0	75.6	1.5	80.5	-1.5	79.7	-1.3	73.5	-1.6	74.2	-2.7
PALO 7B	39.0	-2.5	45.0	-6.4	38.4	-2.1	40.7	-6.8	38.4	-2.3	41.4	-4.7	38.5	-2.0	42.2	-3.1	36.8	-2.0	36.8	-5.5	37.9	-2.0	39.5	-6.3	42.1	-2.5	43.6	-9.2	35.3	-2.8	38.1	-9.7
MAYA 8B	39.0	-1.2	45.3	-6.0	39.4	-1.8	42.2	-6.2	39.2	-1.8	42.0	-3.8	39.8	-2.0	43.0	-2.4	37.9	-1.9	37.7	-5.0	39.5	-2.3	40.3	-4.9	42.9	-2.0	44.4	-8.1	35.8	-2.0	38.9	-8.6
PALO 13B	40.3	-2.3	42.5	-15.9	39.8	-2.0	39.2	-15.8	39.6	-2.0	39.2	-13.8	40.9	-2.5	40.0	-10.8	38.6	-2.3	34.9	-13.7	39.4	-2.0	38.0	-14.9	43.4	-2.1	42.1	-16.9	35.3	-1.0	36.2	-19.0
One-shot, *Without Rationales*																																
GPT-4o	**89.5**	-1.4	86.5	-0.6	89.8	-1.7	87.4	-0.5	88.6	-1.6	86.7	2.7	87.8	-1.1	84.6	-0.5	86.8	-1.2	78.2	-2.1	88.7	-2.0	86.1	0.5	89.7	-1.1	89.8	-0.7	88.3	-1.5	86.3	0.4
GPT-4o-mini	80.6	-1.0	78.5	0.5	81.3	-1.6	78.5	-0.6	78.2	-1.0	77.4	1.8	80.1	-1.0	77.2	1.2	78.6	-1.0	72.3	-1.3	79.2	-1.4	76.2	1.4	82.0	-1.4	81.8	-0.6	80.3	-1.0	78.2	1.8
Gemini 2.0 Flash Lite	85.3	-2.0	82.2	0.7	84.7	-1.3	82.2	-0.5	81.9	-1.0	81.1	2.0	84.4	-1.6	80.9	2.1	82.8	-1.5	76.0	-1.9	82.9	-1.4	79.9	1.7	85.3	-1.0	85.5	-0.8	84.6	-1.6	81.9	1.6
Gemini 1.5 Flash 8B	79.6	-1.7	76.8	0.8	79.6	-1.6	76.8	-0.5	76.2	-0.7	75.8	2.0	78.3	-0.9	75.5	2.0	77.6	-1.7	70.6	-1.3	77.7	-1.6	74.5	1.2	79.6	-0.7	80.1	-0.6	79.3	-1.7	76.5	1.9
PALO 7B	40.5	-2.0	41.2	-6.7	40.6	-2.0	40.7	-8.0	38.1	-2.0	39.5	-6.2	40.3	-2.3	39.2	-6.0	38.3	-1.8	34.9	-6.4	38.7	-2.0	38.3	-10.9	42.0	-2.5	44.0	-8.4	39.7	-1.5	40.4	-7.1
MAYA 8B	41.1	-1.3	41.5	-6.0	42.4	-2.5	41.5	-5.7	39.8	-2.4	40.3	-5.2	41.7	-2.4	40.3	-4.7	40.0	-2.2	35.3	-5.7	39.8	-1.8	39.2	-10.2	42.1	-1.3	44.8	-7.2	41.3	-1.8	41.2	-5.2
PALO 13B	41.7	-1.2	38.7	-16.2	42.1	-1.8	38.7	-15.8	38.5	-0.9	37.6	-14.6	41.3	-1.9	37.0	-14.9	40.7	-2.6	32.5	-17.6	40.2	-2.0	36.1	-16.0	42.8	-1.2	42.2	-16.5	41.2	-1.8	38.2	-16.0
One-shot, *With Rationales*																																
GPT-4o	89.4	-1.0	**89.1**	1.3	**90.0**	-1.7	**88.1**	0.6	**90.1**	-2.1	**88.4**	1.8	**89.9**	-1.5	**88.4**	2.3	**87.2**	-1.1	**84.2**	0.8	**88.6**	-0.7	**87.6**	-0.7	90.4	-1.1	**90.6**	-0.5	**90.3**	-1.9	**90.0**	2.6
GPT-4o-mini	81.6	-1.5	82.6	2.3	81.6	-1.7	79.5	1.7	80.7	-1.5	81.1	2.8	81.7	-1.5	79.8	2.3	78.8	-0.9	75.9	1.9	80.5	-1.0	76.1	-3.0	81.8	-0.9	80.4	-2.5	82.2	-2.0	79.1	-0.7
Gemini 2.0 Flash Lite	85.4	-1.6	86.3	2.2	84.5	-1.0	83.1	1.7	83.9	-1.0	84.8	3.1	84.8	-0.9	83.5	2.3	82.3	-0.7	79.6	2.0	84.8	-1.6	79.8	-3.5	86.5	-1.9	84.1	-2.1	84.8	-0.9	82.8	-0.5
Gemini 1.5 Flash 8B	79.9	-1.5	80.9	3.2	80.1	-2.0	77.8	1.7	78.9	-1.4	79.4	2.4	80.5	-2.0	78.1	2.7	77.1	-0.9	74.2	2.0	79.8	-2.0	74.4	-2.9	81.3	-2.1	78.7	-1.8	79.3	-0.8	77.4	-0.7
PALO 7B	41.3	-2.1	45.3	-4.8	40.7	-2.0	41.6	-6.7	40.2	-2.1	42.7	-6.0	41.4	-2.3	42.0	-7.0	38.6	-1.8	38.1	-6.5	40.4	-2.0	38.3	-10.4	41.8	-2.0	42.6	-9.4	41.2	-2.1	41.3	-7.5
MAYA 8B	42.3	-2.0	46.2	-4.1	42.0	-2.0	42.5	-4.8	41.2	-1.8	44.1	-4.3	42.4	-2.0	42.8	-4.5	40.2	-2.1	38.9	-4.9	42.2	-2.5	39.0	-8.9	42.1	-1.0	43.4	-7.8	41.6	-1.2	42.1	-6.2
PALO 13B	41.7	-1.2	42.6	-14.6	42.5	-2.2	39.7	-14.5	41.8	-2.3	41.2	-13.5	42.4	-1.8	39.8	-14.7	40.4	-2.1	36.0	-14.2	41.3	-2.0	35.5	-16.2	43.3	-2.3	41.4	-17.7	42.5	-2.5	39.3	-15.4
Human Performance	–	–	98.4	–	–	–	99.4	–	–	–	98.8	–	–	–	99.2	–	–	–	98.7	–	–	–	84.9	–	–	–	87.4	–	–	–	90.5	–

Table 17: Performance of VLMs on **eight VL tasks** under zero-shot and one-shot settings, measured by estimated Accuracy (%) from LLM-as-a-judge. **Input:** Urdu Texts + Images; **Output:** En, Ur responses. Shaded columns represent {Ur} in input and {Ur} in output VLM results. En ΔAcc. = {En score from Table 5 (En-Input) – En score from this Table (Ur-Input)}. Ur ΔAcc. = {Ur score from Table 5 (En-Input) – Ur score from this Table (Ur-Input).} Positive Δ in blue, negative Δ in red. We exclude Qwen2VL 7B, and LlaVa Mistral 7B/13B from experiments due to limited Urdu support, from our initial experiments. Zeros indicate unintelligible responses or no support language.

Model	Object Recognition				Scene Understanding				Relation Understanding				Semantic Segmentation			
	En	Δ_{En}	Ur	Δ_{Ur}	En	Δ_{En}	Ur	Δ_{Ur}	En	Δ_{En}	Ur	Δ_{Ur}	En	Δ_{En}	Ur	Δ_{Ur}
Zero-shot, *Without Rationales*																
PALO 7B	42.3	-3.0	48.1	-11.4	42.2	-3.2	45.1	-12.6	40.3	-3.8	40.5	-13.9	39.0	-3.6	38.1	-14.7
MAYA 8B	48.0	4.7	51.3	-7.3	47.6	4.7	48.3	-9.0	44.8	4.7	43.7	-10.6	46.5	4.7	41.3	-11.6
PALO 13B	49.7	0.4	52.5	0.2	50.0	-0.5	49.4	0.4	48.7	-0.5	44.8	0.4	47.0	-0.6	42.4	-0.5
Zero-shot, *With Rationales*																
PALO 7B	48.1	-0.2	55.1	4.8	47.9	-0.2	50.8	4.8	47.7	-0.2	51.5	4.8	48.1	-0.2	52.3	4.8
MAYA 8B	49.4	-0.2	55.4	-0.2	49.2	-0.2	52.3	-0.2	49.0	-0.2	52.1	-0.2	49.4	-0.2	53.1	-0.2
PALO 13B	49.5	-0.1	52.6	-0.3	49.4	-0.2	49.3	-0.2	49.3	-0.3	49.3	-0.6	50.1	-0.3	50.1	-0.5
One-shot, *Without Rationales*																
PALO 7B	50.1	-0.6	51.3	4.7	50.6	-0.6	50.8	4.4	47.8	-0.3	49.6	5.0	49.9	-0.5	49.3	5.3
MAYA 8B	51.6	-0.4	51.6	-0.4	51.8	-0.5	51.6	-0.1	49.3	-0.5	50.4	-0.4	51.4	-0.7	50.4	-0.3
PALO 13B	52.6	-0.5	48.8	-0.5	52.3	-0.6	48.8	-0.3	49.4	-0.4	47.7	-0.4	50.9	-0.1	47.1	-0.7
One-shot, *With Rationales*																
PALO 7B	50.8	-0.2	55.4	5.6	50.5	-0.4	51.7	5.1	50.0	-0.5	52.8	5.1	50.8	-0.3	52.1	4.6
MAYA 8B	52.5	-0.8	56.3	-0.8	51.7	-0.3	52.6	-0.3	51.7	-0.9	54.2	-0.9	52.7	-0.9	52.9	-0.8
PALO 13B	52.3	-0.4	52.7	-0.6	52.0	-0.3	49.8	-0.6	51.3	-0.4	51.3	-0.5	52.4	-0.4	50.0	-0.5

Model	Image Captioning				Image-Text Matching				Unrelatedness				Visual Question Answering			
	En	Δ_{En}	Ur	Δ_{Ur}	En	Δ_{En}	Ur	Δ_{Ur}	En	Δ_{En}	Ur	Δ_{Ur}	En	Δ_{En}	Ur	Δ_{Ur}
Zero-shot, *Without Rationales*																
PALO 7B	39.9	-3.7	37.1	-13.2	39.5	-3.8	42.6	-15.4	45.0	-3.3	47.2	-12.6	42.4	-3.9	46.6	-9.6
MAYA 8B	42.8	5.7	40.3	-11.6	43.3	4.9	45.8	-12.6	48.3	4.5	50.4	-10.2	46.9	4.3	49.8	-5.7
PALO 13B	47.6	0.4	41.4	0.4	47.9	-0.7	46.9	0.3	52.9	-13.0	51.5	-0.4	50.8	-0.9	51.0	-0.3
Zero-shot, *With Rationales*																
PALO 7B	46.8	-0.6	46.9	4.3	47.8	-0.5	49.6	5.1	51.4	-4.0	53.7	2.6	44.4	-0.5	48.2	5.2
MAYA 8B	47.7	-0.3	47.8	-0.4	49.1	-0.5	50.4	-0.5	52.9	-2.5	54.5	-0.1	45.6	-0.4	49.0	-0.4
PALO 13B	48.1	-0.4	45.0	-0.5	49.1	-0.3	48.1	-0.4	52.5	1.3	52.2	-0.4	46.2	-0.3	46.3	-0.4
One-shot, *Without Rationales*																
PALO 7B	48.3	-0.4	45.0	5.0	48.7	-0.6	48.4	4.2	51.3	-3.1	54.1	2.5	50.2	-0.6	50.5	5.1
MAYA 8B	49.7	-0.5	45.4	-0.4	50.0	-0.6	49.3	-0.1	52.9	-1.7	54.9	-0.7	51.6	-0.7	51.3	-0.4
PALO 13B	50.1	-0.6	42.6	-0.5	50.0	-0.4	46.2	-0.7	53.3	11.9	52.3	-0.7	51.3	-0.5	48.3	-0.3
One-shot, *With Rationales*																
PALO 7B	48.6	-0.4	48.2	5.3	50.5	-0.7	48.4	4.9	51.3	-6.6	52.7	0.9	51.2	-0.7	51.4	4.8
MAYA 8B	50.3	-0.8	49.0	-0.9	52.0	-0.9	49.1	-0.9	52.7	-5.0	53.5	-0.7	52.7	-0.9	52.2	-0.3
PALO 13B	50.4	-0.5	46.1	-0.4	51.3	-0.6	45.6	-0.3	52.2	-11.8	51.5	-0.6	51.5	-0.1	49.4	-0.5

Table 18: Performance of VLMs on **eight VL tasks** under finetuning settings, measured by estimated Accuracy (%) from LLM-as-a-judge. **Input:** Urdu Texts + Images; **Output:** En, Ur responses. Shaded columns represent {Ur} in input and {Ur} in output VLM results. En ΔAcc. = {En score from Table 6 (En-Input) – En score from this Table (Ur-Input)}. Ur ΔAcc. = {Ur score from Table 6 (En-Input) – Ur score from this Table (Ur-Input).} Positive Δ in blue, negative Δ in red. We exclude Qwen2VL 7B, and LlaVa Mistral 7B/13B from experiments due to limited Ur support, from our initial experiments. Zeros indicate unintelligible responses or no support language.

Model	Object Recognition	Scene Understanding	Relation Understanding	Semantic Segmentation	Image Captioning	Image-Text Matching	Unrelatedness	Visual Quesiton Answering
Zero-shot, *Without Rationales*								
GPT-4o	87.0	86.8	82.6	77.0	82.1	89.9	97.3	92.9
GPT-4o-mini	81.3	78.6	73.9	69.1	74.9	78.2	88.4	84.9
Gemini 2.0 Flash Lite	85.0	82.3	77.6	72.8	78.6	81.9	92.1	88.6
Gemini 1.5 Flash 8B	79.6	76.9	72.2	67.4	73.2	76.5	86.7	83.2
Zero-shot, *With Rationales*								
GPT-4o	85.0	87.1	84.6	82.5	85.3	87.6	97.2	90.3
GPT-4o-mini	81.5	78.7	78.5	77.0	78.5	78.9	88.6	80.2
Gemini 2.0 Flash Lite	85.2	82.4	82.3	80.7	82.1	82.6	92.3	83.9
Gemini 1.5 Flash 8B	79.8	77.0	77.1	75.3	76.8	77.2	86.9	78.5
One-shot, *Without Rationales*								
GPT-4o	85.7	86.9	86.1	81.6	82.0	87.7	97.0	90.6
GPT-4o-mini	77.7	78.0	76.8	74.2	76.1	77.8	89.0	82.5
Gemini 2.0 Flash Lite	81.4	81.7	80.5	77.9	79.8	81.5	92.7	86.2
Gemini 1.5 Flash 8B	76.0	76.3	75.2	72.5	74.4	76.1	87.3	80.8
One-shot, *With Rationales*								
GPT-4o	**88.3**	**87.6**	**87.8**	**85.4**	**88.0**	**89.2**	**97.8**	**94.3**
GPT-4o-mini	81.8	79.0	80.5	76.8	79.7	77.7	87.6	83.4
Gemini 2.0 Flash Lite	85.5	82.6	84.2	80.5	83.4	81.4	91.3	87.1
Gemini 1.5 Flash 8B	80.1	77.3	78.8	75.1	77.9	75.7	85.9	81.7

Table 19: Performance of VLMs on **eight VL tasks** under zero-shot and one-shot settings, measured by estimated Accuracy (%) from LLM-as-a-judge. **Input:** Swahili Texts + Images; **Output:** Swahili responses. We do not include LlaVa Mistral 7B, PALO 7B, MAYA 7B, LlaVa Mistral 13B, PALO 13B, because they showed very little understanding of Sw in our initial experiments.

K Robustness of VLMs: En, Ja, Sw, Ur

In §7 we introduced the *VLURes* Difficulty and Cross-Language Stability. Across tasks, proprietary VLMs are consistently more robust to cross-language shifts (En/Ja/Sw/Ur) under the *One-shot, With Rationales* setting: GPT-4o achieves the lowest difficulty on every task (0.107–0.163) with minimal cross-task variation, followed by GPT-4o-mini and the Gemini Flash variants. Open models are markedly less stable, exhibiting substantially higher average difficulty (roughly 0.36–0.69 where measurable) and reduced language coverage (see $^{(n)}$). At the language level (N=4 common models), difficulty increases from English (0.153) to Japanese (0.206) and Swahili (0.214), with Urdu in between (0.179), suggesting a systematic but moderate degradation beyond English even for strong models. When averaging over all available models, English/Japanese/Urdu become far harder (0.382/0.443/0.422), while Swahili remains artificially low (0.214) because only the four strongest models are evaluated there, emphasizing that robustness must be interpreted jointly with model coverage; this pattern aligns with prior findings that multilingual VLM generalization is uneven and gaps widen in non-English, especially low-resource, settings.

