Journal of Machine Learning Research 23 (2025) 1-35 Submitted 1/21; Revised 5/22; Published 9/22

VLURes: Benchmarking VLM Visual and Linguistic Understanding in Low-Resource Languages

Jesse Atuhurra* ATUHURRA.JESSE.AG2@NAIST.AC.JP
Division of Information Science
Nara Institute of Science and Technology
Nara, Japan

Iqra Ali IQRA.ALI@QMUL.AC.UK
Division of Computer Science
Queen Mary University of London
London, United Kingdom

Tomoya Iwakura IWAKURA.TOMOYA@FUJITSU.COM
Artificial Intelligence Laboratory
Fujitsu Limited
Kanagawa, Japan

Hidetaka Kamigaito KAMIGAITO.H@NAIST.AC.JP
Division of Information Science
Nara Institute of Science and Technology
Nara, Japan

Tatsuya Hiraoka* TATSUYA.HIRAOKA@MBZUAI.AC.AE
Department of Computer Science
Mohamed bin Zayed University of Artificial Intelligence
Abu Dhabi, United Arab Emirates
**Corresponding Authors*

Editor: Kai-Wei Chang

Abstract

The evaluation of Vision-Language Models (VLMs) is predominantly limited to English-centric benchmarks with short textual contexts, hindering the assessment of their fine-grained reasoning capabilities in diverse linguistic settings. To address this gap, we introduce VLURes, a new multilingual benchmark designed to evaluate the visual and linguistic understanding of VLMs in long-text scenarios. VLURes comprises 4,000 culturally diverse image-text pairs across four languages: English, Japanese, and the low-resource languages Swahili and Urdu. ***VLURes*** introduces eight fine-grained vision-language tasks, including a novel task for identifying `unrelatedness` between modalities. We evaluate ten prominent VLMs on ***VLURes***, analyzing both their direct responses and generated rationales through both automated scoring and human evaluation by native speakers. Our results reveal significant performance disparities across languages, with even the top-performing model, GPT-4o, lagging human performance by 6.7% on average. This gap is substantially larger for open-source models, highlighting critical areas for improvement in multilingual visual reasoning. Models show high sensitivity to language inputs. ***VLURes*** provides a

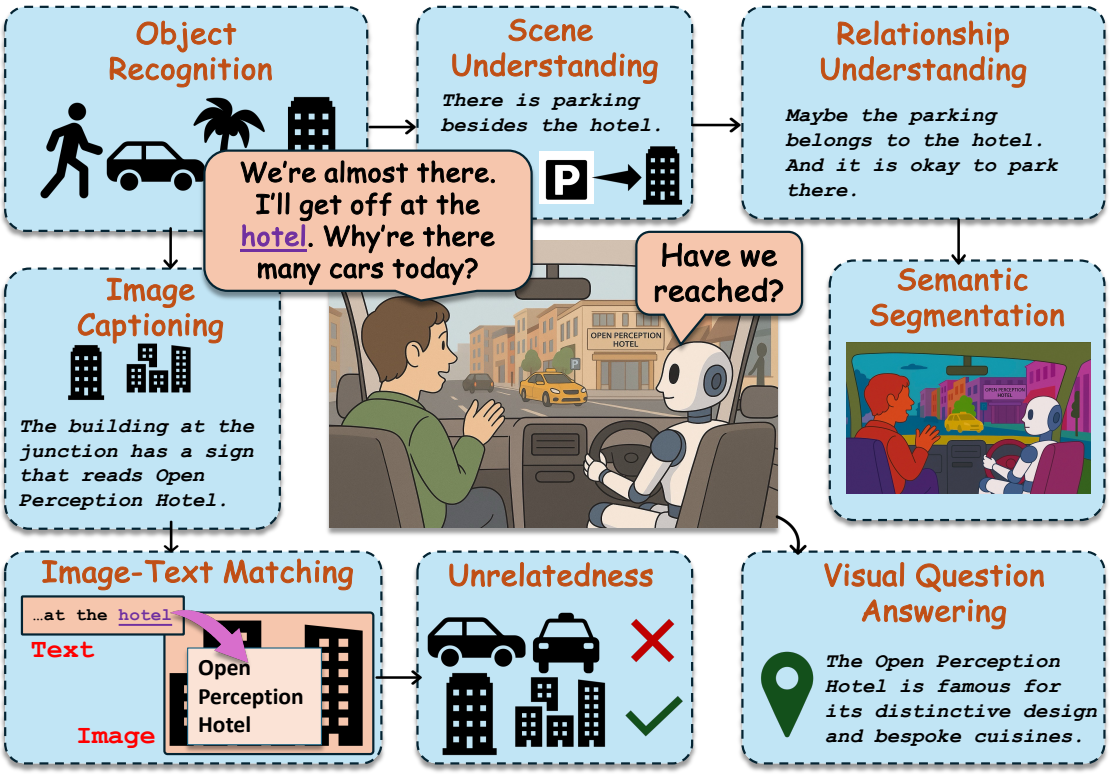

Figure 1: VLURes consists of eight tasks shown above. The tasks enable the intelligent agent to understand its surroundings. We evaluate Vision Language Models' performance on VLURes. It offers image-text pairs in En, Jp, Sw, Ur, and each image is embedded in article-length prose.

crucial new resource for driving the development of more robust, equitable, and globally competent VLMs for real-world intelligent agents, such as robots. Data, code at: https://github.com/jatuhurrra/VLURes.

Keywords: vision-language models, multilingual benchmark, low-resource languages, multimodal reasoning, cross-lingual evaluation

1 Introduction

Vision Language Models (VLMs) have emerged as transformative tools in the field of artificial intelligence, integrating visual and textual understanding to enhance perception in intelligent agents, such as robots (as shown in Figure 1). Agents must interpret complex scenes, recognize objects, and comprehend contextual relationships to navigate and interact effectively in diverse environments. VLMs make up the perception component of modern intelligent agents. Yet, to measure the technological progress made in intelligent agents' perception, the state of the art is to benchmark agents against task-specific data sets. However, existing data sets, such as MSCOCO Zaidan et al. (2007) and Visual Genome Krishna et al. (2017) mainly focus

Table 1: ***VLURes*** includes eight tasks, and prompts specific to each task needed to generate the responses from the VLM. The dots (.....) indicate truncated parts of the VLM response.

Task	Task Question (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.	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.	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...
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.	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... 2. Functional Connections: - The bridge serves... 3. Social Interactions: - There are people walking...
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.	Semantic Regions. 1. Sky: - Location: Top portion of the image. - Description: The sky is mostly clear... 2. Buildings: - Location: Middle section... 3. Bridge/Pathway: ... 4. People: ... 5. Street Lamps: ... 6. Water: ... This breakdown highlights the urban setting and infrastructure...
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.	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... In the background, a skyline of tall, contemporary buildings is visible...
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... 1. Wentworth Point:... 2. Apartment Buildings:...
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... Unrelated Text Elements: - Specific Events: The text describes a specific incident involving a killing... 3. Missing Elements for Illustration: - To illustrate the textual elements, the image would need to show specific individuals, the Toyota Kluger, etc.
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:... 2. Granville:... Significance: - Wentworth Point: Known for its residential developments... - Granville: Known for its diverse community... The image could potentially depict a similar urban environment but does not explicitly show any identifiable location.

on high-resource languages like English. This restricts the assessment of agents' perception in low-resource languages.[1] A further limitation of existing data sets, such as MaRVL Liu et al. (2021) (comprising image-text pairs in Indonesian, Chinese, Swahili, Tamil, and Turkish), is that they offer short *textual descriptions* of the images, hampering efforts to assess the visual and linguistic abilities of agents when thorough textual descriptions exist,[2] for a wide range of vision-language (VL) tasks.

We address the above limitations, to (i) unravel the **V**isual and **L**inguistic **U**nderstanding abilities of VLMs in low **Res**ource languages, (ii) facilitate a thorough analysis of abilities of VLMs' multimodal, especially when rich textual information accompanies the image, (iii) and comprehensively evaluate VLMs at fine-grained VL tasks, shown in Table 1.

Therefore, we first construct a multilingual benchmark, ***VLURes***, containing image-text pairs in *English, Japanese, Swahili, and Urdu.* Second, we embed each image with detailed

1. The languages like Swahili and Urdu, though low-resource, have 200 million and 230 million speakers, respectively, and would benefit from the availability of multimodal evaluation data.
2. Rich textual information supports the agent to perform VL tasks by providing the semantics and context about the environment and the task Zhang et al. (2024a).

context derived from article-length texts. Third, we introduce eight VL tasks, illustrated in Figure 1.