Language	Average Difficulty
English	0.153
Japanese	0.206
Swahili*	0.214
Urdu	0.179

Table 20: Overall Average Difficulty per Language. Calculated from *One-shot, With Rationales* accuracies, averaged over 8 tasks using N=4 common VLMs (GPT-4o, GPT-4o-mini, Gemini 2.0 Flash Lite, Gemini 1.5 Flash 8B). Difficulty = $|1 - \text{Avg. Accuracy}|$. **Lower scores indicate lower average difficulty.**

Language	No. of Models (N)	Average Difficulty
English	10	0.382
Japanese	10	0.443
Swahili*	4	0.214
Urdu	9	0.422

Table 21: Overall Average Difficulty per Language (all available models). Calculated from *One-shot, With Rationales* accuracies. For each language, the average difficulty (1 - Avg. Accuracy of a model on that language) is first computed per model, then these model-specific average difficulties are averaged across all models evaluated for that language. The number of models included in each language's average is noted. **Lower scores indicate lower average difficulty**.

K.1 Detailed Cross-lingual performance shifts (Across four prompting settings): En→Ja, En→Sw, En→Ur.

Tables 23 to 30 quantify how VLM accuracy changes when we keep the task fixed but switch the interaction language from English to Japanese, Swahili, or Urdu under four prompting regimes (ZS/OS $\times$ with-/without rationales). Throughout, negative Δ values (red) indicate a performance drop relative to English, while positive values (blue) indicate an increase; zeros for several open models on Sw/Ur reflect unsupported evaluations rather than true "no change."

1. A persistent cross-lingual gap on image-only understanding tasks. For the five image-only tasks (OR, SU, RU, SS, IC), all four strong closed models (GPT-4o, GPT-4o-mini, Gemini 2.0 Flash Lite, Gemini 1.5 Flash 8B) show consistent cross-lingual degradation when moving away from English, typically on the order of ∼3–5 points for Japanese and Swahili and ∼1–2 points for Urdu (Tables 23–27). This pattern aligns with broader evidence that VLMs remain sensitive to input-language shifts even when the visual content is unchanged, and that multilingual evaluation often exposes sizeable language-conditioned gaps (Geigle et al., 2024; Pfeiffer et al., 2022). Among the open models, the Japanese drops are markedly larger (often ∼8–17 points across OR/SU/RU/SS/IC), indicating weaker robustness to non-English instructions and/or reduced Japanese coverage during pretraining and instruction tuning (Chen et al., 2023b; Geigle et al., 2024).

2. Rationales and one-shot examples raise accuracy, but they rarely *close* the language gap. Across OR/SU/RU/SS/IC, adding rationales and/or a one-shot demonstration does not consistently reduce the En→Ja/Sw/Ur deltas: some tasks show marginal attenuation, others remain unchanged, and in several cases the Swahili gap slightly widens (Tables 23–27). This suggests that rationale prompting mainly helps the model "reason better" given a language it already handles well, but does not directly repair underlying multilingual representation/alignment weaknesses, consistent with prior observations that prompting-based reasoning aids (e.g., chain-of-thought) improve performance without guaranteeing robustness under distribution or language shifts (Wei et al., 2022b). In our setting, the remaining deltas therefore reflect a more structural multilingual limitation rather than a prompt-format artifact.

3. Text-image alignment behaves differently: ITM often *improves* outside English. A striking contrast

Model	Average Difficulty per Task							
	OR	SU	RU	SS	IC	ITM	U	VQA
GPT-4o	0.120	0.128	0.127	0.126	0.163	0.133	0.107	0.114
GPT-4o-mini	0.190	0.213	0.203	0.211	0.246	0.240	0.204	0.216
Gemini 2.0 Flash Lite	0.153	0.177	0.166	0.174	0.209	0.203	0.167	0.179
Gemini 1.5 Flash 8B	0.207	0.230	0.220	0.228	0.263	0.257	0.221	0.233
LlaVa Mistral 7B	$0.630^{(2)}$	$0.654^{(2)}$	$0.648^{(2)}$	$0.651^{(2)}$	$0.686^{(2)}$	$0.680^{(2)}$	$0.643^{(2)}$	$0.657^{(2)}$
Qwen2VL 7B	$0.362^{(2)}$	$0.370^{(2)}$	$0.382^{(2)}$	$0.380^{(2)}$	$0.409^{(2)}$	$0.408^{(2)}$	$0.371^{(2)}$	$0.386^{(2)}$
PALO 7B	$0.566^{(3)}$	$0.579^{(3)}$	$0.586^{(3)}$	$0.588^{(3)}$	$0.621^{(3)}$	$0.615^{(3)}$	$0.581^{(3)}$	$0.592^{(3)}$
MAYA 8B	$0.555^{(3)}$	$0.581^{(3)}$	$0.573^{(3)}$	$0.578^{(3)}$	$0.612^{(3)}$	$0.606^{(3)}$	$0.572^{(3)}$	$0.583^{(3)}$
LlaVa Mistral 13B	$0.579^{(2)}$	$0.606^{(2)}$	$0.603^{(2)}$	$0.606^{(2)}$	$0.642^{(2)}$	$0.639^{(2)}$	$0.600^{(2)}$	$0.620^{(2)}$
PALO 13B	$0.576^{(3)}$	$0.598^{(3)}$	$0.592^{(3)}$	$0.597^{(3)}$	$0.630^{(3)}$	$0.630^{(3)}$	$0.585^{(3)}$	$0.603^{(3)}$

Table 22: Average Task Difficulty, i.e., $|1 - \text{Avg. Accuracy}|$, per VLM across Languages. Scores are obtained from the *One-shot, With Rationales* setting. The average is calculated over available languages for each VLM-task pair (up to 4: En, Ja, Sw, Ur). A $^{(n)}$ superscript indicates the number of languages included in the average if less than 4. **Lower scores indicate lower average difficulty (higher average accuracy)**. Values rounded to three decimal places.

appears for Image–Text Matching (ITM): most models exhibit positive deltas for Ja/Sw/Ur under all four settings, with especially large gains under one-shot prompting (e.g., +10 points for Ja in multiple cases; Table 28). This indicates that, in *VLURes*, ITM is not purely governed by language understanding; it is also shaped by how tightly the article text is grounded to the embedded image within each language's web sources. Concretely, ITM can become easier when the paired discourse in a given language more directly describes the visible content (or contains clearer "anchor" spans), highlighting that cross-lingual performance shifts can reflect both model multilinguality and language-specific data distributions (Geigle et al., 2024; Pfeiffer et al., 2022). At the same time, the fact that ITM improves while image-only tasks drop cautions against treating "good alignment" as synonymous with "deep visual understanding," echoing prior critiques that VLMs can exploit superficial correspondences and shortcuts (Thrush et al., 2022; Hsieh et al., 2023).

4. The proposed *Unrelatedness* task exposes fine-grained grounding differences across models and languages. For Unrelatedness (U), the deltas are smaller-magnitude and more model-dependent than in image-only tasks (Table 29). GPT-4o shows modest positive shifts for Sw/Ur (up to $\sim$+3 points), whereas several other models show slight negatives, especially on Sw/Ur, suggesting that reliably filtering *non-matching* spans requires stronger multilingual semantics plus robust token-level grounding. This is consistent with recent findings that many VLMs remain vulnerable to hallucination and weak faithfulness even when they appear strong on coarse alignment benchmarks (Li et al., 2023b; Hsieh et al., 2023). In this sense, Unrelatedness complements ITM: ITM rewards selecting what matches, while U stress-tests whether a model can suppress plausible but irrelevant textual signals under long-context multimodal inputs.

5. VQA shows mixed shifts, with the largest gains concentrated in Swahili for the strongest model. For VQA, the deltas are generally small for Japanese and Urdu but can be notably positive for Swahili in GPT-4o (Table 30). One plausible interpretation is that Swahili VQA instances in our web-derived setting often produce questions whose answers are visually salient and less confusable with background discourse, whereas Urdu shows a weaker and less consistent pattern. More broadly, cross-lingual VQA is known to be sensitive to both linguistic phenomena and dataset composition (question style, answer space, and cultural/lexical grounding), which can yield language-specific effects beyond raw translation difficulty (Pfeiffer et al., 2022; Geigle et al., 2024).

Summary and implication. Overall, the cross-lingual deltas reveal a consistent story: (i) core image understanding remains English-centric for current VLMs, (ii) rationales and one-shot prompting improve absolute performance but do not reliably eliminate language-conditioned drops, and (iii) alignment-heavy tasks (ITM/U) can move in the opposite direction, underscoring that "multilingual robustness" must be analyzed task-by-task rather than inferred from a single aggregate score. These trends motivate future VLM training that jointly targets multilingual coverage *and* fine-grained grounding/faithfulness, especially for long-text web settings where spurious textual cues and weak visual grounding are common failure modes (Geigle et al., 2024; Thrush et al., 2022; Li et al., 2023b).

Model	ZS w/o rat.			ZS w/ rat.			OS w/o rat.			OS w/ rat.		
	Ja	Sw	Ur	Ja	Sw	Ur	Ja	Sw	Ur	Ja	Sw	Ur
GPT-4o	-4.1	-4.5	-1.5	-3.8	-5.1	-1.3	-4.1	-5.1	-1.4	-3.7	-5.0	-1.0
GPT-4o-mini	-3.9	-4.5	-1.7	-4.3	-4.9	-1.5	-4.2	-4.5	-1.0	-3.5	-5.0	-1.5
Gemini 2.0 Flash Lite	-3.8	-5.2	-1.2	-3.9	-4.8	-1.6	-3.8	-4.8	-2.0	-3.2	-4.3	-1.6
Gemini 1.5 Flash 8B	-3.5	-5.4	-1.6	-3.7	-4.8	-1.6	-4.5	-4.8	-2.0	-3.7	-4.3	-1.6
LlaVa Mistral 7B	-14.3	0.0	0.0	-14.5	0.0	0.0	-14.8	0.0	0.0	-14.0	0.0	0.0
Qwen2VL 7B	-3.5	0.0	0.0	-3.7	0.0	0.0	-4.0	0.0	0.0	-4.9	0.0	0.0
PALO 7B	-3.7	0.0	-2.1	-3.6	0.0	-1.8	-3.7	0.0	-2.5	-3.5	0.0	-2.0
MAYA 8B	-4.0	0.0	-2.4	-4.3	0.0	-2.0	-4.1	0.0	-1.8	-4.2	0.0	-1.9
LlaVa Mistral 13B	-16.3	0.0	0.0	-16.0	0.0	0.0	-16.4	0.0	0.0	-15.7	0.0	0.0
PALO 13B	-3.8	0.0	-1.2	-3.8	0.0	-1.2	-4.1	0.0	-1.8	-3.6	0.0	-1.5

Table 23: Δ English accuracy (En→Ja, En→Sw, En→Ur) for each VLM on **Object Recognition** under four prompting settings: zero-shot without rationales (**ZS w/o rat.**), zero-shot with rationales (**ZS w/ rat.**), one-shot without rationales, (**OS w/o rat.**), one-shot with rationales (**OS w/ rat.**).

All $\Delta_{\text{En,lang}}$ are computed as $\Delta_{\text{En,lang}} = \text{Acc}_{\text{En-input}} - \text{Acc}_{\text{lang-input}}$.

Model	ZS w/o rat.			ZS w/ rat.			OS w/o rat.			OS w/ rat.		
	Ja	Sw	Ur	Ja	Sw	Ur	Ja	Sw	Ur	Ja	Sw	Ur
GPT-4o	-3.8	-4.7	-1.0	-4.0	-5.0	-2.0	-3.8	-4.0	-1.7	-3.7	-4.2	-1.7
GPT-4o-mini	-4.1	-5.4	-1.5	-3.7	-4.8	-1.5	-3.5	-4.3	-1.6	-3.6	-4.1	-1.6
Gemini 2.0 Flash Lite	-3.4	-5.4	-1.7	-3.2	-4.0	-1.6	-3.9	-4.1	-2.0	-3.0	-3.7	-1.5
Gemini 1.5 Flash 8B	-3.8	-5.2	-1.9	-3.7	-4.8	-1.8	-4.2	-4.1	-2.2	-3.0	-3.9	-1.6
LlaVa Mistral 7B	-8.3	0.0	0.0	-8.4	0.0	0.0	-9.0	0.0	0.0	-9.1	0.0	0.0
Qwen2VL 7B	-4.6	0.0	0.0	-4.0	0.0	0.0	-3.5	0.0	0.0	-3.8	0.0	0.0
PALO 7B	-3.4	0.0	-2.0	-3.4	0.0	-2.2	-3.7	0.0	-2.5	-3.6	0.0	-2.1
MAYA 8B	-3.7	0.0	-2.5	-3.8	0.0	-2.4	-3.6	0.0	-2.0	-3.8	0.0	-1.9
LlaVa Mistral 13B	-8.2	0.0	0.0	-7.8	0.0	0.0	-8.0	0.0	0.0	-9.1	0.0	0.0
PALO 13B	-3.8	0.0	-2.2	-3.6	0.0	-2.3	-3.8	0.0	-2.4	-3.6	0.0	-2.0

Table 24: Δ English accuracy (En→Ja, En→Sw, En→Ur) for each VLM on **Scene Understanding** under four prompting settings: zero-shot without rationales (**ZS w/o rat.**), zero-shot with rationales (**ZS w/ rat.**), one-shot without rationales (**OS w/o rat.**), one-shot with rationales (**OS w/ rat.**).

All $\Delta_{\text{En,lang}}$ are computed as $\Delta_{\text{En,lang}} = \text{Acc}_{\text{En-input}} - \text{Acc}_{\text{lang-input}}$.

Model	ZS w/o rat.			ZS w/ rat.			OS w/o rat.			OS w/ rat.		
	Ja	Sw	Ur	Ja	Sw	Ur	Ja	Sw	Ur	Ja	Sw	Ur
GPT-4o	-4.5	-4.6	-1.1	-3.4	-4.8	-1.6	-3.5	-4.4	-1.6	-4.4	-4.0	-2.1
GPT-4o-mini	-3.5	-4.5	-1.7	-3.7	-4.9	-1.5	-3.5	-4.5	-1.6	-3.7	-5.0	-1.6
Gemini 2.0 Flash Lite	-3.7	-5.2	-1.2	-3.7	-4.8	-1.6	-3.5	-4.8	-2.0	-3.7	-4.3	-1.6
Gemini 1.5 Flash 8B	-4.2	-5.4	-1.8	-3.7	-4.8	-1.6	-3.1	-4.8	-2.0	-3.7	-4.3	-1.6
LlaVa Mistral 7B	-11.8	0.0	0.0	-13.0	0.0	0.0	-13.5	0.0	0.0	-13.0	0.0	0.0
Qwen2VL 7B	-3.1	0.0	0.0	-3.5	0.0	0.0	-3.6	0.0	0.0	-2.7	0.0	0.0
PALO 7B	-3.0	0.0	-1.8	-3.2	0.0	-1.8	-3.0	0.0	-2.2	-3.6	0.0	-1.9
MAYA 8B	-4.2	0.0	-1.8	-3.9	0.0	-1.9	-4.0	0.0	-1.8	-3.6	0.0	-1.9
LlaVa Mistral 13B	-17.0	0.0	0.0	-15.9	0.0	0.0	-15.0	0.0	0.0	-17.0	0.0	0.0
PALO 13B	-4.4	0.0	-2.0	-4.0	0.0	-2.3	-4.2	0.0	-2.6	-3.6	0.0	-2.4

Table 25: Δ English accuracy (En→Ja, En→Sw, En→Ur) for each VLM on **Relation Understanding** under four prompting settings: zero-shot without rationales (**ZS w/o rat.**), zero-shot with rationales (**ZS w/ rat.**), one-shot without rationales (**OS w/o rat.**), one-shot with rationales (**OS w/ rat.**).

All $\Delta_{\text{En,lang}}$ are computed as $\Delta_{\text{En,lang}} = \text{Acc}_{\text{En-input}} - \text{Acc}_{\text{lang-input}}$.