Given an image-text pair, the agent must decide whether to use information from the image, text, or both to accomplish the task. The tasks in *VLURes* are suitable for evaluating agent abilities, such as identifying objects (OR), interpreting scenes (SU), and inferring object interactions (RU), which are essential for agent navigation and manipulation. Our novel *unrelatedness* task challenges VLMs to discard irrelevant text about the image and retain relevant texts, enhancing their robustness in noisy, real-world settings.

Then, given an image-text pair as input, we leverage VLMs to generate responses (shown in Table 1) and **rationales**, which are evaluated automatically by LLM judges, and manually by native speakers. Spanning ten image categories, our benchmark provides rich, *article-length descriptions* drawn from diverse sources, such as websites, blogs, news articles, Wikinews, and Wikipedia, contrasting with the short captions in prior benchmarks (see Table 2).

This work makes the **following key contributions. First,** we introduce multi-task, multimodal VL data sets for Swahili and Urdu, addressing the evaluation gap in both languages Thapliyal et al. (2022). **Second,** we introduce article-length prose in every image-text pair in *VLURes*, providing rich context to the image. **Third,** the inclusion of rationales offers transparency into VLM reasoning, aiding evaluation by both native speakers and the LLM-judges. **Fourth,** our work introduces the novel *unrelatedness* task to evaluate the model's ability to discard irrelevant information. **Finally**, we validate the effectiveness of our benchmark by fine-tuning open-source VLMs, demonstrating their potential for developing multilingual intelligent agents. We publicly release all data sets in this benchmark.

2 Related Work

2.1 Multimodal Benchmarks

Many benchmarks have been developed to evaluate the visual and multimodal understanding abilities of VLMs. The benchmarks include text-and-image pairs, for instance, MMMU Yue et al. (2023), TextVQA Singh et al. (2019), MathVista Lu et al. (2024), MME Fu et al. (2023), MMBench Liu et al. (2023), SEED-Bench Li et al. (2023); video, such as Perception Test Pătrăucean et al. (2023), *inter alia*. Table 2 shows prior benchmarks, and compares our benchmark to prior data sets. As shown in the table, IC, REG, VQA data sets and benchmarks above did not introduce "rationale" information about the *caption, reference expression, or answer* generated/provided by the model or human. Moreover, the benchmarks above did not provide thorough textual context for images in the image-text pairs. These drawbacks motivated us to construct a new VL benchmark. In contrast, we develop a multilingual, multimodal, multitask benchmark containing image-text pairs primarily sourced from web resources in four languages. (We describe more details about this decision later in §5.3.) Additionally, recent works, such as Fu et al. (2023); Liu et al. (2023) and others, which leverage a small yet well-curated sample size of images to provide instructions Xu et al. (2023); Fu et al. (2023); Yue et al. (2023); Yu et al. (2023) to VLMs for performing VL tasks on those images, inspire our work. For example, Fu et al. (2023) assembled a set of 1,077 *images* and corresponding *instruction-answer pairs* to perform eleven perception and cognition tasks in the MME benchmark.

data set	Task	#Tasks	Multilingual	Language	#Languages	Rationales	#Images	#Questions	Article-level Prose
VQAv2 Goyal et al. (2017)	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. (2017)	VQA	1	✘	En	1	✘	108K	1.7M	✘
VizWizQA Gurari et al. (2018)	VQA	1	✘	En	1	✘	*	31.1K	✘
TextVQA Singh et al. (2019)	VQA	1	✘	En	1	✘	28K	45.3K	✘
LAION 5B Schuhmann et al. (2022)	IC	1	✔	En, Zh, ..	many	✘	5.85B	*	✘
LAION-COCO	IC	1	✘	En	1	✘	600M	*	✘
Visual Genome Krishna et al. (2017)	IC	1	✘	En	1	✘	108KM	1.77M	✘
MSCOCO Lin et al. (2014)	IC	1	✘	En	1	✘	328K	*	✘
Flickr30k Plummer et al. (2016)	IC	1	✘	En	1	✘	30K	*	✘
Crossmodal-3600 Thapliyal et al. (2022)	IC	1	✔	En, Jp, 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. (2023)	OCR, ..	20	✔	En, Zh	2	✘	2.9K	2.9K	✘
SEED-Bench Li et al. (2023)	OCR,..	12	✘	En	1	✘	19K	19K	✘
MathVista Lu et al. (2024)	Maths	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,IC,VQA,..	8	✔	En, Jp, Sw, Ur	4	✔	4K	8K	✔

Table 2: A comparison between our data set and existing data sets. In this table, ✔ means "availabe" while ✘ means "unavailable". Yet * means the original paper did not explicitly mention the number of images or questions. Compared to all these data sets, only our data set introduces rationales and article-length prose.

2.2 Rationales in LLMs

Rationales Zaidan et al. (2007) serve as a gateway towards understanding the thought process of LLMs, by explaining how LLMs arrived at the final answer Ling et al. (2017); Hsieh et al. (2023). Moreover, rationales break down problems into manageable steps, improving accuracy in tasks like math and reasoning Yao et al. (2023); Zhou et al. (2023); Zhang et al. (2024b); Hao et al. (2023); Dhuliawala et al. (2024); Wang et al. (2024b); Jiang et al. (2024). In this study, we prompted VLMs to generate rationales and hypothesized that the rationales are valuable for human and automatic evaluation of VLM responses.

2.3 How Different is *VLURes* from Previous Benchmarks?

In summary, the benchmarks mentioned above are mainly limited to English or Chinese. Moreover, previous Swahili or Urdu corpora had ≤ 100 images or only one-sentence captions, limiting comprehensive evaluations. On the contrary, *VLURes* introduces sizable VL resources for Swahili and Urdu with 1,000 image-text pairs per language; long-text grounding whereby *VLURes* embeds the image in article-level prose, letting models examine discourse-level grounding; and explicitly checks a model's ability to discard text snippets that have nothing to do with the picture via the unrelatedness task.

3 Vision and Language Task Definitions

The tasks are related to each other as follows.

Hierarchical Dependency: Image-centric tasks like OR and SS provide foundational data (objects and pixel-level details) that feed into higher-level tasks like SU and RU, which

Table 3: Definitions of the eight VL tasks used in VLURes.

Task	Task Definition
Object Recognition (OR)	The object recognition or object detection task involves confirming the presence of an object Lin et al. (2014). In this study, we investigated the ability of GPT-4V to identify and categorize objects within an image based on their visual features. *For example cat, bottle, car, etc.*
Scene Understanding (SU)	The task involves interpreting the context and the overall scene beyond just individual objects Kafle and Kanan (2017). *For example, A girl is sitting on a bench in a park.*
Relationship Understanding (RU)	This task requires models to identify, characterize, and reason about relationships between different objects within an image Krishna et al. (2017), which includes relationships such as spatial proximity, interaction, ownership, causality, social, and others. *For example, A girl sitting on a desk is feeding the cat, and there is a pet-owner relationship between the girl and the cat.*
Semantic Segmentation (SS)	This task involves dividing an image into parts with a semantic meaning Lin et al. (2014), such as identifying roads, buildings, and people in a street scene.
Image Captioning (IC)	In this task, the VLM is required to generate a natural language description of an image Lin et al. (2014), which captures the scene's content, including objects, actions, relationships, emotions, and atmosphere. *For example, This image contains a girl wearing a pink color skirt and feeding a white color cat.*
Image-Text Matching (ITM)	This task requires the VLM to comprehend which parts of the text correspond to the image Xu et al. (2023). Given an image-text pair, we prompted the VLM to select the exact part of the text that best describes the image.
Unrelatedness (U)	This is a new task that we introduced. Herein, we prompted the VLM to select the exact part of the text that is *not relevant* to the image when given an image-text pair.
Visual Question Answering (VQA)	Under this task, the VLM needs to understand a natural language question about an image and generate appropriate answers Agrawal et al. (2015) by integrating visual understanding, language comprehension, and reasoning abilities.

interpret context and relationships. These, in turn, support multimodal tasks like IC and VQA, which require both visual and linguistic synthesis.

Complementary: Each task probes a distinct aspect of comprehension–OR focuses on "what," RU on "how," SU on "where," and SS on "to what extent"–while ITM, U, and VQA test how well this understanding aligns with text (see Table 3).

Shared Input, Diverse Outputs: A single image-text pair (e.g., a car outside a hotel) can be analyzed across all eight tasks. For instance, OR identifies the car, SU places it in a parking context, IC describes it, and VQA answers questions about it, with ITM ensuring text-image consistency throughout.

Evaluation Synergy: Together, they form a comprehensive framework. Success in one task (e.g., accurate OR) enhances performance in others (e.g., precise IC), while failure in a foundational task (e.g., missing an object in OR) cascades to higher-level tasks, such as VQA.

3.1 Formal Definition: Vision and Language Tasks

To study the visual and linguistic abilities of VLMs, we introduce eight tasks where the input comprises a combination of *images, texts, and prompts*, and the outputs are *responses* generated by VLMs (as illustrated in Table 1). To provide a formal structure for our benchmark, *VLURes*, we define its task space. Let $\mathcal{I}$ denote the space of images and $\mathcal{X}_{\text{txt}}$ denote the space of article-length text descriptions relevant to an image. The benchmark comprises a set of eight distinct vision-language tasks, denoted by $\mathcal{T}$:

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

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

- X_k: The input data space for task τ_k.
- Y_k: The output space, representing the expected response format (e.g., object lists, object relations, segmentation masks, image captions).
- Φ_k: The evaluation function (or metric) used to score the performance on task τ_k, mapping predicted outputs and ground truth labels to a scalar score.

The tasks in *VLURes* are broadly categorized based on their primary input modality requirements during visual reasoning. We formally define these subsets as follows: *(1) Image-Only Reasoning Tasks ($\mathcal{T}_{img}$):* These tasks primarily rely on visual understanding of the image content.

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

For any task $\tau_k \in \mathcal{T}_{\text{img}}$, the input space is predominantly $X_k \subseteq \mathcal{I}$.

(2) Image-Text Reasoning Tasks ($\mathcal{T}_{img+txt}$): These tasks require joint reasoning over both the visual information in the image and the semantic content of the accompanying text.

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

For any task $\tau_k \in \mathcal{T}_{\text{img+txt}}$, the input space involves pairs of images and texts, $X_k \subseteq \mathcal{I} \times \mathcal{X}_{\text{txt}}$.

Task Interdependencies: VLURes is designed to exhibit hierarchical and complementary relationships among tasks. This structure can be conceptualized as a Directed Acyclic Graph (DAG), $G = (\mathcal{T}, E)$, where an edge $(\tau_i, \tau_j) \in E$ signifies that the successful execution of task τ_i potentially provides necessary or beneficial information for task τ_j. For instance, accurate Object Recognition (τ_{OR}) and Semantic Segmentation (τ_{SS}) form foundations for Scene Understanding (τ_{SU}) and Relationship Understanding (τ_{RU}). In turn, these contribute to complex generative tasks like Image Captioning (τ_{IC}) and reasoning tasks such as Visual Question Answering (τ_{VQA}) and Image-Text Matching (τ_{ITM}). Quantifying the strength of these dependencies can be approached using information-theoretic measures. For example, the conditional mutual information between the outputs Y_i and Y_j of two tasks, given their respective inputs X_i and X_j (which are often derived from the same underlying image-text pair), indicates how much information one task's output provides about the other, conditional on the inputs:

$$I(Y_i; Y_j \mid X_i, X_j) = \mathbb{E}_{p(x_i, x_j, y_i, y_j)} \left[\log \frac{p(y_i, y_j | x_i, x_j)}{p(y_i | x_i) p(y_j | x_j)} \right]. \tag{4}$$

Here, the expectation is taken over the joint distribution of inputs and outputs. High mutual information between tasks, such as τ_{OR} and τ_{IC}, formally supports the intuition that identifying objects is crucial for generating accurate descriptions. Computing these values empirically across the benchmark could further highlight the structured nature of visual and linguistic understanding required by *VLURes*.

4 Introducing the Novel *Unrelatedness* Task

Advances in VLM development have enabled researchers to investigate a wide range of tasks embedded in the numerous recently proposed benchmarks. Looking at the multimodal

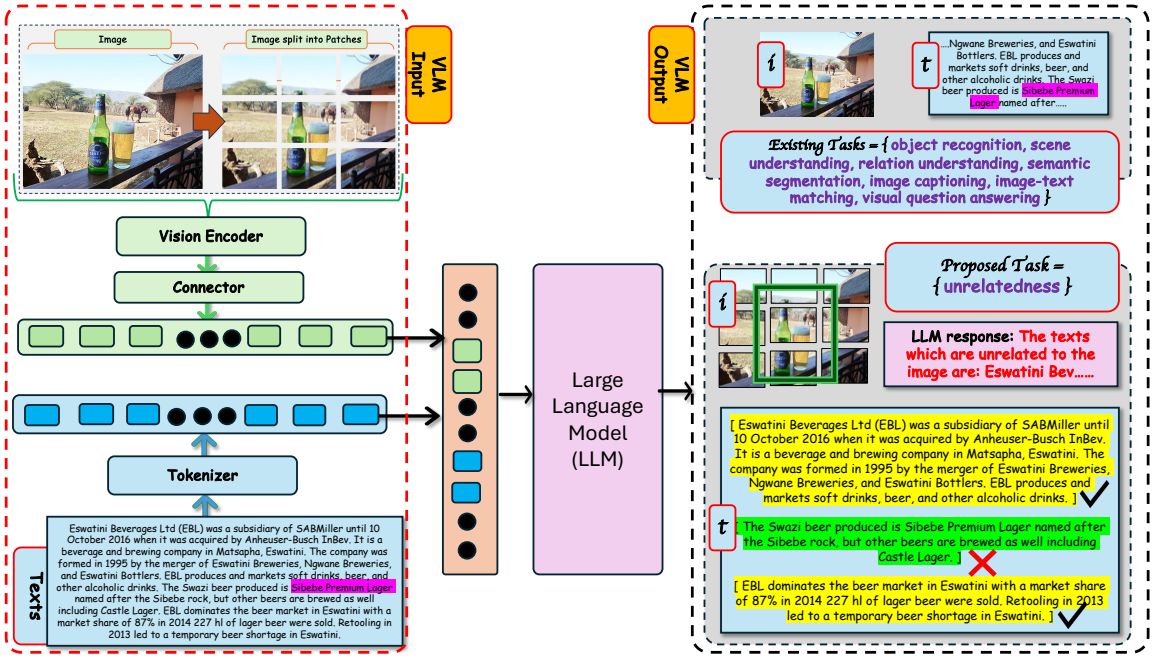

Figure 2: The proposed Unrelatedness task. Left: The VLM inputs consists of two modalities, a pair of image and texts. The image undergoes a series of transformations in the vision encoder and connector, generating visual tokens that are ready for alignment in a shared embedding space. Similarly, a tokenizer tokenizes text, generating textual tokens. Textual and visual tokens are aligned in a shared embedding space and fed as input to the LLM. Right. The LLM uses its multimodal understanding to decide what textual information is relevant to different parts of the image. We see that text painted green (marked with a cross sign) is directly related to the region of the image shown inside a green square box. That is, the text matches the image part shown in green. But in this task, we are interested in text unrelated to the image. Hence, yellow text (marked with a check sign) answers our Unrelatedness task.

benchmarks, that is, benchmarks which evaluate VLM performance when the input to the foundation model is $\boldsymbol{text} + \boldsymbol{Y}$, where $\mathbf{Y} = \{$ **image, video, audio, speech, music, molecule, etc.**$\}$, we observe that there are no benchmarks which explicitly studied the foundation model's abilities to distinguish the information signals from one modality, which are not relevant to the other modality.

From this viewpoint, we deliberate on this effort and define a new task called ***unrelatedness***. The *unrelatedness* task (illustrated in Figure 2) encourages models first to understand the information available from all modalities in the input. Then, via a shared embedding space, the model leverages its multi-modal alignment abilities to align the textual information with the information contained in the other modality, such as image or video. After that, the model must discard the textual information related to the image or video in the shared embedding space and focus on the textual information not relevant to the image or video.

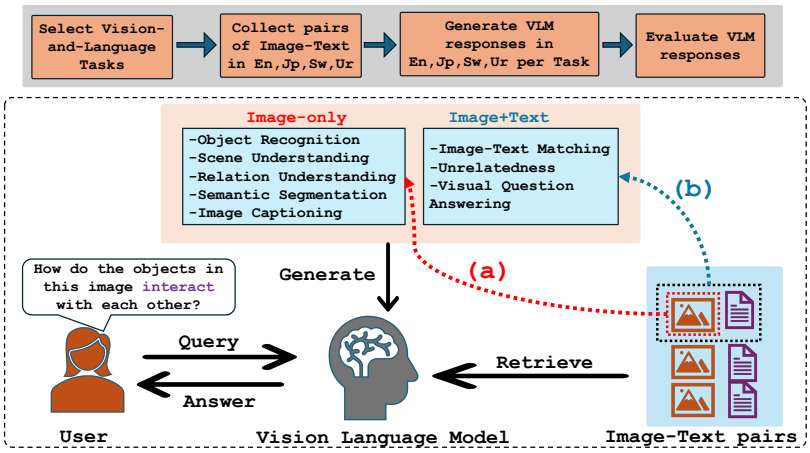

Figure 3: VLURes development process. Top: We selected the VL tasks and constructed data sets for each task and language. We leverage image-text pairs as input to VLMs and prompt VLMs to generate responses for each task. Lastly, we deploy both automatic evaluation using the LLM-as-a-Judge approach and hire native speakers to judge the quality of VLM responses manually. Bottom: We chose the tasks in VLURes to compare two kinds of visual reasoning for each image-text pair in the input: image-only reasoning (shown in (a)) and image+text reasoning (shown in (b)).