Model	ZS w/o rat.			ZS w/ rat.			OS w/o rat.			OS w/ rat.		
	Ja	Sw	Ur	Ja	Sw	Ur	Ja	Sw	Ur	Ja	Sw	Ur
GPT-4o	-5.0	-5.1	-0.8	-4.6	-5.2	-0.9	-5.3	-4.8	-1.0	-4.5	-4.9	-0.7
GPT-4o-mini	-3.2	-5.1	-1.5	-3.0	-5.2	-1.0	-3.4	-4.0	-0.8	-3.6	-5.1	-1.2
Gemini 2.0 Flash Lite	-2.7	-5.0	-1.2	-3.0	-5.0	-1.3	-2.4	-4.3	-1.1	-3.0	-5.2	-1.0
Gemini 1.5 Flash 8B	-3.3	-5.0	-1.8	-2.9	-5.1	-1.3	-3.0	-5.1	-1.5	-3.1	-5.1	-1.6
LlaVa Mistral 7B	-13.9	0.0	0.0	-12.5	0.0	0.0	-13.8	0.0	0.0	-12.2	0.0	0.0
Qwen2VL 7B	-4.1	0.0	0.0	-4.2	0.0	0.0	-3.8	0.0	0.0	-4.0	0.0	0.0
PALO 7B	-4.0	0.0	-2.4	-4.3	0.0	-2.0	-4.1	0.0	-1.8	-4.2	0.0	-1.9
MAYA 8B	-3.7	0.0	-2.0	-3.6	0.0	-1.9	-3.7	0.0	-1.8	-3.5	0.0	-1.7
LlaVa Mistral 13B	-15.0	0.0	0.0	-15.9	0.0	0.0	-15.7	0.0	0.0	-15.0	0.0	0.0
PALO 13B	-4.0	0.0	-2.8	-4.1	0.0	-2.0	-4.3	0.0	-1.8	-3.9	0.0	-1.9

Table 26: Δ English accuracy on **Semantic Segmentation** (En→Ja, En→Sw, En→Ur) for each VLM under four prompting settings: zero-shot w/o rationales (**ZS w/o rat.**), zero-shot w/ rationales (**ZS w/ rat.**), one-shot w/o rationales (**OS w/o rat.**), one-shot w/ rationales (**OS w/ rat.**).
All $\Delta_{\text{En,lang}}$ are computed as $\Delta_{\text{En,lang}} = \text{Accuracy}_{\text{En-input}} - \text{Accuracy}_{\text{lang-input}}$.

Model	ZS w/o rat.			ZS w/ rat.			OS w/o rat.			OS w/ rat.		
	Ja	Sw	Ur	Ja	Sw	Ur	Ja	Sw	Ur	Ja	Sw	Ur
GPT-4o	-3.8	-4.8	-0.5	-3.7	-5.2	-0.2	-3.9	-4.1	-0.7	-3.5	-5.2	-0.1
GPT-4o-mini	-3.2	-4.8	-1.2	-2.9	-5.0	-0.6	-3.5	-3.8	-0.7	-3.6	-4.9	-0.3
Gemini 2.0 Flash Lite	-3.7	-4.5	-1.3	-3.3	-4.5	-1.0	-3.4	-4.1	-0.9	-3.3	-4.6	-0.4
Gemini 1.5 Flash 8B	-3.3	-4.9	-1.8	-3.0	-5.0	-1.3	-3.5	-4.4	-1.5	-3.1	-4.9	-1.2
LlaVa Mistral 7B	-13.9	0.0	0.0	-12.5	0.0	0.0	-13.8	0.0	0.0	-12.9	0.0	0.0
Qwen2VL 7B	-4.1	0.0	0.0	-4.2	0.0	0.0	-3.8	0.0	0.0	-4.0	0.0	0.0
PALO 7B	-4.0	0.0	-2.0	-4.3	0.0	-2.4	-4.1	0.0	-1.8	-4.2	0.0	-2.0
MAYA 8B	-3.3	0.0	-2.0	-3.6	0.0	-2.0	-3.1	0.0	-1.8	-3.7	0.0	-2.3
LlaVa Mistral 13B	-15.9	0.0	0.0	-15.0	0.0	0.0	-15.7	0.0	0.0	-15.8	0.0	0.0
PALO 13B	-4.1	0.0	-2.3	-4.1	0.0	-2.0	-4.3	0.0	-1.8	-3.7	0.0	-1.9

Table 27: Δ English accuracy on **Image Captioning** (En→Ja, En→Sw, En→Ur) for each VLM under four prompting settings: zero-shot without rationales (**ZS w/o rat.**), zero-shot with rationales (**ZS w/ rat.**), one-shot without rationales (**OS w/o rat.**), one-shot with rationales (**OS w/ rat.**).
All $\Delta_{\text{En,lang}}$ are computed as $\Delta_{\text{En,lang}} = \text{Acc}_{\text{En-input}} - \text{Acc}_{\text{lang-input}}$.

Model	ZS w/o rat.			ZS w/ rat.			OS w/o rat.			OS w/ rat.		
	Ja	Sw	Ur	Ja	Sw	Ur	Ja	Sw	Ur	Ja	Sw	Ur
GPT-4o	+0.8	+1.0	+3.5	+6.1	+5.0	+6.0	+10.2	+3.8	+9.4	+5.9	+5.3	+6.8
GPT-4o-mini	+5.7	+2.5	+3.3	+4.6	+8.4	+7.3	+10.0	+8.6	+9.7	+6.7	+7.3	+8.6
Gemini 2.0 Flash Lite	+5.7	+2.5	+3.3	+4.7	+8.4	+7.3	+3.1	+2.1	+6.7	+6.5	+4.7	+4.0
Gemini 1.5 Flash 8B	+5.6	+2.7	+3.7	+4.3	+8.3	+7.1	+7.7	+6.9	+7.3	+5.6	+5.0	+5.1
LlaVa Mistral 7B	+4.8	0.0	0.0	+4.9	0.0	0.0	+4.4	0.0	0.0	+4.5	0.0	0.0
Qwen2VL 7B	+2.9	0.0	0.0	+3.1	0.0	0.0	+1.7	0.0	0.0	+3.2	0.0	0.0
PALO 7B	+1.8	0.0	+1.6	+2.7	0.0	+2.3	+3.8	0.0	+3.4	+1.9	0.0	+2.3
MAYA 8B	+2.5	0.0	+2.1	+1.8	0.0	+3.4	+4.3	0.0	+3.4	+3.0	0.0	+2.9
LlaVa Mistral 13B	+2.2	0.0	0.0	+2.4	0.0	0.0	+2.7	0.0	0.0	+3.1	0.0	0.0
PALO 13B	+2.3	0.0	+2.3	+2.4	0.0	+3.7	+4.8	0.0	+5.0	+2.5	0.0	+2.7

Table 28: Δ English accuracy on **Image-Text Matching** (En→Ja, En→Sw, En→Ur) for each VLM under zero-shot without rationales (**ZS w/o rat.**), zero-shot with rationales (**ZS w/ rat.**), one-shot without rationales (**OS w/o rat.**), and one-shot with rationales (**OS w/ rat.**).
$\Delta_{\text{En,lang}} = \text{Acc}_{\text{En-input}} - \text{Acc}_{\text{lang-input}}$.

Model	ZS w/o rat.			ZS w/ rat.			OS w/o rat.			OS w/ rat.		
	Ja	Sw	Ur	Ja	Sw	Ur	Ja	Sw	Ur	Ja	Sw	Ur
GPT-4o	+0.2	+2.6	+1.9	+0.6	+2.6	+2.0	+0.7	+2.5	+2.8	+0.3	+1.9	+2.9
GPT-4o-mini	+0.4	-0.3	-0.7	+1.1	-1.2	-0.5	+0.6	-0.3	-1.0	+0.3	-0.1	-1.0
Gemini 2.0 Flash Lite	+0.4	-1.1	-0.9	+2.0	-1.5	-0.7	+0.5	-0.6	-1.4	+0.8	-1.7	-0.9
Gemini 1.5 Flash 8B	+0.3	-1.3	-0.7	+2.3	-1.0	-0.5	+0.5	-0.5	-1.4	+0.8	-1.7	-1.0
LlaVa Mistral 7B	+0.8	0.0	0.0	+0.6	0.0	0.0	+0.4	0.0	0.0	+0.5	0.0	0.0
Qwen2VL 7B	+0.9	0.0	0.0	+1.2	0.0	0.0	+1.5	0.0	0.0	+1.3	0.0	0.0
PALO 7B	+2.2	0.0	+2.1	+2.5	0.0	+1.9	+2.8	0.0	+2.9	+2.0	0.0	+2.4
MAYA 8B	+1.5	0.0	+2.3	+1.8	0.0	+3.0	+2.1	0.0	+3.1	+1.9	0.0	+2.2
LlaVa Mistral 13B	+2.1	0.0	0.0	+2.4	0.0	0.0	+2.6	0.0	0.0	+2.3	0.0	0.0
PALO 13B	+4.0	0.0	+2.7	+4.3	0.0	+2.3	+4.6	0.0	+3.5	+4.1	0.0	+2.5

Table 29: Δ English accuracy on **Unrelatedness** (En→Ja, En→Sw, En→Ur) for each VLM under zero-shot without rationales (**ZS w/o rat.**), zero-shot with rationales (**ZS w/ rat.**), one-shot without rationales (**OS w/o rat.**), and one-shot with rationales (**OS w/ rat.**).
$\Delta_{\text{En,lang}} = \text{Acc}_{\text{En-input}} - \text{Acc}_{\text{lang-input}}$.

Model	ZS w/o rat.			ZS w/ rat.			OS w/o rat.			OS w/ rat.		
	Ja	Sw	Ur	Ja	Sw	Ur	Ja	Sw	Ur	Ja	Sw	Ur
GPT-4o	+1.6	+6.4	-1.1	+3.4	+7.4	+0.0	+2.3	+5.2	+1.2	+1.1	+3.8	+2.4
GPT-4o-mini	+0.4	-1.4	+0.6	+1.2	-0.8	+0.2	+0.9	-0.5	+0.8	+0.7	+0.3	+1.1
Gemini 2.0 Flash Lite	+0.5	-0.6	+0.8	+1.0	-0.2	+1.0	+1.2	+0.0	+1.4	+1.0	+0.3	+1.5
Gemini 1.5 Flash 8B	+0.2	-1.0	+0.5	+0.8	-0.4	+0.7	+1.0	-0.2	+0.9	+0.8	+0.1	+1.2
LlaVa Mistral 7B	+3.3	0.0	0.0	+2.9	0.0	0.0	+2.4	0.0	0.0	+2.6	0.0	0.0
Qwen2VL 7B	+1.2	0.0	0.0	+2.1	0.0	0.0	+1.8	0.0	0.0	+1.9	0.0	0.0
PALO 7B	+2.7	0.0	+0.7	+3.0	0.0	+2.5	+3.2	0.0	+2.0	+2.8	0.0	+1.5
MAYA 8B	+2.5	0.0	+2.1	+2.8	0.0	+2.5	+3.0	0.0	+2.2	+2.9	0.0	+1.2
LlaVa Mistral 13B	+3.7	0.0	0.0	+4.0	0.0	0.0	+3.8	0.0	0.0	+3.9	0.0	0.0
PALO 13B	+4.1	0.0	+1.3	+4.3	0.0	+1.9	+4.5	0.0	+3.1	+4.2	0.0	+2.0

Table 30: Δ English accuracy on **Visual Question Answering** (En→Ja, En→Sw, En→Ur) for each VLM under zero-shot without rationales (**ZS w/o rat.**), zero-shot with rationales (**ZS w/ rat.**), one-shot without rationales (**OS w/o rat.**), and one-shot with rationales (**OS w/ rat.**).
$\Delta_{\text{En,lang}} = \text{Acc}_{\text{En-input}} - \text{Acc}_{\text{lang-input}}$.

K.2 Detailed Cross-lingual Results (After Fine-tuning): En→Ja, En→Sw, En→Ur

We now revisit the same cross-lingual analysis after fine-tuning the open VLMs on our English training split, and re-evaluating them under the four prompting regimes (ZS w/o rat., ZS w/ rat., OS w/o rat., OS w/ rat.). Tables 31–38 report the cross-lingual change in accuracy when the *input language* is switched from English to another language. Throughout this subsection, **negative** values indicate an accuracy **drop** relative to English, while **positive** values indicate an **gain** (in percentage points).

Takeaway 1: Fine-tuning yields highly model-dependent cross-lingual transfer. After fine-tuning, models with stronger multilingual pretraining signals exhibit substantially smaller language sensitivity, and in some tasks they even show positive transfer to Japanese. For example, Qwen2VL-7B shows consistent gains on Japanese for the two image-text alignment tasks, with improvements of roughly +4.6 to +5.5 pp on Image Captioning and Image-Text Matching across prompting settings (Tables 35, 36), and similarly positive gains on Japanese VQA (up to about +5.4 pp, Table 38). These results are consistent with the broader trend that multilingual VLM performance depends strongly on the diversity and scale of multilingual image-text pretraining data and objectives (Chen et al., 2023b; Geigle et al., 2024).

Takeaway 2: English-only fine-tuning can amplify cross-lingual gaps for English-centric models. In contrast, LLaVa-Mistral-7B remains highly sensitive to Japanese inputs after fine-tuning, with large drops that persist across core image-only tasks. Under zero-shot prompting, Japanese accuracy decreases by roughly -11 to -12 pp for Object Recognition, Scene Understanding, Relation Understanding, and Semantic Segmentation (Tables 31–34). One-shot prompting partially mitigates these gaps (typically reducing drops to the -4.5 to -6.8 pp range), but substantial deficits remain (Tables 31–33). This pattern suggests that fine-tuning on English data can further specialize the model to English-centric lexical and instruction distributions, improving English performance more than non-English performance, rather than equalizing cross-lingual robustness.

Takeaway 3: Rationales and one-shot examples provide inconsistent post fine-tuning benefits across languages. After fine-tuning, adding rationales does not reliably reduce cross-lingual gaps. For several models and tasks, rationale prompting changes the cross-lingual deltas only marginally, and in some cases it can even worsen transfer (e.g., Qwen2VL-7B on Scene Understanding shifts from a positive gain to a small drop when moving from ZS w/o rat. to ZS w/ rat., Table 32). This aligns with the broader observation that rationale-style prompting primarily affects *how* a model verbalizes intermediate reasoning, and its benefits can be language-dependent, especially when the model's strongest reasoning and instruction-following competence is concentrated in English (Wei et al., 2022b).

Takeaway 4: The proposed *Unrelatedness* task remains brittle after fine-tuning. A salient trend is that *Unrelatedness* often degrades after fine-tuning, even for models that improve on alignment-positive tasks such as ITM. Notably, Qwen2VL-7B exhibits a sizeable drop on Japanese Unrelatedness in the strictest setting (ZS w/o rat., about -5.0 pp, Table 37), and LLaVa-Mistral-7B shows even larger drops (up to -12.6 pp, Table 37). This behavior is consistent with the intuition that fine-tuning on standard instruction data and predominantly *matching-positive* multimodal supervision can reinforce a model's tendency to align text with the image, making it harder to *reject* plausible but irrelevant textual spans. More broadly, prior work has shown that VLMs can be fragile under hard negative or confounder-style evaluation, including contrastive grounding settings and adversarial caption perturbations (Thrush et al., 2022; Hsieh et al., 2023), and that hallucination and overcommitment can persist even when models appear strong on standard benchmarks (Li et al., 2023b). Our post fine-tuning Unrelatedness results reinforce the need to explicitly evaluate, and potentially train for, *negative multimodal alignment*.

Takeaway 5: Fine-tuning does not resolve the Swahili and Urdu bottleneck for open models. Finally, a practical limitation remains unchanged: for several open models, Swahili and Urdu evaluation (and therefore fine-tuning transfer) is not available in our setup, which is reflected by the missing entries in Tables 31–38. This highlights a broader ecosystem constraint: progress on multilingual robustness depends not only on prompt design or English fine-tuning, but also on model-side multilingual coverage, tokenization support, and the presence of high-quality multilingual image-text resources (**?**Pfeiffer et al., 2022; Chen et al., 2023b).

Model	ZS w/o rat.			ZS w/ rat.			OS w/o rat.			OS w/ rat.		
	Ja	Sw	Ur	Ja	Sw	Ur	Ja	Sw	Ur	Ja	Sw	Ur
LlaVa-Mistral-7B	–12.1	0.0	0.0	–12.9	0.0	0.0	–6.4	0.0	0.0	–5.8	0.0	0.0
Qwen2VL-7B	–0.6	0.0	0.0	+0.6	0.0	0.0	0.0	0.0	0.0	–0.3	0.0	0.0
PALO-7B	+0.7	0.0	0.0	+0.2	0.0	0.0	–0.3	0.0	0.0	–0.2	0.0	0.0
MAYA-8B	+0.3	0.0	0.0	+0.4	0.0	0.0	–0.1	0.0	0.0	–0.2	0.0	0.0
LlaVa-Mistral-13B	+8.4	0.0	0.0	+8.5	0.0	0.0	+0.2	0.0	0.0	+0.3	0.0	0.0
PALO-13B	–0.4	0.0	0.0	+0.4	0.0	0.0	–0.6	0.0	0.0	–0.1	0.0	0.0

Table 31: Δ English accuracy on **Object Recognition** under fine-tuning (En→Ja, En→Sw, En→Ur).

Model	ZS w/o rat.			ZS w/ rat.			OS w/o rat.			OS w/ rat.		
	Ja	Sw	Ur	Ja	Sw	Ur	Ja	Sw	Ur	Ja	Sw	Ur
LlaVa Mistral 7B	-11.7	0.0	0.0	-11.8	0.0	0.0	-6.3	0.0	0.0	-4.5	0.0	0.0
Qwen2VL 7B	+2.0	0.0	0.0	-0.3	0.0	0.0	-0.5	0.0	0.0	-0.3	0.0	0.0
PALO 7B	-0.8	0.0	0.0	-0.2	0.0	0.0	-0.6	0.0	0.0	-0.2	0.0	0.0
MAYA 8B	-0.4	0.0	0.0	-0.2	0.0	0.0	-0.1	0.0	0.0	-0.8	0.0	0.0
LlaVa Mistral 13B	+0.0	0.0	0.0	+0.0	0.0	0.0	-0.2	0.0	0.0	+0.0	0.0	0.0
PALO 13B	-0.5	0.0	0.0	+0.2	0.0	0.0	-0.3	0.0	0.0	+0.4	0.0	0.0

Table 32: Δ English accuracy on **Scene Understanding** under fine-tuning (En→Ja, En→Sw, En→Ur).