Finally, the model generates the answer, highlighting all textual information unrelated to the image.

5 *VLURes* Benchmark

Our work is motivated by the need to train intelligent agents to choose the correct modality required to accomplish a VL task, in the presence of both image and textual information. Therefore, the VLM input is always an image text pair (see Figure 3), but the agent must apply appropriate `image` or `image+text` visual reasoning to accomplish various VL tasks.

5.1 Data Curation Principles

We constructed *VLURes* from web pages because they frequently contain multimedia content, such as images, making them a suitable basis for developing a multimodal benchmark. *VLURes* aims to address the significant gap in current benchmarks that primarily assess visual and linguistic understanding in English. Hence, *VLURes* follows these collection principles: (1) encompass multiple tasks and various topics to reflect real-world agent applications; (2) offer *image-text* pairs in *En, Jp, Sw, Ur*, and the data is native to each language, facilitating multilingual comparison; (3) varying difficulty levels of *image-only* or *image+text* reasoning, to identify limitations in existing models effectively; (4) suitability for

(a) **En Wikinews** article, in `animals`.

(b) **Japanese news** article, in `events`.

(c) **Swahili blog** post, in `events`.

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

Figure 4: We collected all images and texts from each article, then used CLIP scores to select the most relevant image. We embed each chosen image with article-level prose. The chosen images and texts constitute the image-text pairs in our benchmark. Texts shown above are truncated for brevity.

the *multimodal input* processing of modern VLMs, and (5) collect lengthy texts to provide thorough context to the image and to facilitate the evaluation on a wide range of VL tasks.

The taxonomy for this work is introduced as follows: we identify two types of visual reasoning (Figure 3): *image-only reasoning* and *image+text reasoning*. *Image-only reasoning* includes five primary tasks: *Object Recognition (OR), Scene Understanding (SU), Relationship Understanding (RU), Semantic Segmentation (SS), and Image Captioning (IC)*. Moreover, *Image+Text reasoning* includes three primary tasks: *Image-Text Matching (ITM), Unrelatedness (U), and Visual Question Answering (VQA)*. All eight tasks are formally defined in §3.1. In addition, we account for diverse visual contexts, including archival artifacts, photographic images, data visualizations, maps, infographics, and more, from diverse web resources. Lastly, recognizing the lack of multi-task image-text data sets in Sw and Ur, we aimed to evaluate VLMs across various VL tasks.

5.2 Language Selection

We chose languages in *VLURes* under this criteria: (1) languages from distinct families (En is Indo-European, Jp is Japonic, Sw is Niger-Congo, Ur is Indo-European), (2) geographically diverse under the Dryer and Haspelmath (2013) distinction (En from Eurasia, Jp from Eurasia, Sw from Africa, Ur from Eurasia),[3] (3) comprise different writing scripts (En uses Latin, Jp uses Kanji, Sw uses Latin, Ur uses Nastaliq), and (4) includes low-resource languages (Sw, Ur) Nigatu et al. (2024), alongside high resource languages (En, Jp).

5.3 Data Sources and Data Collection

Our data set comprises En, Jp, Sw, and Ur image-text pairs. However, unlike image captioning with short texts Lin et al. (2014); Plummer et al. (2016), we provide "rich context" for each image. An article accompanies every image (see Figure 4).

3. Continent-wise: En, Jp, Sw, Ur originate from Europe, Asia, Africa, and Asia, respectively.

5.3.1 Text+Image Collection

We collected articles from Wikinews and Wikipedia due to their permissive licenses, as well as from news websites, travel blogs, restaurant review forums, and personal blogs. We kept all the texts and images available in each article. Examples are shown in Figure 4.[4][5] To diversify our data set, we gathered image-text pairs for ten categories: *animals, products, buildings, locations, events, food, drinks, hobbies, works of art, and organization*, because articles about these categories are useful to elicit cultural nuances per language, and the articles are readily available on web resources, such as Wikinews, Wikipedia, web forums, and the like. We collected over 1,000 image-text pairs per language and filtered them in the subsequent clean-up process.

Each 'text' in *VLURes* consists of the texts gathered from one URL. During the collection of images from URLs, we restrict the image extension to only `png/jpeg/jpg`, and exclude image URLs that contain the following tokens: `logo, button, icon, plugin, widget`. Moreover, we exclude any articles that contain no valid, downloadable images. We manually checked all images and confirmed no NSFW images in our data sets.

We conducted several data-cleaning steps, for example, removing articles whose main language is not the target language, and URLs containing zero words inside. Moreover, none of the collected documents contained any word on the 'List of Dirty, Naughty, Obscene or Otherwise Bad Words.'[6]

Table 4: Statistics of VLURes.

Metric	En	Jp	Sw	Ur
#Images	1000	1000	1130	996
#Texts	1000	1000	1130	996
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

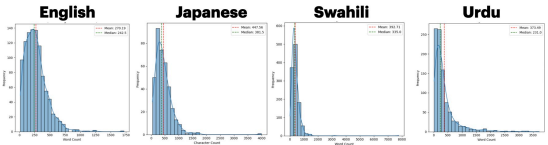

Figure 5: Distribution of text lengths in VLURes indicated by the number of words, for En, Sw, Ur; and the number of characters, for Jp.

5.3.2 Aligning Images with Texts

To assign an image, among many images, to the texts contained in a single article, we treat each article as a bipartite assignment problem Hessel et al. (2019); Kuhn (1955). Then, we deploy CLIP Radford et al. (2021) ViT-L/14 to calculate the pairwise similarity between all text in the article and every image. Images that do not achieve a CLIP cosine similarity of at least 0.15 with any sentence are discarded. Finally, we sorted the CLIP similarity and selected only the image with the highest similarity score to create the *image-text* pair from that article.

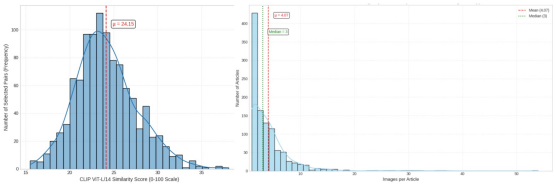

Figure 6: Example distribution of CLIP scores for Sw image-text pairs (left), and number of images in each Sw article (right).

4. This is a link to Wikinews articles in English `https://en.wikinews.org/wiki/Main_Page`.
5. Data dumps from Wikipedia are available at `https://dumps.wikimedia.org/enwiki/`.
6. https://github.com/LDNOOBW/List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words

5.4 Data Statistics

We show detailed statistics about data created in this study in Table 4 and Figures 5, 6. We ensured an even distribution of image-text pairs (1K) across languages. On average, the texts contained in *VLURes* are long enough: 270 words for En, 392 words for Sw, 373 words for Ur, and 447 characters for Jp. These lengthy texts address the limited context per image-text pair in previous data sets. In Figure 5, we visualize the length distributions of the texts across En, Jp, Sw, and Ur. Moreover, Figure 6 shows the distribution of CLIP scores for Sw image-text pairs, and the number of images in each Sw article.

5.5 Benchmark Difficulty and Robustness

Beyond basic statistics, we formalize aspects of *VLURes'* difficulty and its potential to evaluate model robustness, particularly in terms of its multilingual nature and the inclusion of complex reasoning tasks involving long-form text.