Model	ZS w/o rat.			ZS w/ rat.			OS w/o rat.			OS w/ rat.		
	Ja	Sw	Ur	Ja	Sw	Ur	Ja	Sw	Ur	Ja	Sw	Ur
LlaVa Mistral 7B	-11.2	0.0	0.0	-11.4	0.0	0.0	-6.8	0.0	0.0	-5.9	0.0	0.0
Qwen2VL 7B	+0.7	0.0	0.0	+0.7	0.0	0.0	-0.5	0.0	0.0	-0.7	0.0	0.0
PALO 7B	-0.5	0.0	0.0	-0.2	0.0	0.0	+0.3	0.0	0.0	+0.4	0.0	0.0
MAYA 8B	-0.1	0.0	0.0	-0.2	0.0	0.0	+0.5	0.0	0.0	+0.7	0.0	0.0
LlaVa Mistral 13B	+0.0	0.0	0.0	+0.0	0.0	0.0	-0.2	0.0	0.0	-0.8	0.0	0.0
PALO 13B	-0.5	0.0	0.0	+0.4	0.0	0.0	-0.6	0.0	0.0	+1.3	0.0	0.0

Table 33: Δ English accuracy on **Relation Understanding** under fine-tuning (En→Ja, En→Sw, En→Ur).

Model	ZS w/o rat.			ZS w/ rat.			OS w/o rat.			OS w/ rat.		
	Ja	Sw	Ur	Ja	Sw	Ur	Ja	Sw	Ur	Ja	Sw	Ur
LlaVa Mistral 7B	-12.1	0.0	0.0	-11.3	0.0	0.0	-6.7	0.0	0.0	-6.5	0.0	0.0
Qwen2VL 7B	+0.3	0.0	0.0	-0.3	0.0	0.0	-0.5	0.0	0.0	+1.3	0.0	0.0
PALO 7B	-0.5	0.0	0.0	-0.1	0.0	0.0	+0.3	0.0	0.0	-0.6	0.0	0.0
MAYA 8B	-0.1	0.0	0.0	-0.2	0.0	0.0	-0.7	0.0	0.0	-0.5	0.0	0.0
LlaVa Mistral 13B	+0.0	0.0	0.0	+0.0	0.0	0.0	-0.2	0.0	0.0	+0.0	0.0	0.0
PALO 13B	-0.3	0.0	0.0	-0.3	0.0	0.0	-0.1	0.0	0.0	+0.3	0.0	0.0

Table 34: Δ English accuracy on **Semantic Segmentation** under fine-tuning (En→Ja, En→Sw, En→Ur).

Model	ZS w/o rat.			ZS w/ rat.			OS w/o rat.			OS w/ rat.		
	Ja	Sw	Ur	Ja	Sw	Ur	Ja	Sw	Ur	Ja	Sw	Ur
LlaVa-Mistral-7B	6.5	0.0	0.0	6.3	0.0	0.0	2.1	0.0	0.0	1.9	0.0	0.0
Qwen2VL-7B	+5.5	0.0	0.0	+5.1	0.0	0.0	+5.4	0.0	0.0	+4.6	0.0	0.0
PALO-7B	+2.4	0.0	0.0	+2.6	0.0	0.0	+2.5	0.0	0.0	+5.1	0.0	0.0
MAYA-8B	–0.5	0.0	0.0	–2.5	0.0	0.0	–0.7	0.0	0.0	–0.7	0.0	0.0
LlaVa-Mistral-13B	–8.9	0.0	0.0	0.0	0.0	0.0	–0.6	0.0	0.0	–0.8	0.0	0.0
PALO-13B	–13.0	0.0	0.0	+1.3	0.0	0.0	–0.3	0.0	0.0	+4.5	0.0	0.0

Table 35: Δ English accuracy on **Image Captioning** under fine-tuning (En→Ja, En→Sw, En→Ur).

Model	ZS w/o rat.			ZS w/ rat.			OS w/o rat.			OS w/ rat.		
	Ja	Sw	Ur	Ja	Sw	Ur	Ja	Sw	Ur	Ja	Sw	Ur
LlaVa Mistral 7B	+0.5	0.0	0.0	-0.5	0.0	0.0	-6.2	0.0	0.0	-6.3	0.0	0.0
Qwen2VL 7B	+5.3	0.0	0.0	+4.9	0.0	0.0	+5.0	0.0	0.0	+4.6	0.0	0.0
PALO 7B	+2.4	0.0	0.0	+2.4	0.0	0.0	+2.5	0.0	0.0	+5.3	0.0	0.0
MAYA 8B	-0.1	0.0	0.0	-8.7	0.0	0.0	-0.1	0.0	0.0	-5.0	0.0	0.0
LlaVa Mistral 13B	-8.9	0.0	0.0	+0.0	0.0	0.0	-0.5	0.0	0.0	-0.7	0.0	0.0
PALO 13B	-0.9	0.0	0.0	+0.3	0.0	0.0	-0.6	0.0	0.0	-0.1	0.0	0.0

Table 36: Δ English accuracy on **Image-Text Matching** under fine-tuning (En→Ja, En→Sw, En→Ur).

Model	ZS w/o rat.			ZS w/ rat.			OS w/o rat.			OS w/ rat.		
	Ja	Sw	Ur	Ja	Sw	Ur	Ja	Sw	Ur	Ja	Sw	Ur
LlaVa Mistral 7B	-12.6	0.0	0.0	-7.6	0.0	0.0	-6.0	0.0	0.0	-6.3	0.0	0.0
Qwen2VL 7B	-5.0	0.0	0.0	-0.5	0.0	0.0	-0.5	0.0	0.0	-0.3	0.0	0.0
PALO 7B	-2.1	0.0	0.0	-0.3	0.0	0.0	-0.6	0.0	0.0	+4.9	0.0	0.0
MAYA 8B	-0.1	0.0	0.0	-2.5	0.0	0.0	-0.6	0.0	0.0	-0.9	0.0	0.0
LlaVa Mistral 13B	-8.4	0.0	0.0	-0.6	0.0	0.0	-0.5	0.0	0.0	-0.7	0.0	0.0
PALO 13B	-0.7	0.0	0.0	+0.3	0.0	0.0	-0.7	0.0	0.0	-0.3	0.0	0.0

Table 37: Δ English accuracy on **Unrelatedness** under fine-tuning (En→Ja, En→Sw, En→Ur).

Model	ZS w/o rat.			ZS w/ rat.			OS w/o rat.			OS w/ rat.		
	Ja	Sw	Ur	Ja	Sw	Ur	Ja	Sw	Ur	Ja	Sw	Ur
LlaVa Mistral 7B	-6.7	0.0	0.0	-6.5	0.0	0.0	-3.0	0.0	0.0	-2.6	0.0	0.0
Qwen2VL 7B	+2.0	0.0	0.0	+4.2	0.0	0.0	+5.4	0.0	0.0	+5.3	0.0	0.0
PALO 7B	-0.4	0.0	0.0	+1.3	0.0	0.0	+0.4	0.0	0.0	-0.7	0.0	0.0
MAYA 8B	-0.5	0.0	0.0	-5.0	0.0	0.0	-0.7	0.0	0.0	-0.3	0.0	0.0
LlaVa Mistral 13B	-0.4	0.0	0.0	+0.0	0.0	0.0	-0.4	0.0	0.0	-0.8	0.0	0.0
PALO 13B	-0.6	0.0	0.0	+1.3	0.0	0.0	-0.6	0.0	0.0	-0.5	0.0	0.0

Table 38: Δ English accuracy on **Visual Question Answering** under fine-tuning (En→Ja, En→Sw, En→Ur).

K.3 Cross-lingual Accuracy with Japanese Input and English Output (Ja→En)

Next, we change the *input* language to Japanese while keeping the *output* in English, and we quantify how much performance changes relative to the English-input condition. Table 39 reports per-task deltas under the four prompting settings. For clarity in interpretation, we treat **negative** values as a **drop under Japanese input** (worse than English input), and **positive** values as a **gain under Japanese input** (better than English input).

Model	OR	SU	RU	SS	IC	ITM	U	VQA
Zero-shot, Without Rationales								
GPT-4o	-4.1	-3.8	-4.5	-5.0	-3.9	-3.5	-4.0	-4.0
GPT-4o-mini	-3.9	-3.9	-3.5	-3.2	-4.0	-3.3	-3.5	-3.5
Gemini 2.0	-3.8	-3.8	-3.7	-2.7	-3.7	-3.0	-3.0	-3.0
Gemini 1.5	-3.5	-3.8	-4.2	-3.3	-3.8	-4.1	-4.0	-4.0
LlaVa Mistral 7B	-14.3	-13.5	-11.8	-13.9	-8.6	-13.6	-8.5	-8.7
Qwen2VL 7B	-3.5	-4.6	-3.1	-4.1	+1.9	+2.4	+3.0	+2.1
PALO 7B	-3.7	-3.4	-3.0	-4.0	+1.7	+0.9	+2.3	+2.3
MAYA 8B	-4.0	-3.7	-2.8	-4.0	-3.3	-4.1	-4.1	-4.2
LlaVa Mistral 13B	-16.3	-15.9	-17.0	-15.0	-15.9	-16.4	-16.4	-17.0
PALO 13B	-3.8	-3.8	-4.4	-4.0	-3.8	-4.0	-4.7	-4.1
Zero-shot, With Rationales								
GPT-4o	-3.8	-4.0	-3.4	-3.7	-2.9	-3.6	-3.5	-3.4
GPT-4o-mini	-4.3	-3.7	-3.7	-3.0	-2.5	-3.6	-3.6	-3.9
Gemini 2.0	-3.9	-3.7	-3.7	-3.0	-2.9	-3.7	-3.5	-3.0
Gemini 1.5	-3.7	-3.7	-3.7	-3.0	-2.9	-4.0	-3.5	-3.0
LlaVa Mistral 7B	-14.5	-13.6	-13.0	-12.5	-9.0	-12.9	-14.1	-9.3
Qwen2VL 7B	-3.7	-4.0	-3.5	-3.0	+1.7	+2.0	+2.3	+2.7
PALO 7B	-3.6	-5.0	-3.7	-4.3	+2.1	+1.7	+1.7	+1.9
MAYA 8B	-3.9	-3.8	-3.5	-4.1	-3.3	-3.6	-3.6	-3.5
LlaVa Mistral 13B	-16.0	-15.4	-17.3	-15.9	-15.8	-15.6	-15.9	-17.0
PALO 13B	-3.5	-4.2	-3.6	-4.2	-3.7	-4.3	-3.6	-3.2
One-shot, Without Rationales								
GPT-4o	-4.1	-3.8	-3.5	-4.1	-3.3	-3.6	-3.7	-3.9
GPT-4o-mini	-4.2	-3.5	-3.5	-4.2	-3.2	-3.3	-3.8	-3.5
Gemini 2.0	-3.8	-3.3	-3.5	-4.0	-2.6	-3.0	-2.9	-3.5
Gemini 1.5	-4.5	-3.8	-3.1	-5.1	-3.9	-4.2	-3.7	-3.8
LlaVa Mistral 7B	-14.8	-13.7	-13.5	-13.8	-9.3	-14.1	-14.0	-9.0
Qwen2VL 7B	-4.0	-3.5	-3.6	-4.0	-2.4	-4.0	-4.0	-3.9
PALO 7B	-3.5	-3.5	-3.0	-4.2	-3.6	-4.1	-4.3	-3.8
MAYA 8B	-4.0	-3.8	-4.7	-3.8	-3.8	-4.2	-3.8	-3.9
LlaVa Mistral 13B	-16.4	-15.2	-15.2	-15.7	-15.5	-15.9	-16.0	-15.9
PALO 13B	-3.8	-4.1	-3.6	-4.3	-3.7	-4.2	-4.0	-4.0
One-shot, With Rationales								
GPT-4o	-3.7	-3.9	-4.4	-3.6	-3.8	-3.8	-3.8	-3.7
GPT-4o-mini	-4.0	-3.5	-3.7	-3.8	-3.0	-3.6	-3.5	-3.7
Gemini 2.0	-4.1	-4.0	-3.7	-3.6	-3.0	-3.7	-3.6	-3.7
Gemini 1.5	-3.7	-3.8	-3.5	-3.7	-3.0	-4.0	-3.5	-3.8
LlaVa Mistral 7B	-14.8	-12.7	-13.5	-12.7	-9.2	-14.4	-13.5	-8.3
Qwen2VL 7B	-4.9	-3.8	-3.6	-3.2	-3.0	-3.8	-3.4	-3.0
PALO 7B	-3.2	-3.7	-3.7	-4.0	-4.2	-4.0	-3.5	-3.5
MAYA 8B	-3.5	-3.7	-3.7	-4.1	-4.1	-3.3	-3.6	-4.0
LlaVa Mistral 13B	-16.0	-10.0	-17.3	-15.9	-15.8	-15.6	-15.9	-17.0
PALO 13B	-3.8	-4.1	-3.6	-4.3	-4.0	-4.3	-3.6	-3.8

Table 39: Cross-lingual drop in English accuracy when switching from English input to Japanese input (Ja→En). $\Delta_{\text{En}} = \text{Acc}_{\text{En-input}} - \text{Acc}_{\text{Ja-input}}$. Positive Δs in blue, negative Δs in red.

Takeaway 1: Closed-source models show a consistent, moderate Japanese-input penalty across tasks. Across GPT-4o, GPT-4o-mini, Gemini 2.0, and Gemini 1.5, switching from English to Japanese inputs yields a relatively uniform drop of roughly 3 to 5 percentage points on most tasks, across all prompting regimes. This pattern holds for both image-only tasks (OR, SU, RU, SS, IC) and image-text reasoning tasks (ITM, U, VQA), suggesting that the primary bottleneck is not task-specific visual reasoning, but rather the language interface that conditions the multimodal reasoning pipeline. Such language sensitivity is aligned with prior evidence that multilingual VLM performance depends strongly on the coverage and balance of multilingual pretraining data and evaluation settings.

Takeaway 2: The largest drops for closed-source models tend to appear on structure-heavy visual tasks. Although the closed-source models degrade fairly uniformly, their worst drops often appear on tasks that require structured visual grounding or fine-grained compositional reasoning, particularly Semantic Segmentation (SS) and Relation Understanding (RU). For example, GPT-4o exhibits one of its largest penalties on SS in the zero-shot setting, and comparable penalties on RU across settings (Table 39). This is consistent with the intuition that any degradation in instruction comprehension can cascade more severely for tasks that require precise grounding decisions rather than a single coarse classification.

Takeaway 3: Rationales and one-shot examples do not reliably reduce the Japanese-input gap. Adding rationales sometimes slightly shrinks the gap for particular model-task pairs, but the overall effect is inconsistent across tasks and models. Similarly, one-shot prompting does not systematically improve robustness under Japanese inputs, and in several cases the Japanese-input penalty remains nearly unchanged relative to zero-shot. This aligns with the broader observation that rationale prompting primarily changes the *reasoning trace* a model produces, and its benefits can be language-dependent when the model's strongest instruction-following behavior is concentrated in English.

Takeaway 4: Open models split into two regimes, catastrophic degradation vs. partial robustness. The LLaVa-Mistral family exhibits very large drops under Japanese input, reaching double-digit declines across most tasks (for example, around 14 to 17 points for LLaVa-Mistral 13B across tasks and settings). This indicates limited Japanese robustness in these models, plausibly driven by English-heavy instruction tuning, weaker Japanese tokenization coverage, or insufficient multilingual vision-language supervision. In contrast, several other open models (for example, PALO and MAYA variants) show smaller, more closed-source-like penalties, typically in the 3 to 5 point range (Table 39), consistent with the view that multilingual pretraining diversity is a key determinant of cross-lingual robustness.

Takeaway 5: Qwen2VL and PALO show an intriguing zero-shot anomaly on text-grounded tasks, but it disappears with longer prompts. Under *zero-shot* prompting, Qwen2VL-7B and PALO-7B exhibit **positive** deltas on IC, ITM, U, and VQA, while still showing **negative** deltas on the more purely visual tasks (OR, SU, RU, SS). This suggests that Japanese prompts can sometimes elicit better text-image alignment behavior in short-context conditions. However, once we move to *one-shot* prompting (with or without rationales), these gains largely vanish, and the deltas become negative again. A plausible explanation is that longer Japanese contexts, such as demonstrations and rationales, increase linguistic load and expose limitations in Japanese instruction-following, even when short Japanese prompts are manageable. This reinforces that multilingual robustness cannot be inferred from a single prompting regime, and should be evaluated across multiple context lengths and instruction styles (Pfeiffer et al., 2022; Chen et al., 2023b).

Takeaway 6: Unrelatedness remains challenging, and is sensitive to Japanese prompting, especially for weaker multilingual models. The proposed Unrelatedness (U) task shows drops comparable to other tasks for strong closed-source models, but it degrades sharply for LLaVa-Mistral 7B when rationales are included. This is consistent with prior findings that VLMs are brittle under hard-negative or confounder-style evaluations, where models must reject plausible but incorrect alignments, rather than simply find a best match (Thrush et al., 2022; Hsieh et al., 2023; Li et al., 2023b). Overall, the Japanese-input setting strengthens our central motivation for including Unrelatedness as a first-class capability, because language shifts can exacerbate a model's tendency to over-align text with images.