Task and Language Difficulty. While inherent task difficulty is complex to define absolutely, we quantify the *empirical difficulty* for a given VLM on a specific task τ_i and language $\ell \in \mathcal{L} = \{\text{En, Jp, Sw, Ur}\}$. Using the primary accuracy metric reported in our experiments (§6), we define a difficulty score

$$\text{Diff}(\tau_i, \ell|\text{VLM}) = 1 - \text{Acc}_{\text{VLM}}(\tau_i, \ell), \tag{5}$$

where $\text{Acc}_{\text{VLM}}(\tau_i, \ell)$ is the accuracy achieved by the specific VLM on task τ_i in language ℓ. A higher Diff score indicates greater difficulty for that model on that specific task-language pair. We hypothesize that factors, such as the (1) low-resource nature of Swahili and Urdu, (2) requirement for grounding long textual contexts (§5.1), (3) and complexity of tasks like RU (τ_{RU}) and U (τ_{U}) contribute to higher average difficulty scores on *VLURes* compared to benchmarks focused solely on high-resource languages or short captions. Analyzing $\text{Diff}(\tau_i, \ell|\text{VLM})$ across VLMs (using results in §7) enables a quantitative comparison of model capabilities on these challenging dimensions. Overall, English is easier for VLMs but Urdu, and Swahili are most difficult (more details in §12).

Cross-Lingual Robustness. A key goal of *VLURes* is to assess how robustly VLMs perform across different languages. We can quantify the robustness of a VLM on a specific task τ_i by measuring the consistency of its performance across the languages $\mathcal{L}$. A simple measure uses the variance of the accuracy scores

$$\text{Robustness}(\tau_i|\text{VLM}) = 1 - \text{Var}_{\ell \in \mathcal{L}} \left(\text{Acc}_{\text{VLM}}(\tau_i, \ell) \right), \tag{6}$$

where $\text{Var}_{\ell \in \mathcal{L}}(\cdot)$ calculates the sample variance across the four languages. A robustness score closer to 1 indicates stable performance across languages for task τ_i. In contrast, a lower score indicates significant performance disparities, potentially highlighting weaknesses in the model's multilingual capabilities or biases towards high-resource languages, such as English. *VLURes* is designed to elicit lower robustness scores compared to English-centric benchmarks, thereby providing a clearer signal on the need for improved language-inclusive perception.

Alternatively, the distribution of performance across languages can be captured using the entropy of the normalized accuracy scores. Let $p(\ell|\tau_i, \text{VLM}) = \frac{\text{Acc}_{\text{VLM}}(\tau_i, \ell)}{\sum_{\ell' \in \mathcal{L}} \text{Acc}_{\text{VLM}}(\tau_i, \ell')}$ be the

normalized performance in language ℓ. The performance entropy for task τ_i is

$$H(\tau_i|\text{VLM}) = -\sum_{\ell \in \mathcal{L}} p(\ell|\tau_i, \text{VLM}) \log p(\ell|\tau_i, \text{VLM}), \tag{7}$$

Higher entropy suggests more evenly distributed performance across languages (higher effective robustness), while lower entropy indicates performance concentrated in only a few languages.

By applying these metrics (Eq. 5, 6, 7) to the results obtained in §7, we provide a rigorous quantitative assessment of model performance profiles, highlighting that *VLURes* is suitable to evaluate the VLMs because it is sufficiently complex and generalizable. Notably, robustness scores could only be computed for models demonstrating broad multilingual support (GPT-4o, GPT-4o-mini, Gemini 2.0 Flash Lite, Gemini 1.5 Flash 8B), as other models lacked comprehensive evaluation data across all four benchmark languages (En, Jp, Sw, Ur) primarily due to missing Swahili results. GPT-4o is the most robust VLM in our evaluation (details in §12).

6 Experiments

6.1 Evaluation Criteria

To evaluate the quality of responses generated by VLMs, researchers may use human-curated rules or deploy template-matching rules Lu et al. (2022). However, we take inspiration from Lu et al. (2024) and follow their evaluation steps, that is, (1) prompt VLMs to generate responses for a task, (2) extract the answer relevant to the task from the model's output, and (3) assign scores on a scale of 1 to 100 to quantify the quality of VLM response. We deploy *accuracy* as the primary metric.

Automatic Evaluation. For automatic/quantitative evaluation, we deploy *Gemini 1.5 Pro* via the Gemini API as the *LLM-judge*.[7] To assign scores for the model's response to any task, the input to the LLM-judge includes the image, text, response from VLM, and a prompt describing the evaluation criteria. Then, the LLM-judge outputs a score, on a scale of one to one hundred, such that one is the lowest and one hundred is the highest.

Human Evaluation. We conducted human evaluations to incorporate a human perspective in evaluating cultural nuance and multilingual understanding of VLMs and to validate rationales. We recruited eight evaluators, two native speakers per language (En, Jp, Sw, Ur), and they possess sufficient visual-linguistic knowledge. Each data sample (consisting of an image-text pair, task instructions, and the VLM's response/rationale) was judged independently by the two native speakers assigned to that language. These native speakers assign a score between 1 and 100, aligning with the accuracy metric reported in our experiments. Full evaluation guidelines are in our online Appendix.

6.2 Model Selection

We chose ten VLMs that have shown competitive results in recent studies. We categorize VLMs as: (1) *proprietary VLMs*, including GPT-4o OpenAI (2024a), GPT-4o-mini OpenAI

7. We made this decision due to: the need to minimize API costs, the model's ability to understand both image and text inputs, large context window, and the high token/sec throughput of Gemini models Duan et al. (2024); Zheng et al. (2023); Artificial Analysis (2025).

Table 5: Performance of VLMs on eight VL tasks under zero-shot and one-shot settings, measured by Accuracy (%). Input: English Texts + Images; Output: En, Jp, Sw, Ur responses. Shaded columns represent {En} in input and {En} in output VLM results.

Model	Object Recognition				Scene Understanding				Relation Understanding				Semantic Segmentation				Image Captioning				Image-Text Matching				Unrelatedness				Visual Question Answering			
	En	Jp	Sw	Ur	En	Jp	Sw	Ur	En	Jp	Sw	Ur	En	Jp	Sw	Ur	En	Jp	Sw	Ur	En	Jp	Sw	Ur	En	Jp	Sw	Ur	En	Jp	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	–	–	–

(2024b), Gemini 2.0 Flash Lite Google (2025), Gemini 1.5 Flash 8B Google (2024), and (2) *open-source VLMs*, including LlaVa-NeXT Mistral 7B Liu et al. (2024), Qwen2VL 7B Wang et al. (2024a), PALO 7B Rasheed et al. (2025), MAYA 8B Alam et al. (2024), LlaVa-NeXT Mistral 13B Liu et al. (2024), and PALO 13B Rasheed et al. (2025).

6.3 Experiment Setup

We evaluate the models under *zero-shot* and *one-shot* settings, before and after *fine-tuning*.[8] Note that we only fine-tuned the open models. We access proprietary models via the relevant API and open-source models via the Hugging Face Transformers Wolf et al. (2020) library. In all three settings, we further compare the results *without rationales* against *with rationales*.

7 Results

We compare the performance of several models on *VLURes*. We set the *zero-shot no rationale* setting as our baseline VLM assessment. In addition, we establish a baseline for human performance in En, Jp, Sw, and Ur by recruiting eight native speakers. All human evaluators are graduate students, and they possess sufficient knowledge of vision and language.

8. We provide the detailed experiment settings, hyperparameters, and data access instructions at this repository: `https://github.com/jatuhurrra/VLURes`.

In Tables 5, 6, we show the performance for all VLMs when the input contains image-text pairs and those texts are in En. Then, VLMs generate responses in En, Jp, Sw, Ur. This setting results in *En-En* as well as *En-Jp, En-Sw*, and *En-Ur* outputs, making it possible to evaluate VLM performance when the language of the VLM response is different from the language in the input texts.[9]

Amongst all settings in Table 5, the *1-shot with rationale* setting results in the best performance, such that GPT-4o achieves slightly above 90% accuracy across all eight tasks. GPT-4o achieves the best accuracy compared to other VLMs but still lags human performance in almost all tasks. For example, in OR, GPT-4o scores a 90.8% accuracy but falls 6.7% short of human performance. The gap highlights the room for improvement on *VLURes*. In contrast, open models achieve weaker performance on *VLURes* than proprietary models. In OR, the best open-source model, Qwen2VL, achieves 62.5%, lagging behind GPT-4o by 28.3 points. The performance deficit may be attributed to factors including limited access to high-quality data during training, a lower level of sophistication in the model architecture, the absence of advanced multimodal fusion techniques needed for seamless interaction between modalities, and the lack of resources or data necessary for extensive fine-tuning.

Additionally, we observe that none of the open models generated intelligible responses for *Sw*. As a result, the entries for *Sw* under LlaVa-NeXT Mistral 7B, Qwen2VL 7B, PALO 7B, MAYA 8B, LlaVa-NeXT Mistral 13B, and PALO 13B in Table 5 are empty. In contrast, all open models generated intelligible responses for *Ur* except the two 7B and 13B versions of LlaVa-NeXT Mistral. From Table 5, we deduce that the level of increasing difficulty on our benchmark is *En*→*Jp*→*Ur*→*Sw*, under the selection of VLMs used in this study.

After fine-tuning, the results in Table 6 show that the Qwen2VL accuracy increased from 62.5% to 71.3%, promising additional benefits in accuracy with additional fine-tuning. Similar to the zero-shot and one-shot settings in Table 5, open models in Table 6 failed to generate intelligible responses for *Sw*. Moreover, we observe a consistent trend in zero-shot, one-shot, and fine-tuning settings where all models achieve higher accuracy when the language in the input text and output text is *En*, that is *En-En*. There is a consistent drop in accuracy for *En-Jp, En-Sw, En-Ur* input-output language pairs under all settings. This finding is consistent with previous findings that VLMs perform best when there is language alignment between the input and the output.

Tables for additional results, given the same language in both input and output, that is **Jp-Jp, Sw-Sw**, and **Ur-Ur**, are detailed in §8, 9, 10.

8 Experiments with *VLURes* Japanese Data

Unlike the experiments in §7 above in which the input texts to VLMs were in *En*, we conducted further experiments with *Jp* texts in the input, alongside each image. We implemented similar settings, that is, zero-shot, one-shot, and fine-tuning settings. There is one distinction between the experiments in this section and the experiments in §7: We prompt the VLMs to generate outputs only in two languages, *En, Jp*, instead of four languages *En, Jp, Sw, Ur* because open VLMs lack strong lingusitc understanding of *Jp* texts like they do for *En*.

9. We conduct this evaluation because VLMs are known to perform better with English prompts due to training data bias than with prompts in other languages Geigle et al. (2024); Chen et al. (2023). We check for any VLM performance drops.

Table 6: Performance of VLMs on eight VL tasks under finetuning settings, measured by Accuracy (%). Input: English Texts + Images. Output: En, Jp, Sw, Ur responses. Shaded columns represent {En} in input and {En} in output VLM results.

Model	Object Recognition				Scene Understanding				Relation Understanding				Semantic Segmentation				Image Captioning				Image-Text Matching				Unrelatedness				Visual Question Answering			
	En	Jp	Sw	Ur	En	Jp	Sw	Ur	En	Jp	Sw	Ur	En	Jp	Sw	Ur	En	Jp	Sw	Ur	En	Jp	Sw	Ur	En	Jp	Sw	Ur	En	Jp	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 7 shows the results for both zero-shot and one-shot settings. Similar to the observations above, the table shows a drop in accuracy among models across all tasks, indicated by Δ_{Jp} values in the shaded columns. The accuracy drop is less severe for *Jp-Jp* texts in the input and output, respectively, but more severe in the *Jp-En* input-output setting. Again, this result is consistent with previous works, which stated that VLMs perform best when language is aligned between the input and the output.

Compared to the results in Table 7, the highest accuracy is 87.9% achieved by GPT-4o under the unrelatedness task, and the second highest accuracy is 87.3% under VQA. The results are also consistent with prior findings in the above section, whereby VLMs achieve higher accuracy when prompted to generate *rationales* in addition to the task's answer. Moreover, providing illustrative task examples in the VLM inputs, under one-shot settings, leads to higher task accuracies than VLMs with no task examples, that is, zero-shot settings.

Proprietary VLMs achieve higher accuracy than open VLMs. GPT-4o is the best-performing proprietary VLM across all tasks, with 87.9% accuracy, while Qwen2VL 7B is the best-performing open VLM, with 65.1% accuracy.

Humans demonstrated a better understanding of the VL tasks in this benchmark than VLMs; hence, human accuracies are higher than the best-performing VLM (GPT-4o) accuracies for all tasks, except VQA. We hypothesize that this could be because GPT-4o better memorizes places and landmarks, along with their names, than humans. However, we do not have any empirical evidence to support this claim.

Moving to the results in Table 8, where open models are fine-tuned with *Jp* image-text pairs, two VLMs (Qwen2VL 7B and PALO 7B) benefited significantly from fine-tuning. These VLMs achieved an increase in accuracy across all tasks compared to the setting in which VLMs were fine-tuned with *En* image-text pairs and prompted to generate *Jp* responses for all tasks. Qwen2VL 7B is the best-performing open VLM after finetuning compared to other VLMs. The observation that eliciting *rationales* in addition to VLM responses in the

Table 7: Performance of VLMs on eight VL tasks under zero-shot and one-shot settings, measured by Accuracy (%). Input: Japanese Texts + Images. Output: En, Jp responses. Shaded columns represent {Jp} in input and {Jp} in output VLM results. En ΔAcc. = {En score from Table 5 (En-Input) – En score from this Table (Jp-Input)}. Jp ΔAcc. = {Jp score from Table 5 (En-Input) – Jp score from this Table (Jp-Input).} Positive Δ in blue, negative Δ in red.

Model	Object Recognition				Scene Understanding				Relation Understanding				Semantic Segmentation				Image Captioning				Image-Text Matching				Unrelatedness				Visual Question Answering			
	En	Δ_{En}	Jp	Δ_{Jp}	En	Δ_{En}	Jp	Δ_{Jp}	En	Δ_{En}	Jp	Δ_{Jp}	En	Δ_{En}	Jp	Δ_{Jp}	En	Δ_{En}	Jp	Δ_{Jp}	En	Δ_{En}	Jp	Δ_{Jp}	En	Δ_{En}	Jp	Δ_{Jp}	En	Δ_{En}	Jp	Δ_{Jp}
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	-3.3	76.1	-3.7	73.4	-3.4	77.5	-3.7	76.0	-3.8	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	–

output leads to better performance than when no *rationales* are generated by the VLM still holds even after fine-tuning.

9 Experiments with *VLURes* Swahili Data

From our initial experiments and analysis, we found that all the open models: LlaVa Mistral 7B, PALO 7B, MAYA 7B, Qwen2VL 7B , LlaVa Mistral 13B, and PALO 13B, understand little about *Sw*. Therefore, we excluded the models from further experiments.

We conducted experiments with *Sw* texts in the input, alongside each image, and we implemented similar settings (zero-shot, one-shot, and fine-tuning). Another distinction between the experiments in this section and the experiments in §7 is: We prompt the VLMs to generate outputs only in two languages, *En, Sw*, instead of four languages *En, Jp, Sw, Ur* because open VLMs lack strong lingusitc understanding of *Sw* texts like they do for *En*.

Table 9 shows the results for both zero-shot and one-shot settings. Consistent with the observations above, the table shows a drop in accuracy among models across all tasks, indicated by Δ_{Sw} values in the shaded columns. Again, the accuracy drop is less severe for *Sw-Sw* texts in the input and output, respectively, but more severe in the *Sw-En* input-output

Table 8: Performance of VLMs on eight VL tasks under finetuning settings, measured by Accuracy (%). Input: Japanese Texts + Images. Output: En, Jp responses. Shaded columns represent {Jp} in input and {Jp} in output VLM results. En ΔAcc. = {En score from Table 6 (En-Input) – En score from this Table (Jp-Input)}. Jp ΔAcc. = {Jp score from Table 6 (En-Input) – Jp score from this Table (Jp-Input).} Positive Δ in blue, negative Δ in red.

Model	Object Recognition				Scene Understanding				Relation Understanding				Semantic Segmentation				Image Captioning				Image-Text Matching				Unrelatedness				Visual Question Answering			
	En	Δ_{En}	Jp	Δ_{Jp}	En	Δ_{En}	Jp	Δ_{Jp}	En	Δ_{En}	Jp	Δ_{Jp}	En	Δ_{En}	Jp	Δ_{Jp}	En	Δ_{En}	Jp	Δ_{Jp}	En	Δ_{En}	Jp	Δ_{Jp}	En	Δ_{En}	Jp	Δ_{Jp}	En	Δ_{En}	Jp	Δ_{Jp}
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

setting. As stated earlier, this result is consistent with previous works, which have shown that VLMs perform best when the language is aligned between the input and the output.

Compared to the results in Table 5, the highest accuracy is 87.1% achieved by GPT-4o under the unrelatedness task, while the second highest accuracy is 86.5% under VQA. Prompting VLMs to generate *rationales* in addition to the task's answer leads VLMs to achieve higher accuracy than when no *rationales* are generated by the VLMs. Similar to the results in the above sections, providing illustrative task examples in the VLM inputs, under one-shot settings, leads to higher task accuracies than VLMs with no task examples, that is, zero-shot settings.

Proprietary VLMs achieve high accuracies, yet open VLMs failed to generate intelligible responses when provided with *Sw* image-text pairs in the input. GPT-4o is the best-performing proprietary VLM across all tasks, with 87.1% accuracy.