Summary. In the Ja→En setting, closed-source models exhibit a stable but non-trivial Japanese-input penalty, open models vary widely from catastrophic degradation to moderate robustness, and prompting interventions do not consistently close the gap. These results further support the need for multilingual VLM training and evaluation paradigms that explicitly stress-test language robustness and negative alignment behaviors, rather than assuming English-centric performance generalizes across input languages (Geigle et al., 2024; Pfeiffer et al., 2022; Chen et al., 2023b).

K.4 Cross lingual accuracy with Swahili input and English output (Sw→En)

Model	OR	SU	RU	SS	IC	ITM	U	VQA
Zero-shot, Without Rationales								
GPT-4o	-4.5	-4.7	-4.6	-5.1	-6.9	-5.3	-4.8	-4.2
GPT-4o-mini	-4.5	-4.1	-4.3	-4.1	-5.4	-4.2	-5.2	+0.8
Gemini 2.0 Flash	-5.2	-5.4	-5.1	-5.0	-5.0	-4.2	-5.0	+0.5
Gemini 1.5 Flash	-5.4	-5.2	-5.2	-5.1	-5.3	-5.0	-5.5	+0.6
LlaVa Mistral 7B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Qwen2VL 7B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
PALO 7B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MAYA 8B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
LlaVa Mistral 13B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
PALO 13B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Zero-shot, With Rationales								
GPT-4o	-5.1	-5.0	-4.8	-4.6	-4.7	-5.0	-5.0	-4.2
GPT-4o-mini	-4.9	-4.9	-5.1	-4.4	-0.7	-5.0	-3.6	-4.5
Gemini 2.0 Flash	-4.8	-5.0	-4.8	-4.3	-0.5	-4.9	-3.5	-3.0
Gemini 1.5 Flash	-5.0	-5.0	-4.8	-5.2	-0.8	-5.1	-3.9	-3.0
LlaVa Mistral 7B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Qwen2VL 7B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
PALO 7B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MAYA 8B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
LlaVa Mistral 13B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
PALO 13B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
One-shot, Without Rationales								
GPT-4o	-5.1	-4.0	-4.4	-4.5	-5.2	-5.1	-5.0	-5.6
GPT-4o-mini	-4.5	-4.3	-4.2	-4.3	-4.0	-5.4	-5.0	-3.5
Gemini 2.0 Flash	-4.8	-4.0	-4.9	-5.0	-4.0	-5.0	-5.0	-3.5
Gemini 1.5 Flash	-4.5	-5.1	-4.7	-5.1	-3.9	-4.2	-4.2	-3.8
LlaVa Mistral 7B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Qwen2VL 7B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
PALO 7B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MAYA 8B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
LlaVa Mistral 13B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
PALO 13B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
One-shot, With Rationales								
GPT-4o	-5.0	-4.2	-4.0	-4.6	-5.0	-4.5	-4.5	-3.6
GPT-4o-mini	-5.0	-2.0	-2.6	-4.7	-4.0	-4.9	-5.8	-4.0
Gemini 2.0 Flash	-4.3	-2.1	-2.7	-3.8	-3.6	-4.4	-5.5	-3.4
Gemini 1.5 Flash	-3.7	-3.0	-2.7	-3.7	-3.0	-4.0	-5.2	-2.7
LlaVa Mistral 7B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Qwen2VL 7B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
PALO 7B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MAYA 8B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
LlaVa Mistral 13B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
PALO 13B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table 40: Cross-lingual drop in English accuracy when switching from English-input to Swahili-input (Sw→En). $\Delta_{\text{En}} = \text{Acc}_{\text{En-input}} - \text{Acc}_{\text{Sw-input}}$. Positive Δs in blue, negative Δs in red.

Table 40 reports how model accuracy changes when we keep the output language fixed to English, but switch the *input* language from English to Swahili under four prompting regimes. Across the four large, closed-source models, the dominant pattern is a consistent *performance degradation* under Swahili inputs, typically on the order of a few percentage points per task. This aligns with prior evidence that multilingual multimodal competence remains uneven across languages, especially for lower-resource languages that are underrepresented in large-scale vision–language pretraining and evaluation pipelines (Geigle et al., 2024; Chen et al., 2023b; Pfeiffer et al., 2022).

Swahili inputs induce broad drops, with the largest penalties on language-sensitive tasks. In the zero-shot setting without rationales, GPT-4o drops on *every* task, with particularly large decreases for Image Captioning (IC, -6.9) and Image-Text Matching (ITM, -5.3), compared to smaller but still substantial decreases on perception-heavy tasks like Object Recognition (OR, -4.5) and Semantic Segmentation (SS, -5.1). A similar pattern holds for GPT-4o-mini and the Gemini models, where most tasks fall in roughly the -4 to -5.5 range. This task profile is expected: IC and ITM require stronger language grounding and cross-lingual alignment between the prompt and the visual representation, while OR and SS can sometimes be solved with weaker reliance on the instruction language (Pfeiffer et al., 2022; Geigle et al., 2024).

Rationales provide mixed robustness gains, and do not consistently close the gap. Adding rationales does not uniformly reduce the Swahili penalty. For GPT-4o, drops remain consistently large across tasks in the zero-shot with rationales condition (roughly -4.2 to -5.1), suggesting that rationale prompting alone does not repair the underlying multilingual alignment issues. This is consistent with the general view that chain-of-thought style prompting can improve reasoning when the model already has strong language understanding, but is not guaranteed to fix cross-lingual representation gaps (Wei et al., 2022b). Interestingly, GPT-4o-mini and Gemini show an unusually small drop on Image Captioning under rationales (around -0.5 to -0.8), which may indicate that (i) IC is more image-driven than instruction-driven in your evaluation setup, or (ii) rationale prompts incidentally stabilize output structure for caption scoring. Because this effect is not mirrored across the other tasks, it should be discussed as a task-specific interaction rather than a general "rationales solve multilinguality" conclusion (Wei et al., 2022b; Chen et al., 2023b).

One-shot helps selectively, and the best robustness is typically one-shot with rationales for Gemini. Moving from zero-shot to one-shot without rationales does not systematically improve Swahili robustness: for GPT-4o, average drops remain large and even worsen for VQA (down to -5.6). By contrast, one-shot with rationales reduces the drop for Gemini on several core understanding tasks, for example SU and RU improve to about -2 to -3 (Gemini 2.0: SU -2.1, RU -2.7; Gemini 1.5: SU -3.0, RU -2.7). This suggests that demonstration plus rationale can partially mitigate instruction-language friction for some models and tasks, but the residual gap remains substantial, especially for Unrelatedness (U), which stays strongly negative across all four settings (Wei et al., 2022b; Pfeiffer et al., 2022).

Coverage limitations of open models remain a practical barrier for Swahili. All open-source models in Table 40 are reported as 0.0 across settings, which (in your experimental context) reflects missing support or missing evaluations rather than genuine invariance. This itself is an important robustness finding: beyond accuracy, multilingual deployment depends on basic model support for the language, including tokenization, instruction tuning coverage, and multimodal pretraining exposure, which have historically been skewed toward high-resource languages (Geigle et al., 2024; Chen et al., 2023b). We therefore interpret Swahili robustness as primarily evidenced by the closed-source models here, and we treat "no result" for open models as an actionable limitation and a motivation for future multilingual VLM work (Chen et al., 2023b; Pfeiffer et al., 2022).

K.5 Cross-lingual accuracy with Urdu input and English output (Ur→En)

Table 41 reveals a markedly different transfer profile for Urdu inputs than what we observed for Japanese and Swahili: for frontier closed models (GPT-4o / GPT-4o-mini / Gemini), the *core recognition-and-understanding tasks* exhibit only *mild cross-lingual sensitivity*, while the *language-heavy generation/verification tasks* show the largest volatility. Concretely, under **zero-shot without rationales**, OR/SU/RU/SS mostly remain within roughly -0.8 to -2.0 for the closed models, indicating that switching prompts from English to Urdu does not strongly disrupt their visual grounding pipeline for these tasks. At the same time, **VQA** behaves unusually well in this direction: GPT-4o (+1.6), GPT-4o-mini (+3.0), Gemini 2.0 Flash (+4.0), and Gemini 1.5 Flash (+3.5)

Model	OR	SU	RU	SS	IC	ITM	U	VQA
Zero-shot, Without Rationales								
GPT-4o	-1.5	-1.0	-1.1	-0.8	-0.5	-1.8	+2.4	+1.6
GPT-4o-mini	-1.7	-2.0	-2.0	-1.5	-2.7	-1.5	-1.0	+3.0
Gemini 2.0 Flash	-1.2	-1.2	-1.2	-1.2	-2.4	-1.5	-1.5	+4.0
Gemini 1.5 Flash	-1.6	-1.4	-1.3	-0.9	-2.8	-1.6	-1.7	+3.5
LlaVa Mistral 7B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Qwen2VL 7B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
PALO 7B	-2.1	-2.0	-2.1	-2.0	-9.8	-1.8	-11.1	-3.6
MAYA 8B	-2.4	-2.0	-2.0	-2.3	-8.9	-2.3	-10.2	-3.6
LlaVa Mistral 13B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
PALO 13B	-1.5	-2.5	-2.0	-2.3	-17.3	-2.0	-18.8	-12.9
Zero-shot, With Rationales								
GPT-4o	-1.3	+1.6	-1.6	-1.3	+2.0	-1.5	-2.0	-0.2
GPT-4o-mini	-1.5	+0.3	-1.9	-0.9	+4.2	-1.6	+2.3	-2.9
Gemini 2.0 Flash	-1.6	+1.4	-1.5	-1.5	+3.8	-1.6	+1.5	-2.7
Gemini 1.5 Flash	-1.7	+1.0	-1.6	-2.1	+5.0	-2.0	+2.3	-2.7
LlaVa Mistral 7B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Qwen2VL 7B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
PALO 7B	-2.5	-2.1	+2.0	-2.0	-3.1	-2.0	-5.5	-7.1
MAYA 8B	-1.2	-1.8	-1.8	-2.0	-2.4	-1.9	-5.0	-2.7
LlaVa Mistral 13B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
PALO 13B	-1.2	-1.9	-0.9	-1.9	-10.8	-2.6	-16.9	-19.0
One-shot, Without Rationales								
GPT-4o	-1.4	-1.7	-1.6	-1.1	-0.5	-1.2	+0.5	+0.4
GPT-4o-mini	-1.0	-1.6	-1.0	-1.0	-1.2	-1.0	+1.4	+1.8
Gemini 2.0 Flash	-2.0	-1.3	-1.0	-1.6	-1.6	-1.5	-1.4	-1.6
Gemini 1.5 Flash	-1.7	-1.0	-1.0	-0.9	-1.7	-1.7	-1.6	-1.9
LlaVa Mistral 7B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Qwen2VL 7B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
PALO 7B	-2.0	-2.1	-2.0	-2.3	-1.1	-2.0	-2.0	-7.1
MAYA 8B	-1.3	-2.5	-2.4	-2.0	-2.8	-2.2	-1.8	-5.2
LlaVa Mistral 13B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
PALO 13B	-1.2	-2.5	-0.9	-1.9	-10.8	-2.5	-16.0	-19.0
One-shot, With Rationales								
GPT-4o	-1.0	-1.7	-2.1	-1.5	+2.3	-1.1	-0.7	+2.6
GPT-4o-mini	-1.5	+2.3	-1.9	-1.5	+2.1	-0.9	-3.0	-2.9
Gemini 2.0 Flash	-2.0	+1.4	-2.7	-1.6	+2.1	-1.5	+1.5	-3.4
Gemini 1.5 Flash	-1.7	+1.0	-2.0	-1.6	+5.0	-1.7	+1.9	-2.7
LlaVa Mistral 7B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Qwen2VL 7B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
PALO 7B	-2.0	-1.8	-2.1	-2.3	-3.1	-2.0	-5.5	-7.1
MAYA 8B	-1.2	-1.9	-1.8	-2.0	-2.4	-1.9	-5.0	-2.7
LlaVa Mistral 13B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
PALO 13B	-1.2	-2.6	-0.9	-1.9	-10.8	-2.6	-16.9	-19.0

Table 41: Cross-lingual drop in English accuracy when switching from English-input to Urdu-input (Ur→En). $\Delta_{\text{En}} = \text{Acc}_{\text{En-input}} - \text{Acc}_{\text{Ur-input}}$. Positive Δs in blue, negative Δs in red.

show gains, suggesting that (at least for this benchmark) Urdu prompts can preserve, and sometimes even sharpen, question grounding for VQA compared to English prompts. This is an important counterpoint to the common assumption that non-Latin scripts inevitably incur uniform degradation: the bottleneck is not "non-English" per se, but task-dependent interactions between prompt language, model instruction priors, and the linguistic structure required by each task (cf. cross-lingual VQA observations in prior multilingual vision-language evaluation, e.g., (Pfeiffer et al., 2022)).

Rationales shift which tasks benefit, and can flip the direction of transfer. Adding rationales in **zero-shot** systematically *improves SU* for all closed models (e.g., GPT-4o +1.6, Gemini 2.0 Flash +1.4, Gemini 1.5 Flash +1.0) and yields strong gains on **IC** (GPT-4o +2.0; GPT-4o-mini +4.2; Gemini 1.5 Flash +5.0). This pattern is consistent with the idea that rationale-style prompting can act as an auxiliary scaffold that helps the model structure intermediate linguistic representations (including implicit translation/normalization steps) before producing the final English output (Wei et al., 2022b). However, the same rationale prompting *does not uniformly help* across tasks: VQA shifts from large gains in ZS w/o rat. to small drops in ZS w/ rat. for GPT-4o-mini and Gemini, and **Unrelatedness** becomes unstable (e.g., GPT-4o drops to -2.0 while GPT-4o-mini becomes +2.3). Overall, rationales appear to reallocate model capacity toward language-mediated deliberation, which benefits tasks where the target requires richer generation (IC) or structured scene reasoning (SU), but can interfere with compact decision-style outputs (some U/VQA regimes) when the rationale introduces extra degrees of freedom or distracts attention from the discriminative signal.

One-shot does not dominate zero-shot; it mainly reduces extremes for closed models. Under **one-shot without rationales**, the closed models become more conservative: most tasks cluster around small drops (≈ -1 to -2), and the large VQA gains largely disappear (Gemini 2.0 Flash becomes -1.6 on VQA; Gemini 1.5 Flash -1.9). With **one-shot with rationales**, the picture becomes model-specific rather than uniformly better: GPT-4o recovers strong gains on IC (+2.3) and VQA (+2.6), while GPT-4o-mini improves SU (+2.3) but shows a notable VQA drop (-2.9). This asymmetry suggests that "demonstrations + rationales" is not a universally safe recipe in cross-lingual prompting; rather, it amplifies each model's internal preference for how to use the extra context (translation-first vs. reasoning-first vs. pattern-matching), which can help one task family and hurt another (Wei et al., 2022b).

Open models: degenerate vs. brittle cross-lingual behavior. A striking signal in Table 41 is the presence of **exact 0.0 entries across all tasks** for LLaVa-Mistral (7B/13B) and Qwen2VL-7B. Given the improbability that every task is perfectly unchanged, these rows likely indicate a *degenerate evaluation outcome* (e.g., failed Urdu prompting, fallback behavior, or a pipeline artifact) and should be treated as *non-diagnostic* until verified with raw generations and per-sample logs. In contrast, the PALO/-MAYA family produces non-trivial deltas and exposes a clear failure mode: while OR/SU/RU/SS tend to degrade moderately (typically around -2), **IC and U collapse**, especially for PALO-7B/MAYA-8B in ZS w/o rat. (IC $-9.8/-8.9$; U $-11.1/-10.2$) and catastrophically for PALO-13B (IC -17.3; U -18.8; VQA -12.9). This separation between "visual recognition” vs. “language-conditioned generation/verification” indicates that, for these models, Urdu prompt handling is the dominant bottleneck when the task requires producing or judging English text with nuanced semantic constraints. In other words, cross-lingual transfer here fails primarily in the *language interface* (prompt understanding + English response formation), not in extracting visual evidence, precisely the kind of bottleneck multilingual VLM benchmarking aims to isolate (Geigle et al., 2024).

Takeaway. Ur→En is *not uniformly harder* than Ja→En or Sw→En: for strong closed models it is often *surprisingly robust*, with small drops on OR/SU/RU/SS and occasional gains on VQA/IC depending on prompting. The dominant risk factor is prompting strategy: rationales can substantially help SU/IC yet destabilize U/VQA, and one-shot context can either regularize or mislead depending on the model. For open models, the results underscore two distinct concerns, *evaluation degeneracy* (all-zeros rows) and *true cross-lingual brittleness* (large collapses on IC/U/VQA), both of which motivate careful auditing of language-specific prompting and generation traces before drawing final conclusions.

K.6 Fine-tuned Cross-lingual Accuracy with Japanese Input and English Output (Ja→En)

Table 42 reports the post fine-tuning cross-lingual gap, $\Delta_{\text{En}} = \text{Acc}^{\text{ft}}_{\text{En-input}} - \text{Acc}^{\text{ft}}_{\text{Ja-input}}$. Thus, $\Delta_{\text{En}} > 0$ means the model still performs better when prompted in English (a residual gap against Japanese prompts), while $\Delta_{\text{En}} < 0$ means Japanese prompting is competitive or even better after adaptation. Across tasks, we observe three consistent patterns.

1) Fine-tuning changes the “preferred prompt language”, but not uniformly across models. The LLaVa-Mistral family shows a strong shift toward Japanese prompting under zero-shot without rationales: LlaVa Mistral 7B remains strongly negative across all eight tasks (roughly -8 to -12 points), and LlaVa Mistral 13B is similarly negative (roughly -8 to -9 points). This indicates that, after fine-tuning, Japanese prompts can match or surpass English prompts for these models in this Ja→En setting. Such sharp reversals are consistent with the broader continual-learning literature, where specialization to a new input distribution can change performance tradeoffs on previously strong regimes (e.g., interference and forgetting effects) (McCloskey and

Model	OR	SU	RU	SS	IC	ITM	U	VQA
Zero-shot, Without Rationales								
LlaVa Mistral 7B	-11.5	-8.5	-11.2	-8.3	-8.0	-11.8	-8.9	-9.1
Qwen2VL 7B	+0.6	+5.1	+0.7	-0.3	+5.3	-0.5	+5.3	+5.0
PALO 7B	-0.7	+5.2	-0.5	-0.5	+5.1	-0.3	+5.2	+5.2
MAYA 8B	-0.3	-0.4	-0.1	-0.1	-0.3	+2.4	+2.4	-0.5
LlaVa Mistral 13B	-8.4	-8.5	-8.1	-8.9	-8.1	-8.7	-8.4	-9.1
PALO 13B	+0.4	+0.2	-0.5	-0.6	+0.4	-0.7	+0.3	-0.3
Zero-shot, With Rationales								
LlaVa Mistral 7B	-11.5	-6.3	-11.4	-11.3	-9.2	-11.4	-7.5	-6.7
Qwen2VL 7B	-0.3	+4.5	+4.2	+0.0	+4.4	-0.6	+4.5	+4.2
PALO 7B	-0.2	+4.8	+4.8	-0.2	+4.8	-0.5	+5.1	+5.2
MAYA 8B	-0.2	-0.2	-0.2	-0.2	-0.2	-0.5	-0.5	-0.4
LlaVa Mistral 13B	+0.0	+0.0	-0.4	+0.0	+0.0	+0.0	+0.0	-0.4
PALO 13B	-0.1	-0.3	-0.3	-0.3	-0.5	+1.3	-0.4	-0.4
One-shot, Without Rationales								
LlaVa Mistral 7B	-6.4	-1.5	-6.8	-6.7	-2.1	-6.2	-4.1	-1.8
Qwen2VL 7B	-0.6	+5.1	-0.5	-0.5	+5.2	-0.5	+5.0	+5.4
PALO 7B	-0.6	+4.7	-0.3	-0.5	+5.3	-0.4	+4.2	+5.1
MAYA 8B	-0.4	-0.4	-0.5	-0.7	-0.3	-0.6	-0.1	-0.4
LlaVa Mistral 13B	-0.2	-0.2	-0.2	-0.2	-0.2	+0.0	-0.5	-0.5
PALO 13B	-0.5	-0.5	-0.4	-0.7	-3.1	+11.9	-0.7	-0.3
One-shot, With Rationales								
LlaVa Mistral 7B	-5.8	-0.7	-5.9	-5.7	-0.7	-6.3	-0.7	-0.7
Qwen2VL 7B	**-0.4**	**+5.3**	**-0.7**	**+5.2**	**-0.5**	**+5.0**	**-0.3**	**+5.1**
PALO 7B	-0.2	-0.4	-0.5	-0.3	-0.4	-6.6	-0.7	-0.7
MAYA 8B	-0.8	-0.8	-0.9	-0.9	-0.7	-5.0	-0.9	-0.3
LlaVa Mistral 13B	-0.6	-0.8	-0.9	-0.7	-0.5	-0.8	-0.7	-0.7
PALO 13B	-0.4	-0.6	-0.4	-0.4	-0.5	-11.8	-0.6	-0.5

Table 42: Δ in English accuracy after fine-tuning, Ja→En. $\Delta_{\text{En}} = \text{Acc}^{\text{ft}}_{\text{En-input}} - \text{Acc}^{\text{ft}}_{\text{Ja-input}}$. Positive Δs in blue, negative Δs in red.

Cohen, 1989; Kirkpatrick et al., 2017). In contrast, Qwen2VL 7B and PALO 7B often retain large *positive* gaps on language-heavy tasks, especially Scene Understanding (SU), Image Captioning (IC), Unrelatedness (U), and VQA (typically around $+4$ to $+5$ points in zero-shot), showing that fine-tuning does not automatically make Japanese prompting equally effective for all architectures.

2) One-shot demonstrations can collapse large gaps that persist in pure zero-shot. For LlaVa Mistral 13B, the gap is dramatic in zero-shot without rationales (large negative deltas), but it nearly vanishes in one-shot without rationales (mostly around -0.2 to -0.5), suggesting that a single in-context example can calibrate output format and task intent across languages. This aligns with the general observation that in-context demonstrations often stabilize instruction following and reduce prompt ambiguity, especially when models have been instruction tuned (Wei et al., 2022a). In our table, the effect is model-dependent: Qwen2VL 7B and PALO 7B still show persistent positive gaps on SU, IC, U, and VQA even in one-shot, suggesting that their remaining failure modes are not only “prompt calibration” problems but also reflect deeper cross-lingual instruction-following limitations.

3) Rationales are not a monotonic fix, and sometimes amplify the cross-lingual gap. Rationale prompting is widely used because chain-of-thought style prompting can improve reasoning on many tasks (Wei et al., 2022b; Kojima et al., 2022). However, Table 42 shows that adding rationales can either reduce gaps (e.g., LlaVa Mistral 13B in zero-shot becomes close to 0 on many tasks) or amplify them (e.g., Qwen2VL 7B and PALO 7B show large positive deltas on SU, RU, IC, U, and VQA under zero-shot with rationales). This is consistent with evidence that generated explanations are not always faithful and can introduce additional degrees of freedom that hurt reliability under distribution shift (Turpin et al., 2023). Practically, this argues against treating "with rationales" as universally better in cross-lingual evaluation: rationales should be reported as a separate operating point, not as a default.

Task-level takeaway. Across models, Object Recognition (OR), Relation Understanding (RU), and Semantic Segmentation (SS) tend to exhibit smaller absolute gaps than SU, IC, U, and VQA, which are more instruction and language sensitive in our setup. Overall, the fine-tuned Ja→En results suggest that multilingual robustness after adaptation depends on *how* the model was instruction tuned and fine-tuned, a phenomenon well documented in instruction tuning for LLMs and multimodal instruction tuning for VLMs (Wei et al., 2022a; Liu et al., 2023c; Dai et al., 2023).

K.7 Fine-tuning for Cross-lingual Robustness: Urdu→English (Ur→En)

Table 43 reports $\Delta_{\text{En}} = \text{Acc}^{\text{ft}}_{\text{En-input}} - \text{Acc}^{\text{ft}}_{\text{Ur-input}}$, so *negative* values mean the fine-tuned model performs better when the input is Urdu than when the input is English, while *positive* values indicate the opposite. A first practical observation is that several baselines (LLaVa-Mistral-7B, Qwen2VL-7B, and LLaVa-Mistral-13B) show exact zeros across all tasks and prompting regimes. This pattern is unlikely to reflect true invariance across eight tasks, and we treat it as "no measurable or no reported cross-lingual change" in this setting, focusing the analysis on models with consistent non-zero signal (PALO and MAYA).

PALO-7B shows prompt-sensitive specialization to Urdu. Under zero-shot without rationales, PALO-7B exhibits uniformly negative deltas across all eight tasks (roughly -3 to -4). This is the cleanest signature in the table that fine-tuning can induce *Urdu-specialized* behavior: after training, Urdu inputs systematically outperform English inputs. However, this effect is not stable to prompting. Once we

Model	OR	SU	RU	SS	IC	ITM	U	VQA
Zero-shot, Without Rationales								
LlaVa Mistral 7B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Qwen2VL 7B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
PALO 7B	-3.0	-3.2	-3.8	-3.6	-3.7	-3.8	-3.3	-3.9
MAYA 8B	+4.7	+4.7	+4.7	+4.7	+5.7	+4.9	+4.5	+4.3
LlaVa Mistral 13B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
PALO 13B	+0.4	+0.2	-0.5	-0.6	+0.4	-0.7	+0.3	-0.3
Zero-shot, With Rationales								
LlaVa Mistral 7B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Qwen2VL 7B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
PALO 7B	-0.2	+4.8	-0.2	-0.2	-0.2	-0.6	+5.1	+5.2
MAYA 8B	-0.2	-0.2	-0.2	-0.2	-0.2	-0.3	-2.5	-0.4
LlaVa Mistral 13B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
PALO 13B	-0.1	-0.3	-0.2	-0.3	-0.3	-0.6	+1.3	-0.4
One-shot, Without Rationales								
LlaVa Mistral 7B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Qwen2VL 7B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
PALO 7B	-0.6	+4.7	-0.6	-0.5	+5.3	-0.4	+5.0	+5.1
MAYA 8B	-0.4	-0.5	-0.5	-0.7	-0.3	-0.6	-0.6	-0.7
LlaVa Mistral 13B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
PALO 13B	-0.5	-0.6	-0.4	-0.7	-0.7	+11.9	-0.7	-0.3
One-shot, With Rationales								
LlaVa Mistral 7B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Qwen2VL 7B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
PALO 7B	-0.2	+5.6	-0.4	+5.1	-0.5	+4.6	+5.3	+4.8
MAYA 8B	-0.8	-0.3	-0.9	-0.9	-0.8	-0.8	-5.0	-0.7
LlaVa Mistral 13B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
PALO 13B	-0.4	-0.3	-0.3	-0.6	-0.4	-11.8	-0.6	-0.5

Table 43: Change in English accuracy after fine-tuning on Urdu input (Ur → En). $\Delta_{\text{En}} = \text{Acc}^{\text{ft}}_{\text{En-input}} - \text{Acc}^{\text{ft}}_{\text{Ur-input}}$. Positive Δs in blue, negative Δs in red.

add rationales, the direction flips on several tasks, especially SU, U, and VQA, where deltas become strongly positive (around $+5$). The same inversion persists in one-shot with rationales, where SU, SS, ITM, U, and VQA again show large positive deltas. This is strong evidence that "reasoning scaffolds" are not language-neutral in practice, and that adding rationale-style instructions can introduce a code-switching and formatting mismatch that overwhelms the gains from Urdu fine-tuning. In other words, the model may learn Urdu task competence, but still be brittle to the language and structure of the *prompt wrapper* that elicits that competence.

MAYA-8B improves mainly when the prompt supplies structure. MAYA-8B displays the opposite pattern. In zero-shot without rationales, it has consistently large positive deltas (about $+4.3$ to $+5.7$), meaning Urdu inputs remain substantially worse than English despite fine-tuning. Yet with even mild prompt structure, the cross-lingual gap collapses: in one-shot without rationales, the deltas become small and mostly negative (around -0.3 to -0.7), and in one-shot with rationales, the model shows a pronounced Urdu advantage on the Unrelatedness task (-5.0). This suggests MAYA's fine-tuned Urdu competence is "latent" and requires either exemplars or explicit stepwise structure to be reliably activated. Importantly, this is a robustness result: the same fine-tuned parameters can look ineffective or effective depending purely on the inference protocol, so reporting only one prompt setting would be misleading.

PALO-13B is mostly stable, except for an extreme ITM flip. PALO-13B largely stays near zero (often within ± 0.7), indicating better retention of bilingual behavior and less dramatic specialization. The notable exception is ITM, which is extremely prompt-dependent: it shows a very large positive delta in one-shot without rationales ($+11.9$), but a very large negative delta in one-shot with rationales (-11.8). This sign flip implies that, for ITM under Urdu fine-tuning, rationales can act as a powerful "alignment stabilizer", possibly by forcing the model to externalize intermediate correspondences between the Urdu text and visual evidence before committing to a decision. This effect is large enough that ITM becomes the clearest task where rationale prompting changes not just the magnitude but the *direction* of cross-lingual transfer after fine-tuning.

Task-level takeaway: cross-lingual stability varies by task family. Across models, the most volatile tasks are the ones that heavily depend on structured generation or multimodal alignment cues (IC, ITM, U, VQA). In contrast, OR and RU are comparatively closer to zero in many regimes (except when a model strongly specializes, as in PALO-7B zero-shot without rationales). This pattern supports a concrete evaluation recommendation: after fine-tuning, cross-lingual claims should be made only after checking multiple prompt protocols, because the "Urdu advantage" can be real (negative deltas), but it can also be erased or reversed by seemingly benign changes such as adding rationales.

Overall conclusion. Fine-tuning can create genuine Urdu-input gains, but these gains are not automatically robust to prompting. Some models (PALO-7B) appear to over-specialize and become highly prompt-sensitive, consistent with the classic stability plasticity tension in continual adaptation, while others (MAYA-8B) require explicit inference-time scaffolding to expose the benefits of fine-tuning. Consequently, for ACL-level reporting, the key finding here is not just whether Δ_{En} is negative on average, but whether it remains negative across inference regimes, especially when rationales are introduced.

L Language support of Open Vision Language Models

Table 44 summarizes the practical **language coverage** of popular open-source Vision-Language Models by separating (i) the **base LLM's multilingual text capability** from (ii) the model's **actual vision-language (VL) alignment language support**. This distinction matters because strong multilingual text understanding does not automatically translate into strong multilingual VL performance: some models remain primarily English-aligned for image–text tasks unless they receive explicit multilingual VL fine-tuning, while others (e.g., Qwen2-VL and MAYA) report broader cross-lingual VL readiness across multiple languages.

Model	Base LLM	Language Support (VL Tasks)
LLaVA 7B/13B Mistral	Mistral (multilingual text)	Primarily English. VL performance in other languages varies without multilingual VL fine-tuning.
PALO 7B/13B	Llama-based (multilingual text)	Primarily English. VL performance in other languages depends on fine-tuning.
MAYA 8B	Aya-based (broad language support)	Strong support for English and Indic languages (Hindi, Tamil, Telugu, etc.). Extensive VL alignment.
Qwen2-VL	Qwen series (multilingual)	Strong: English, Chinese. Good support: Fr, Es, De, Ar, Ja, Kr, Sw, Ur.

Table 44: Language support and base LLMs for open-source vision-language models.

M Open-source VLM Pretraining Data

Table 48 presents an overview of the pretraining datasets used by the four open-source VLMs evaluated in this work: LLaVA 7B/13B Vicuna, PALO 7B/13B, MAYA 8B, and Qwen2-VL. The datasets are organized by task type, including text, dialogue caption, visual question answering (VQA), grounding, optical character recognition (OCR), image captioning, visual spatial reasoning, video question answering, image classification, and knowledge-grounded image QA. Notably, LLaVA and PALO share substantial overlap in their pretraining data, particularly in VQA tasks (VQA2, GQA, OKVQA, OCRVQA, A-OKVQA, DVQA, TextVQA, DocVQA, ChartQA) and grounding tasks (GRIT, VisualGenome). Qwen2-VL demonstrates the broadest coverage across task types, incorporating diverse datasets including multilingual sources (LAION-en/zh) and in-house data. MAYA 8B uniquely relies on the Aya multilingual dataset across all task cate-

Model	Generation Settings
GPT-4o	`gpt-4o-2024-08-06`, temp=0, max_tokens=1024, batch=100
GPT-4o-mini	`gpt-4o-mini-2024-07-18`, temp=0, max_tokens=1024, batch=100
Gemini 2.0 FL	`gemini-2.0-flash-lite`, temp=0, max_tokens=1024
Gemini 1.5 F 8B	`gemini-1.5-flash-8b`, temp=0, max_tokens=1024
LLaVA 7B	`llava-hf/llava-v1.6-vicuna-7b-hf`, max_tokens=1024
LLaVA 13B	`llava-hf/llava-v1.6-vicuna-13b-hf`, max_tokens=1024
MAYA 8B	`maya`, max_tokens=1024
PALO 7B	`palo`, max_tokens=1024
PALO 13B	`palo`, max_tokens=1024
Qwen2-VL 7B	`Qwen/Qwen2-VL-7B-Instruct`, max_tokens=1024

Table 45: Model generation settings.

gories, reflecting its focus on multilingual capabilities. The table reveals that while proprietary models benefit from extensive pretraining across multiple task types, the choice of pretraining data significantly influences each model's strengths in specific vision-language tasks and language support.

M.1 Details of Model Parameters and Finetuning Settings

The hyperparameters for VLMs used in the experiments in §5 are shown in Table 45.