Humans further showed a better understanding of the VL tasks in this benchmark than the best-performing VLM (GPT-4o); hence, human accuracies are higher than VLM accuracies for all tasks except VQA. Our earlier hypothesis suggests that GPT-4o may better memorize places and landmarks, along with their names, than humans.

We did not fine-tune open models with *Sw* image-text pairs, as we found that the models lacked strong *Sw* support during our preliminary experiments.

10 Experiments with *VLURes* Urdu Data

From our initial experiments and analysis, we found that the open models: LlaVa Mistral 7B, Qwen2VL 7B, and LlaVa Mistral 13B, understand little about *Ur*. Therefore, we excluded the models from further experiments.

Table 9: Performance of VLMs on eight VL tasks under zero-shot and one-shot settings, measured by Accuracy (%). 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.

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	–

We conducted experiments with *Ur* texts in the input, alongside each image, and we implemented similar settings (zero-shot, one-shot, and fine-tuning). Another distinction between the experiments in this section and the experiments in §7 is: We prompt the VLMs to generate outputs only in two languages, *En, Ur*, instead of four languages *En, Jp, Sw, Ur* because open VLMs lack strong lingusitc understanding of *Ur* texts like they do for *En*.

Table 10 shows the results for both zero-shot and one-shot settings. Consistent with the observations above, the table shows a drop in accuracy among models across all tasks, indicated by Δ_{Ur} values in the shaded columns. Again, the accuracy drop is less severe for *Ur-Ur* texts in the input and output, respectively, but more severe in the *Ur-En* input-output setting. As stated earlier, this result is consistent with previous works, which have shown that VLMs perform best when the language is aligned between the input and the output.

Compared to the results in Table 5, the highest accuracy is 90.6% achieved by GPT-4o under the unrelatedness task, while the second highest accuracy is 90.0% under VQA. Prompting VLMs to generate *rationales* in addition to the task's answer leads VLMs to achieve higher accuracy than when no *rationales* are generated by the VLMs. Similar to the results in the above sections, providing illustrative task examples in the VLM inputs, under one-shot settings, leads to higher task accuracies than VLMs with no task examples, that is, zero-shot settings.

Proprietary VLMs achieve high accuracies, yet open VLMs failed to generate intelligible responses when provided with *Ur* image-text pairs in the input. GPT-4o is the best-performing proprietary VLM across all tasks, with 90.6% accuracy.

Humans further showed a better understanding of the VL tasks in this benchmark than the best-performing VLM (GPT-4o); hence, human accuracies are higher than VLM accuracies for all tasks except VQA. Our earlier hypothesis suggests that GPT-4o may better memorize places and landmarks, along with their names, than humans.

Table 10: Performance of VLMs on eight VL tasks under zero-shot and one-shot settings, measured by Accuracy (%). 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 our experiments due to limited Urdu support in our initial experiments.

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	–

Moving to the results in Table 11, where open models are fine-tuned with *Ur* image-text pairs, PALO 7B benefited significantly from fine-tuning. This VLM achieved increased accuracy across all tasks compared to the setting in which VLMs were fine-tuned with *En* image-text pairs and prompted to generate *Ur* responses for all tasks. However, after fine-tuning, MAYA 8B is the best-performing open VLM across all eight tasks compared to other VLMs. The observation that eliciting *rationales* in addition to VLM responses in the output leads to better performance than when no *rationales* are generated by the VLM still holds even after fine-tuning.

11 Comparison with MaRVL Data set

Liu et al. (2021) introduced the MaRVL data set, mentioned in §1. Moreover, Bugliarello et al. (2022) created IGLUE, a benchmark comprising 20 languages, including *Sw*, and four tasks: natural language inference, question answering, reasoning, and cross-modal retrieval. However, the *Sw* portion of IGLUE uses data from MaRVL. Hence, we compare VLM performance on *Sw* data in our benchmark to VLM performance on the *Sw* data in MaRVL. The results in Table 13 indicate that GPT-4o achieves better accuracy on MaRVL than on *VLURes*, mainly due to the short captions which are found in MaRVL. GPT-4o achieves

Table 11: Performance of VLMs on eight VL tasks under finetuning settings, measured by Accuracy (%). 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.

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																																
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	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	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	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	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																																
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
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																																
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
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																																
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
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 12: Performance of VLMs on eight VL tasks under zero-shot and one-shot settings, measured by Accuracy (%). Input: Swahili Texts + Images; Output: Swahili responses. We do not include LlaVa Mistral 7B, PALO 7B, MAYA 7B, Qwen2VL 7B LlaVa Mistral 13B, PALO 13B, because they showed little understanding of *Sw* in our initial experiments.

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

more significant gains on *U, VQA*, that is 10.7% and 7.8%, respectively, because the model comprehends the short captions in MaRVL much better. Detailed results in §11.

The MaRVL data set 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 data set and the captions accompanying those images.[10]

10. MaRVL data set is available at the URL `https://marvl-challenge.github.io/download`.

Table 13: Performance of VLMs on eight VL tasks, measured by Accuracy (%). Input: Swahili Texts + Images; Output: Sw responses. All images and texts are from the MaRVL data set. We report the best results under one-shot with rationales setting, from GPT-4o. We show the difference in VLM performance between VLURes-Sw and MaRVL-Sw data.

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 14: 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, Jp, 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.

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	$0.576^{(3)}$	$0.598^{(3)}$	$0.592^{(3)}$	$0.597^{(3)}$	$0.630^{(3)}$	$0.630^{(3)}$	$0.585^{(3)}$	$0.603^{(3)}$

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 12.

Under zero-shot settings, the **best accuracy previously reported was 55.5%**, achieved by xUNITER, a variant of the UNITER Chen et al. (2020) 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.

12 Robustness of VLMs across Languages

To unravel the strengths of VLMs across tasks and languages, we calculate the average difficulty scores. A lower average difficulty score indicates that the model is robust to changes in the language of prompt instructions and textual data in its input. On the other hand,

a higher difficulty score means that the model struggles to understand the task when the language of *prompt instructions and textual data* in the model's input changes.

From Table 14, we can see that proprietary models are robust across the eight tasks and the four languages, En, Jp, Sw, and Ur, as indicated by the low difficulty scores. However, GPT-4o is the most robust model among all ten models because it achieved the lowest difficulty scores. For example, under the *scene understanding (SU)* task, the GPT-4o difficulty is 0.128, which implies that GPT-4o can perform this task successfully 89.87% of the time, across four languages. Among proprietary models, the ranking from most robust to least robust is *GPT-4o→Gemini 2.0 Flash Lite→GPT-4o-mini→Gemini 1.5 Flash 8B*.

Unlike the proprietary models, the picture is murkier for the open models. First, none of the open models in our selection supports all four languages in this study. For example, PALO 7B, PALO 13B, and MAYA 8B demonstrated a reasonable understanding of *prompt instructions and textual data* in three languages: En, Jp, and Ur. Yet, LlaVa Mistral 7B, LlaVa Mistral 7B, and Qwen2VL 7B generated intelligible responses in only En and Jp. Open models are less robust to changes in prompting or data language than proprietary models. For example, without any fine-tuning, MAYA 8B can successfully perform the *scene understanding (SU)* task only 41.9% of the time. This pales in comparison with the GPT-4o performance on the same task, which succeeds 89.87% of the time. There is a 47.97% gap between GPT-4o and MAYA 8B under this task.

Because open models are crucial in deploying foundation models to intelligent agents, our results indicate that further improvement in the linguistic and visual abilities of open models is needed to make them robust and pragmatic choices when developing intelligent agents.

13 Sensitivity of VLMs to Language Inputs

VLMs and LLMs are known to exhibit performance drops or increases based on the language in the input. In light of this, we set out to measure the extent to which VLMs in our study are susceptible to performance changes when the language of both input prompt instructions and textual data input is altered. We systematically measure the changes in accuracy using the cross-lingual setting. That is, compared to the accuracy given *En* input data and *En* prompts, what is the change in accuracy when the language is replaced with *Jp, Sw, Ur*. We focus on the decrease or increase of such accuracy. In the tables that follow, accuracy drops are shown in red while increases are shown in blue.

Results are shown in Tables 15, 16, and 17. Table 15 shows the cross-lingual change in English accuracy when switching from English input to Japanese input; Table 16 shows the cross-lingual drop in English accuracy when switching from English input to Swahili input; and Table 17 shows the cross-lingual drop in English accuracy when switching from English input to Urdu input. The overall trend is that models show performance declines in terms of accuracy, for changes from *En* to *Jp,Sw,Ur*.