M.1.1 Inference Settings

All the experiments are conducted on two NVIDIA A100 80GB GPUs. The complete list of VLMs used includes: GPT-4o, GPT-4o-mini, Gemini 2.0 Flash-Lite, Gemini 1.5 Flash-8B, LLaVA-NeXT-Vicuna-7B [4], LLaVA-NeXT-Vicuna-13B [5], PALO 7B/13B [6], MAYA 8B [7], and Qwen2VL 7B [8]. We deploy all the open models via the HuggingFace API, and employ 8-bit quantization for model generation. We deploy GPT *batch API*[9] settings for both `gpt-4o-2024-08-06` and `gpt-4o-mini-2024-07-18` models. Moreover, we deploy *concurrent processing* for Gemini models: `gemini-2.0-flash-lite`, `gemini-1.5-flash-8b`.

[4]https://huggingface.co/llava-hf/llava-v1.6-vicuna-7b-hf
[5]https://huggingface.co/llava-hf/llava-v1.6-vicuna-13b-hf
[6]https://github.com/mbzuai-oryx/PALO
[7]https://github.com/nahidalam/maya
[8]https://huggingface.co/Qwen/Qwen2-VL-7B-Instruct
[9]OpenIA batch API settings: https://platform.openai.com/docs/guides/batch?lang=python.

Dependency/Library	Version
accelerate	0.34.2
bitsandbytes	0.44.1
datasets	3.0.
deepspeed	0.15.1
CUDA	12.1
flash-attn	2.5.8
huggingface_hub	0.24.6
ninja	1.12.1
pillow	10.4.0
requests	2.32.3
safetensors	0.4.4
torch	2.4.0
torchaudio	2.4.0
tqdm	4.66.5
transformers	4.45.1
trl	0.10.1
wandb	0.17.7

Table 46: VLM experiment settings.

Language	Total Pairs	Train (80%)	Dev (10%)	Test (10%)
English (En)	1000	800	100	100
Japanese (Ja)	1000	800	100	100
Swahili (Sw)	1130	904	113	113
Urdu (Ur)	996	796	99	101

Table 47: Data splits for VLURes Finetuning. Counts are identical for image-text pairs.

M.1.2 Fine-tuning Settings

During fine-tuning, we set the number of training epochs to 10, per device train batch size to 2, we deploy gradient accumulation and gradient checkpointing, we use the Adam optimizer, and the learning rate is set to 2^{e-4}. We deploy a constant learning rate scheduler. Our LoRA configuration includes LoRA matrices of rank 8, and the dropout probability is set to 0.05. We conduct three fine-tuning and evaluation runs for each experiment and average the results over three runs. Details about major environment settings used during our experiments are shown in Table 46.

M.1.3 Finetuning Data Splits

Table 47 reports the image–text pair counts used to fine-tune the open VLMs in each language. For every language, we use a standard **80/10/10** split for **train/dev/test**, yielding 800/100/100 pairs for English and Japanese. Swahili contains slightly more data (1,130 pairs), resulting in 904/113/113 pairs, while Urdu contains 996 pairs and is split into 796/99/101 pairs (the test set differs by two examples due to integer rounding). This design keeps the training portion dominant while reserving sufficient held-out data for model selection (dev) and final reporting (test), and it ensures that cross-lingual fine-tuning comparisons are driven primarily by language effects rather than differences in split policy or evaluation size.

Data Type	Data Name	Model			
		LLaVA 7B/13B Vicuna	PALO 7B/13B	MAYA 8B	Qwen2-VL
Text	ShareGPT-4V (Chen et al., 2024)	✓	✓		✓
	LAION-GPT-V (gpt, 2023)	✓	✓		✓
	SlimOrca (Open-Orca / SlimOrca contributors, 2023)				✓
	Aya multilingual dataset (Singh et al., 2024)			✓	
	In-house Data				✓
Dialogue Caption	COCO (Lin et al., 2014)				✓
	TextCaps (Sidorov et al., 2020)				
	SBU (Ordonez et al., 2011)				✓
	DataComp (Gadre et al., 2023)				
	CC12M (Changpinyo et al., 2021) & 3M				
	LAION-en & zh (Schuhmann et al., 2022b)				✓
	LLaVA (Liu et al., 2023a,b)	✓			
	Aya multilingual dataset (Singh et al., 2024)			✓	
VQA	VQA2 (Goyal et al., 2017b)	✓	✓		✓
	GQA (Hudson and Manning, 2019)	✓	✓		✓
	OKVQA (Marino et al., 2019)	✓	✓		✓
	OCRVQA (Mishra et al., 2019)	✓	✓		
	A-OKVQA (Schwenk et al., 2022)	✓	✓		
	DVQA (Kafle et al., 2018)	✓	✓		✓
	TextVQA (Singh et al., 2019)	✓	✓		✓
	DocVQA (Mathew et al., 2021)	✓	✓		
	ChartQA (Masry et al., 2022)	✓	✓		
	A12D (Kembhavi et al., 2016)				✓
	Aya multilingual dataset (Singh et al., 2024)			✓	
Grounding	A12D (Kembhavi et al., 2016)				✓
	GRIT (Peng et al., 2023)	✓	✓		
	VisualGenome (Krishna et al., 2017b)	✓	✓		✓
	RefCOCO (Kazemzadeh et al., 2014)				✓
	RefCOCO+ (Kazemzadeh et al., 2014)				✓
	RefCOCOg (Mao et al., 2016)				✓
	LLaVA (Liu et al., 2023a,b)	✓			
	Aya multilingual dataset (Singh et al., 2024)			✓	
OCR	SynthDoG-en&zh (Kim et al., 2022)	✓	✓		✓
	DocVQA (Mathew et al., 2021)	✓	✓		
	Aya multilingual dataset (Singh et al., 2024)			✓	
Image Captioning	Web CapFilt (Li et al., 2022)				✓
	NoCaps (Agrawal et al., 2019)				✓
	LAION-CC-SBU subset (Liu et al., 2023b)	✓	✓		
	LLaVA-Instruct-150K (Liu et al., 2023a,b)	✓	✓		
	Aya multilingual dataset (Singh et al., 2024)			✓	
Visual Spatial Reasoning	Flickr30K (Young et al., 2014)				✓
	IconQA (Lu et al., 2021)				✓
	ChartQA (Masry et al., 2022)	✓	✓		
	A12D (Kembhavi et al., 2016)				✓
	Aya multilingual dataset (Singh et al., 2024)			✓	
Video Question Answering	MSVD-QA (Xu et al., 2017)				
	MSRVT-QA (Xu et al., 2017)				
	iVQA (Yang et al., 2022)				✓
	Aya multilingual dataset (Singh et al., 2024)			✓	
Image Classification	VizWiz (Gurari et al., 2018b)				✓
	Aya multilingual dataset (Singh et al., 2024)			✓	
Knowledge-Grounded Image QA	ScienceQA (Lu et al., 2022)	✓	✓		✓
	Aya multilingual dataset (Singh et al., 2024)			✓	

Table 48: Overview of pretraining datasets used by different models.

N Project Costs

N.1 API Costs: Inference

To ensure transparency and facilitate reproducibility, we provide a detailed analysis of the computational costs associated with our experiments. While our study includes several open-source models run on local infrastructure, this section focuses on the expenses incurred from using proprietary models accessed via their respective APIs: `gpt-4o`, `gpt-4o-mini`, `gemini-2.0-flash-lite`, and `gemini-1.5-flash-8b`. Our experimental design encompassed a comprehensive evaluation across our full benchmark, which includes four languages (English, Japanese, Swahili, and Urdu), four distinct prompting settings, and eight vision-language tasks, culminating in 32,000 API calls per model. Table 49 provides a granular breakdown of these expenses, detailing the pricing tiers, the total number of input and output tokens processed, and the resulting cost for each model. The cumulative estimated cost for this extensive inference stage totals approximately **$221.04**. We note that this figure exclusively covers the model inference phase of our study. A separate analysis detailing the costs associated with our LLM-as-a-judge evaluation framework is presented in the subsequent section. This detailed breakdown serves as a crucial data point for community, highlighting resource investment required for large-scale, multilingual VLM benchmarking.

N.2 API Costs: Evaluation

Following the inference stage, we employed an LLM-as-a-judge framework to systematically evaluate the quality of the generated outputs. This section details the costs associated with this evaluation phase, which was conducted using the `Gemini 1.5 Pro` model. The evaluation was comprehensive, covering the outputs from all ten models (both proprietary and open-source) across the full experimental matrix: eight tasks, four languages, and four prompting settings for each of our 1,000 benchmark items. This extensive scope resulted in 1,280,000 individual outputs requiring evaluation, with each judged by a separate API call. The input to the `Gemini 1.5 Pro` judge for each call comprised a detailed evaluation prompt (≈490 tokens) and the specific VLM output being assessed. The size of the VLM output varied depending on whether rationales were included (≈154 tokens without, ≈315 tokens with). Summing across all evaluation calls, the total input token count is estimated at 927.36 million. Conversely, the judge's output was designed for efficiency, consisting of a minimal JSON object containing only the score (≈5 tokens), leading to a total of 6.4 million output tokens. As detailed in Table 50, the total estimated cost for this LLM-as-a-judge evaluation phase is **$1,191.20**. While substantial, this investment was crucial for achieving robust, automated, and scalable evaluation of generative model outputs across vast and diverse landscape of our multilingual benchmark.

N.3 Total API Costs

The total computational cost for this study is composed of two primary phases: model inference and LLM-as-a-judge evaluation. The inference phase, detailed in Part 1, amounted to approximately $221.04. The subsequent evaluation phase, detailed in Part 2, incurred a significantly larger cost of approximately $1,191.20, primarily due to the vast number of individual outputs requiring judgment. Cumulatively, as summarized in Table 51, the total estimated expenditure for the project is **$1,412.24**. This figure underscores the substantial financial investment required to conduct comprehensive, large-scale, and multilingual benchmarking of modern foundation models using automated evaluation frameworks.

N.4 Human Evaluation Costs

We hired two native speakers per language (for En, Ja, Sw, and Ur) to rate the quality of VLM-generated outputs, for eight vision and language tasks. The evaluators followed clear guidelines to rate the performances of ten VLMs used in this study on a scale of 1 to 100. All the evaluators were compensated fairly based on prevailing market prices.

O Domain Names used for Data Collection

Table 52 shows the web resources we used to gather image-text pairs, for each language, in *VLURes*.

Table 52 summarizes the **domain-level provenance** of the image-text pairs used to build *VLURes*, organized by language. For **English**, we curate data from broad, high-coverage sources spanning encyclopedic and news-style content (*Wikipedia*, *Wikinews*, *BBC*) and community Q&A/discussion (*StackExchange*) (Atuhurra et al., 2024), ensuring diversity across factual descriptions and naturally occurring user-authored text. For **Japanese**, we intentionally widen source coverage beyond Wikipedia by including major public broadcasters and news outlets (*NHK*, *Yomiuri*, *Nikkei*), large-scale user platforms

Model	Price per 1M Tokens		Input Tokens	Output Tokens	Est. Cost
	Input ($)	Output ($)	(Millions)[a]	(Millions)[b]	($)
GPT-4o	1.2500	5.0000	33.68	30.02	$192.20
GPT-4o-mini	0.0750	0.3000	33.68	30.02	$11.54
Gemini 2.0 Flash	0.0750	0.3000	33.68	30.02	$11.54
Gemini 1.5 Flash	0.0375	0.1500	33.68	30.02	$5.76
Total			**Cumulative Cost:**		**$221.04**

Table 49: Estimated inference cost for proprietary models on the entire *VLURes* benchmark. Costs are based on Batch API pricing (Aug 2024) and are calculated across four languages, four prompting settings, and eight tasks for our 1,000-item benchmark.

Note: Open-source models were run on local infrastructure and are excluded from this API cost analysis.
[a] Total input tokens are estimated as 32,000 calls (1k items × 8 tasks × 4 languages) × (avg. 171 prompt tokens + avg. 93 text tokens) = 33,680,000 tokens per model.
[b] Total output tokens are estimated as 16,000 calls × 154 tokens (w/o rationales) + 16,000 calls × 315 tokens (w/ rationales) = 30,016,000 tokens per model.

Metric	Value
LLM-as-a-Judge Model	`Gemini 1.5 Pro`
Total Evaluation Calls	1,280,000
Input Token Calculation	
Avg. Input Tokens / Call (w/o rationales)	$\approx$ 644 tokens
Avg. Input Tokens / Call (w/ rationales)	$\approx$ 805 tokens
Total Input Tokens (Millions)[a]	927.36
Output Token Calculation	
Avg. Output Tokens / Call	$\approx$ 5 tokens
Total Output Tokens (Millions)[b]	6.40
Cost Calculation	
Input Cost ($)	$1,159.20
Output Cost ($)	$32.00
Total Estimated Evaluation Cost	**$1,191.20**

Table 50: Estimated cost for the LLM-as-a-judge evaluation phase using Gemini 1.5 Pro. The cost covers the evaluation of 1,280,000 individual model outputs from all 10 VLMs across the entire benchmark.

Note: Pricing is based on Gemini 1.5 Pro rates ($1.25/1M input$,5.00/1M output for prompts $\leq$ 128k tokens).
[a] Total input tokens = (640,000 calls $\times$ 644 tokens) + (640,000 calls $\times$ 805 tokens) = 927,360,000.
[b] Total output tokens = 1,280,000 calls $\times$ 5 tokens = 6,400,000.

Component	Estimated Cost ($)
Part 1: Model Inference	$221.04
Part 2: LLM-as-a-Judge Evaluation	$1,191.20
Total	**$1,412.24**

Table 51: Summary of total estimated project costs, combining both the model inference and the LLM-as-a-judge evaluation phases.

for everyday language and grounded entities (*Cookpad*, *Tabelog*, *Kakaku*), and domain-specific institutions that contribute specialized terminology and visual grounding, sports (*J.League*, *NPB*), culture/museums (*TNM*, *Tokyo Gendai*), public venues (*Tokyo Zoo*), and government/industry bodies (*METI*, *AIJ*). For **Swahili**, sources combine *Swahili Wikipedia* with prominent regional news and media portals (*BBC Swahili*, *Mwananchi*, *Taifa Leo*, *Bongo5*, *Michuzi*, *Swahili Times*), emphasizing naturally occurring references to local events, people, and places. For **Urdu**, we similarly mix *Urdu Wikipedia* with major news publishers (*Dawn*, *Express*, *BBC Urdu*) and popular informational sites (*UrduPoint*, *Kfoods*), capturing both formal reporting language and practical, everyday-domain content. Overall, the table clarifies that *VLURes* is constructed from a **deliberate blend of encyclopedic, news, and high-traffic community or lifestyle domains**, with language-specific additions chosen to increase topical breadth and visual grounding in each language.

Language	Domain Name
English	www.wikipedia.org www.wikinews.org www.stackexchange.com www.bbc.com
Japanese	ja.wikipedia.org www.nhk.or.Ja www.cookpad.com www.tabelog.com www.jleague.Ja www.npb.Ja www.tnm.Ja www.tokyogendai.com/ja/ www.nikkei.com www.tokyo-zoo.net www.aij.or.Ja www.yomiuri.co.Ja www.meti.go.Ja www.kakaku.com
Swahili	sw.wikipedia.org www.bbc.com/swahili www.michuzi.co.tz www.bongo5.com www.mwananchi.co.tz www.mwanahalisionline.com taifaleo.nation.co.ke swahilitimes.co.tz
Urdu	ur.wikipedia.org www.dawnnews.tv www.express.pk www.bbc.com/urdu www.urdupoint.com www.kfoods.com

Table 52: Overview of sources of data used to construct the ***VLURes*** benchmark, for each language.

P Human Evaluation Alignment and Prompts

P.1 Human Evaluation Alignment

We assessed the usefulness of *Gemini 1.5 Pro* as the main evaluator in our work. First, we selected 100 responses of GPT-4o-mini, from the *relation understanding* task, for appraisal by human evaluators (i.e., we chose two native speakers for each language, who are PhD students in the computer science area). Then, we recruited two human annotators to review the VLM responses and assign their scores for each response. The responses are graded on a scale from one to one hundred, the same as with the *Gemini 1.5 Pro judge*. We compare the agreement between their evaluation and the *Gemini 1.5 Pro judge*. By calculating the intraclass correlation coefficient, the agreement between human annotators and the LLM judge is 82.3%. The high level of agreement suggests *Gemini 1.5 Pro* performs comparably to human annotators in assessing the perception tasks.

We leave the full evaluation across all tasks for future work due to the high costs of recruiting human evaluators and payment of API access to VLMs.

P.2 Prompts used for Evaluation in this study

In what follows, we show several prompts we used throughout the study.

The Prompt for Gemini 1.5 Pro to Rate the Performance of VL Tasks

Below is a vision and language task with responses from both the vision language model and the ground-truth provided by human annotation. Based on the 3 criteria below, rate the model's performance on a scale of 1-100. Only provide the scores without explanations.

Accuracy: The response provided by the LLM is accurate and has no factual errors. Conclusions not made arbitrarily.

Helpfulness: The model's response provides clear information the task.

Linguistic Quality: The response is logical. The model correctly understands the task, and the expressions smooth and natural.

Please ensure that you do not let the length of the text influence your judgment, do not have a preference for any AI assistant names that might appear in the dialogue, do not let irrelevant linguistic habits in the conversation influence your judgment, and strive to remain objective. Your scoring should be strict enough and do not give a perfect score easily.

Human Evaluation Instructions

Based on the 3 criteria below, rate the model's performance on a scale of 1-100. Only provide the scores without explanations.

Accuracy: The response provided by the LLM is accurate and has no factual errors. Conclusions not made arbitrarily.

Helpfulness: The model's response provides clear information the task.

Linguistic Quality: The response is logical. The model correctly understands the task, and the expressions smooth and natural.

Please ensure that you do not let the length of the text influence your judgment, do not have a preference for any AI assistant names that might appear in the dialogue, do not let irrelevant linguistic habits in the conversation influence your judgment, and strive to remain objective. Your scoring should be strict enough and do not give a perfect score easily.

Components of our LLM-as-a-Judge Prompt

1. Role and Persona

The prompt begins by assigning a specific role to the LLM-judge: *"You are a meticulous evaluator of Vision-Language AI responses."* This sets the context and encourages a detailed, analytical mode of operation.

---

2. Core Task and Context

The judge is explicitly told what to evaluate, including the specific **task description**, the **language of the response**, and the **experimental setting** under which the response was generated (e.g., *zero-shot with rationales*).

---

3. Ground Truth Information

The prompt includes the ground-truth materials (**image-text pairs**) that the original VLM used to provide full context for the evaluation. This consists of:

- The original **text** content associated with the image (*e.g., "text_content"*).
- A reference to the **image** file itself (*e.g., "Image file: img859.Jaeg"*).

---

4. Candidate Response

The specific **VLM-generated output** that needs to be scored is clearly demarcated and presented to the judge, labeled with its unique **Text ID**.

---

5. Scoring Rubric and Scale

The judge is instructed to rate the response on a scale from 0 (lowest quality) to 100 (highest quality) based on three explicit criteria:

- **Accuracy:** The correctness and factual integrity of the response.
- **Helpfulness:** The clarity, practicality, and relevance of the assistance provided for the given task.
- **Linguistic Quality:** The logical coherence, naturalness, and correct understanding of the task.

---

6. Objectivity and Best Practices

A set of important guidelines is provided to minimize bias and ensure objective scoring. The judge is explicitly instructed to disregard response length, not to prefer specific AI assistant names, and to remain strict and objective in its scoring.

---

7. Strict Output Formatting

Finally, the prompt concludes with strict instructions for the output format to ensure programmatic parsing. The judge must return **only** a valid JSON object containing the numerical score, with no additional explanation: {"score": <number>}

Table 53: Structure of the Evaluation Prompt for the LLM-as-a-Judge, that is Gemini 1.5 Pro, used in scoring the outputs generated by VLMs in this study. Each component is designed to provide comprehensive context and ensure consistent, high-quality evaluation. The full prompt is shown in Table 54.

Evaluation Prompt for Gemini 1.5 Pro LLM Judge

System Persona and Core Task

You are a meticulous and objective AI expert specializing in the evaluation of Vision-Language Model (VLM) responses. Your task is to evaluate the quality of a VLM's response based on the provided context and scoring criteria.

Evaluation Context and Candidate Response

Please evaluate the following VLM response for the task: "{task_description}" in {language}.
The response was generated {settings_description} and is based on the following input image-text pair.

GROUND TRUTH CONTEXT:

- Image File: {image_info}
- Associated Text: "{text_info}"

CANDIDATE VLM RESPONSE TO EVALUATE (Text ID {text_id}):
"{response}"

Scoring Criteria and Output Format

Please rate the quality of the candidate response on a scale from 0 (lowest quality) to 100 (highest quality) based on the following three criteria:

1. **Accuracy:** The response is factually correct and logically sound. Conclusions are well-supported and not arbitrary.
2. **Helpfulness:** The response is clear, instructive, and directly addresses the specific vision-language task.
3. **Linguistic Quality:** The response correctly interprets the task, and the language is fluent, natural, and coherent.

Important Evaluation Guidelines:

- Remain strictly objective and do not let the length of the response influence your judgment.
- Do not show preference for any AI assistant names that might appear in the response.
- Strive to be strict in your scoring; do not award a perfect score easily.

REQUIRED OUTPUT FORMAT:
Return ONLY a single, valid JSON object with the score. Do not include any other text, explanations, or markdown formatting. Your entire response must be a single, valid JSON object.
Example: `{"score": 85}`

Table 54: The complete **Evaluation Prompt** for the **Gemini 1.5 Pro LLM-as-a-Judge**. This structured prompt was used for all 1.28 million evaluation calls. Placeholders (highlighted in blue) were programmatically filled with the relevant information for each specific VLM output being judged.

Q Four Prompt Settings used for Response Generation in this study

Prompt Setting 1: Zero-shot, No Rationales (in English)

```
LANGUAGE_CONFIGS = {
  "English": {
    "code": "En",
    "system_prompt": "You are an AI assistant that analyzes images and text.",

    "prompt_template_image_only": (
      "You are an intelligent assistant tasked with analyzing an image.",
      "Task description: {task_description}",
      "Provide only the analysis for this task, clearly labeled."
    ),

    "prompt_template_image_text": (
      "You are an intelligent assistant tasked with analyzing the relationship between an image and text.",
      "Examine both the image and the provided text carefully.",
      "Text associated with the image: {text_content}",
      "Task description: {task_description}",
      "Provide your analysis based on both sources, citing specific visual and textual evidence."
    ),

    "tasks": {
      1: "Analyze the image and list all objects present. Categorize each object (e.g., furniture, electronic devices, clothing) and be specific.",
      2: "Describe the overall scene. What is the setting, and what activities or events are taking place?",
      3: "Identify interactions or relationships between objects or entities. Explain spatial, functional, or social connections.",
      4: "Divide the image into semantic regions. Label each region (e.g., sky, buildings, people, street) and briefly describe its contents.",
      5: "Provide a detailed natural-language description of what is happening in the image, as if for someone who cannot see it.",
      6: "Extract parts of the text that directly reference entities, objects, or scenes in the image, and explain the visual evidence.",
      7: "Identify which parts of the text are not represented in the image and explain why they are unrelated.",
      8: "List places mentioned in the text or shown in the image. For each, indicate whether it appears in the text, the image, or both, and note if it is well known."
    }
  }
}
```

Table 55: Language configuration and prompts for ***zero-shot evaluation without rationales*** in English.

Prompt Setting 2: Zero-shot, With Rationales (in English)

```
LANGUAGE_CONFIGS = {
  "English": {
    "code": "En",

    "system_prompt": "You are an AI assistant that analyzes images and text.",

    "prompt_template_image_only": (
      {task_description},
      "Please provide your analysis along with your rationale. Explain step by step how you arrived at your answer."
    ),

    "prompt_template_image_text": (
      "Text associated with the image: {text_content}",
      {task_description},
      "Please provide your analysis and rationale. Explain step by step how you derived your response by referencing evidence from both the image and text."
    ),

    "tasks": {
      1: (
        "Let's analyze this image step by step to identify and categorize all objects present.",
        "Steps to follow: 1. Scan the image systematically (left to right, top to bottom). 2. List each identified object. 3. Group objects into categories (furniture, electronics, clothing, etc.). 4. Verify even small or partially visible objects are included.",
        "After these steps, provide your detailed answer and then your rationale explaining how you identified and categorized the objects. Think step by step."
      ),

      2: (
        "Let's analyze the scene step by step for a comprehensive description.",
        "Steps to follow: 1. Identify the primary setting or location. 2. Note the time of day and overall atmosphere. 3. Identify main activities or events occurring. 4. Observe any additional background activities.",
        "After these steps, provide your detailed description and then your rationale describing what visual cues led to your interpretation. Think step by step."
      ),

      3: (
        "Let's analyze the interactions and relationships in this image step by step.",
        "Steps to follow: 1. Identify every entity (object or person) in the image. 2. Examine spatial relationships among the entities. 3. Look for direct interactions. 4. Consider any functional or implied social connections.",
        "After these steps, provide your detailed analysis and then your rationale on how you established these relationships. Think step by step."
      ),

      4: (
        "Let's divide this image into semantic regions step by step.",
        "Steps to follow: 1. Identify major spatial divisions in the image. 2. Determine the function or nature of each region. 3. Describe the contents within each region. 4. Observe transitions between regions.",
        "After these steps, provide your detailed breakdown and then your rationale explaining how and why you segmented the image this way. Think step by step."
      ),

      5: (
        "Let's create a detailed narrative description of this image step by step.",
        "Steps to follow: 1. Establish the main subject or focus. 2. Describe the context and surroundings. 3. Note any actions or movement. 4. Include key details about appearance and condition.",
        "After these steps, provide your detailed narrative and then your rationale describing what guided your narrative construction. Think step by step."
      ),

      6: (
        "Let's analyze the relationship between the image and text step by step.",
        "Steps to follow: 1. Carefully observe all visual elements in the image. 2. Read the text thoroughly to understand its description. 3. Create a list of key entities or concepts mentioned in the text. 4. For each entity, identify its visual counterpart in the image.",
        "After these steps, provide your detailed analysis and then your rationale explaining how the text and visual elements correspond. Think step by step."
      ),

      7: (
        "Let's identify which parts of the text are not related to the image step by step.",
        "Steps to follow: 1. Document all key visual elements in the image. 2. Read through the text to list all statements or claims. 3. For each textual element, check if there is visual evidence. 4. Identify text elements that do not correspond to anything in the image.",
        "After these steps, provide your detailed analysis and then your rationale explaining why those parts lack visual support. Think step by step."
      ),

      8: (
        "Let's analyze the places mentioned in the image and text step by step.",
        "Steps to follow: 1. Identify any locations visible in the image. 2. Look for explicit mentions of places in the text. 3. For each place, determine whether it is present in the image, the text, or both.",
        "After these steps, provide your detailed analysis and then your rationale explaining how you verified each location and its significance. Think step by step."
      )
    }
  }
}
```

Table 56: Language configuration and prompts for ***zero-shot evaluation with rationales*** in English.

Prompt Setting 3: One-shot, No Rationales (in English)

```
EXAMPLES = {
  1: { "question": "Analyze this image and list all objects present. Categorize each object into groups such as furniture, electronic devices, clothing, etc.",
    "example_response": (
      Example Analysis:
      1. Furniture: Airport benches
      2. Electronics: Cameras, Smartphones
      3. Clothing: Jackets, Hats
      4. Misc: Luggage, Backpacks ) },
  2: { "question": "Describe the overall scene in this image. What is the setting, and what activities or events are taking place?",
    "example_response": (
      Example Analysis:
      "An airport terminal with a crowd holding protest signs, some seated and others standing." ) },
  3: { "question": "Identify any interactions or relationships between objects or entities in this image. Explain their connections.",
    "example_response": (
      Example Analysis:
      "Protesters clustered around signs indicate coordinated demonstration." ) },
  4: { "question": "Divide this image into semantic regions. Label each region and describe its contents.",
    "example_response": (
      Example Analysis:
      - Foreground: Protesters and signs
      - Middle: Dense crowd
      - Background: Airport interior ) },
  5: { "question": "Provide a detailed natural language description of what is happening in this image.",
    "example_response": (
      Example Analysis:
      "A protest is taking place in an airport terminal; participants hold signs and gather in groups." ) },
  6: { "question": "Extract and list the specific parts of the text that reference entities, objects, or scenes depicted in the image.",
    "example_response": (
      Example Analysis:
      "The text mentions 'protest signs' which are clearly visible in the crowd." ) },
  7: { "question": "Identify which parts of the text are not represented in the image.",
    "example_response": (
      Example Analysis:
      "Mentions of judicial rulings are not depicted in the image." ) },
  8: { "question": "List the places mentioned in the text or shown in the image.",
    "example_response": (
      Example Analysis:
      "JFK Airport is implied by the terminal setting." ) }
}

LANGUAGE_CONFIGS = {
  "English": {
    "code": "En",

    "system_prompt": "You are an AI assistant that analyzes images and text.",

    "prompt_template_image_only": (
      "Example:",
      "Question: {example_question}",
      "Response: {example_response}",
      "
",
      "Now, analyze the following task:",
      "{task_description}",
      "
",
      "Provide your analysis."
    ),

    "prompt_template_image_text": (
      "Example:",
      "Question: {example_question}",
      "Response: {example_response}",
      "Text associated with the image:",
      "{text_content}",
      "Task:",
      "{task_description}",
      "Provide your analysis based on both the image and the text."
    ),

    "tasks": {
      1: "Analyze this image and list all objects present. Categorize each object into groups such as furniture, electronic devices, and clothing.",
      2: "Describe the overall scene in this image, including the setting and any visible activities.",
      3: "Identify any interactions or relationships between entities in this image and explain their connections.",
      4: "Divide this image into semantic regions. Label each region and describe its contents.",
      5: "Provide a detailed natural language description of what is happening in this image.",
      6: "Extract from the text the parts that reference entities, objects, or scenes depicted in the image and explain the visual evidence.",
      7: "Identify which parts of the text are not represented in the image and explain why.",
      8: "List the places mentioned in the text or shown in the image; for each, indicate if it appears in the text, image, or both and explain its significance."
    },

    "examples": EXAMPLES
  }
}
```

Table 57: Language configuration and prompts for ***one-shot evaluation without rationales*** in English.

Prompt Setting 4: One-shot, With Rationales (in English)

```
EXAMPLES = {
  1: {
    "question": "Analyze this image and list all objects present. Categorize each object into groups such as furniture, electronic devices, clothing, etc.",
    "example_response": (
      "Example Analysis: Objects present include airport seating, electronic devices such as cameras and smartphones, and various articles of clothing."
      "Rationale: The image was scanned systematically; objects were identified based on clear visual boundaries and grouped according to their common functions and appearances." )
  },
  2: {
    "question": "Describe the overall scene in this image. What is the setting, and what activities or events are taking place?",
    "example_response": (
      "Example Analysis: The scene depicts an airport terminal with a large group of people, likely engaged in a protest. Key elements include seating areas and visible signage."
      "Rationale: Visual cues such as the arrangement of people, signage, and environmental details indicate a public gathering in an airport." )
  },
  3: {
    "question": "Identify any interactions or relationships between objects or entities in this image. Explain any spatial, functional, or social connections.",
    "example_response": (
      "Example Analysis: Individuals holding similar protest signs are clustered together, indicating both a social bond and a unified demonstration."
      "Rationale: The proximity and matching visuals (like similar signs) support the inference of coordinated activity." )
  },
  4: {
    "question": "Divide this image into different semantic regions and describe each region's contents.",
    "example_response": (
      "Example Analysis: The image is segmented into foreground (people and protest signs), middle ground (a dense crowd), and background (airport terminal features)."
      "Rationale: Clear visual separations allow the division into regions, helping structure the overall scene analysis." )
  },
  5: {
    "question": "Provide a detailed, natural language description of what is happening in this image.",
    "example_response": (
      "Example Analysis: The image shows a protest at an airport terminal with diverse participants; some are seated while others stand, holding signs that express protest messages."
      "Rationale: Observing the various elements (signs, group formations, terminal features) supports a comprehensive narrative of the event." )
  },
  6: {
    "question": "Extract and list the specific parts of the text that reference entities, objects, or scenes depicted in the image. Explain the visual evidence.",
    "example_response": (
      "Example Analysis: The text mentions 'protest signs' which are clearly visible in the image among the crowd."
      "Rationale: The correspondence between the text and the visible signs confirms the match." )
  },
  7: {
    "question": "Identify which parts of the text are not represented in the image. Explain what is missing.",
    "example_response": (
      "Example Analysis: Details about judicial rulings are mentioned in the text but are not depicted in the image, which shows only protest activity."
      "Rationale: The absence of legal or courtroom elements in the image indicates no visual support for those text parts." )
  },
  8: {
    "question": "What places are mentioned in the text or shown in the image? Indicate whether they appear in the text, the image, or both; and explain their significance.",
    "example_response": (
      "Example Analysis: JFK Airport is inferred from the airport scene in the image and is mentioned in the text."
      "Rationale: The combined visual and textual context highlights the significance of the location as a major travel hub." )
  }
}
LANGUAGE_CONFIGS = {
  "English": {
    "code": "En",

    "system_prompt": "You are an AI assistant that analyzes images and text.",

    "prompt_template_image_only": (
      "Example:",
      "Question: {example_question}",
      "Response: {example_response}",
      "Now, analyze this task:",
      "{task_description}",
      "Provide your analysis."
    ),

    "prompt_template_image_text": (
      "Example:",
      "Question: {example_question}",
      "Response: {example_response}",
      "Text associated with the image:",
      "{text_content}",
      "Task:",
      "{task_description}",
      "Provide your analysis based on both the image and the text."
    ),

    "tasks": {
      1: "Analyze this image and list all objects present. Categorize the objects into groups such as furniture, electronic devices, and clothing.",
      2: "Describe the overall scene in this image, including the setting and visible activities.",
      3: "Identify any interactions or relationships between entities in this image and explain their connections.",
      4: "Divide this image into semantic regions. Label each region and briefly describe its contents.",
      5: "Provide a detailed natural language description of what is happening in this image.",
      6: "Extract from the text the parts that reference entities, objects, or scenes depicted in the image and explain the visual evidence.",
      7: "Identify the parts of the text that are not represented in the image and explain why.",
      8: "List the places mentioned in the text or shown in the image; for each, indicate if it appears in the text, image, or both and explain its significance."
    },

    "examples": EXAMPLES
  }
}
```

Table 58: Language configuration and prompts for English ***one-shot evaluation with rationales*** in English.