Table 15: Cross-lingual change in English accuracy when switching from English input to Japanese input (Jp→En). $\Delta_{\text{En}} = \text{Acc}_{\text{En-input}} - \text{Acc}_{\text{Jp-input}}$. Positive Δs in blue, negative Δs in red.

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 16: 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.

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 17: 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.

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

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

14 Project Costs

14.1 Total API Costs

The total computational cost for this study comprises 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 18, 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.

14.2 Human Evaluation Costs

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

15 Discussion

The results presented in §7 demonstrate that VLMs still lag humans on *VLURes* by 6.7%. Hence, VLMs can be improved, and our benchmark can be used to further develop them. Moreover, whereas VLMs struggle with tasks in *VLURes*, humans efficiently perform such tasks, highlighting the gap between humans and VLMs. Additionally, incorporating rationales into VLM reasoning enhances performance on our benchmark, in line with previous findings. In Table 5, the human performance is quite low at 81.9%, compared to the other tasks where human accuracy is near perfect. We do not have explicit reasons, but this probably happened due to fatigue or human error during the evaluation. We observed that VLMs performed best when the input instructions or prompts were given in English. This is a previously reported behavior and is probably caused by data bias at the time of model pretraining Geigle et al. (2024); Chen et al. (2023). Moreover, only a few models were robust across tasks and languages, that is, GPT-4o, GPT-4o-mini, Gemini 2.0 Flash Lite, and Gemini 1.5 Flash 8B. For the other VLMs, a low robustness is due to the limited ability to understand Swahili and Urdu. For example, despite the impressive performance of Qwen2 VL 7B on

English and Japanese data, the model generated unintelligible responses when Swahili or Urdu data were fed as input. Despite the gap between proprietary and open-source models, fine-tuning models such as Qwen2VL has improved performance on all tasks (see Table 6).

Relevance of Rationales. During the human evaluation, evaluators emphasized that the rationales provided more evidence necessary for them to make an informed decision about VLM performance.

***VLURes* Usability.** This benchmark is useful in the following ways. (i) Diagnostic evaluation: Drop your favourite VLM into the eight prompts and instantly see where it falls apart in Swahili or Urdu. (ii) Data-efficient tuning: Because each item carries rich text and multiple tasks, a few hundred examples can already teach cross-modal alignment in low-resource languages. (iii) Noise-robustness research: The Unrelatedness task can serve as a training objective for filtering hallucinated or irrelevant content. (iv) *VLURes* contains long-text grounding, which aligns with real-world agents that must ground paragraphs, not phrases. (v) The unrelatedness task forces models to discard irrelevant text instead of blindly aligning everything-mirrors noisy, open-world scenarios.

16 Limitations

VLURes is limited by the following factors.

Language Scope: *VLURes* covers four languages, but we hope to expand it to more languages in the future. Moreover, though there is coverage of scripts (e.g., Nastaliq for Ur), we did not conduct any script-based (i.e., right-to-left) evaluation. We leave that for future work.

Data Size: 1,000 images per language is still modest for modern deep learning, and models may overfit quickly if used for training rather than evaluation. However, *VLURes* is suited for data-efficient tuning because each item contains rich text and multiple tasks, and only a few hundred examples can teach cross-modal alignment in low-resource languages.

Licence and redistribution of web images: We acquired all images from public sources, and we have done our best to mitigate copyright infringement. Therefore, we encourage everyone to use our benchmark freely, respecting the licence terms. We release the benchmark with a CC BY-NC-SA 4.0 licence.

LLM Judges: We chose a Gemini-based model as the LLM-judge, which could introduce benchmark bias. Previous works, such as Panickssery et al. (2024) have reported the existence of a self-preference bias among LLMs. However, we have done our best to compare LLM judges with human evaluators. As the agreement between VLMs and human evaluators over all tasks and languages shows, the LLM judge selected in this work is reliable for evaluation.

Few-shot Prompts. We recognize that under few-shot prompting, there are many possible values of k, such as, 1,2,3,5,10, etc. However, k=*1* was sufficient for our experiments, and we leave the evaluations with higher k values for future work.

17 Conclusion

This work introduced *VLURes*, a comprehensive multilingual benchmark designed to systematically evaluate the fine-grained reasoning capabilities of Vision-Language Models beyond the confines of English-centric, short-text paradigms. By curating 4,000 culturally diverse,

long-text image pairs in English, Japanese, Swahili, and Urdu, and defining eight challenging vision-language tasks, including a novel unrelatedness task, we have created a robust testbed for assessing the true multimodal and multilingual capabilities of modern VLMs.

Our extensive evaluation of ten state-of-the-art models on VLURes reveals several critical insights. First, we quantify significant performance disparities across languages, demonstrating that even the top-performing proprietary models lag human performance by a notable margin (e.g., a 6.7% gap for GPT-4o). Second, this performance gap widens substantially for open-source models, particularly in the low-resource languages Swahili and Urdu, where some models failed to produce intelligible responses, exposing a critical axis of brittleness. Furthermore, our results confirm the utility of generating rationales for enhancing performance and transparency, while also highlighting that the benefits of fine-tuning are not always distributed equally across different languages.

These findings show a crucial challenge for the development of truly global intelligent agents: the limited ability of current models to handle linguistic diversity and complex, contextual grounding simultaneously. The VLURes benchmark serves not only as an evaluation tool but also as a catalyst for future research. We encourage the community to leverage our benchmark and publicly available datasets to develop novel methods for: (1) more effective cross-lingual transfer learning in the multimodal domain; (2) robust fine-tuning strategies that specifically benefit low-resource languages; and (3) models that can more effectively discern and reason over relevant information in noisy, long-form multimodal contexts.

Overall, by providing a challenging new testbed and a clear empirical picture of current limitations, this work will accelerate progress towards more equitable, robust, and capable vision-language models for real-world applications.

Acknowledgments and Disclosure of Funding

This work was conducted during an internship at Fujitsu AI Lab, and we are grateful for the mentorship provided by the Fujitsu AI team. The High-Performance Computing Center of the Nara Institute of Science and Technology provided the GPUs used in this research. The Natural Language Processing Laboratory provided funding in direct support of this work at Nara Institute of Science and Technology.

References

Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. VQA: Visual question answering. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

Nahid Alam, Karthik Reddy, Sarthak Garg, Saurabh Arjun Sawant, Sarthak Jain, Ankit Tiwari, Sarthak Gupta, Rohan Pandey, Saurabh Suman, Ankit Kumar, Saurabh Kumar, Sarthak Sharma, Saurabh Singh, Sarthak Verma, Saurabh Yadav, Sarthak Joshi, Saurabh Mishra, Sarthak Mehta, and Saurabh Chauhan. Maya: An instruction finetuned multilingual multimodal model. *arXiv preprint arXiv:2412.07112*, 2024.

Artificial Analysis. Artificial Analysis Leaderboards. https://artificialanalysis.ai/leaderboards, 2025. Accessed: 2025-05-12.

Emanuele Bugliarello, Fangyu Liu, Jonas Pfeiffer, Siva Reddy, Desmond Elliott, Edoardo Maria Ponti, and Ivan Vulić. IGLUE: A benchmark for transfer learning across modalities, tasks, and languages. *arXiv preprint arXiv:2201.11732*, abs/2201.11732, 2022.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish V Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme Ruiz, Andreas Peter Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. PaLI: A jointly-scaled multilingual language-image model. In *International Conference on Learning Representations (ICLR)*, 2023.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: UNiversal Image-TExt Representation learning. In *European Conference on Computer Vision (ECCV)*, 2020.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. Chain-of-verification reduces hallucination in large language models. In *Findings of the Association for Computational Linguistics: ACL*, pages 3563–3578, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

Matthew S. Dryer and Martin Haspelmath. The World Atlas of Language Structures Online, 2013. Accessed: 2025-05-09.

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

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. MME: A comprehensive evaluation benchmark for multimodal large language models, 2023.

Gregor Geigle, Radu Timofte, and Goran Glavaš. Babel-ImageNet: Massively multilingual evaluation of vision-and-language representations. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5064–5084, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

Google. Gemini 1.5 Flash-8B is now production ready. https://developers.googleblog.com/en/gemini-15-flash-8b-is-now-generally-available-for-use/, 2024. Accessed: 2025-05-09.

Google. Start building with Gemini 2.0 Flash and Flash-Lite. https://developers.googleblog.com/en/start-building-with-the-gemini-2-0-flash-family/, 2025. Accessed: 2025-05-09.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *International Conference on Computer Vision (ICCV)*, 2017.

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

Shibo Hao, Yi Gu, Haodi Ma, Joshua Jiahua Hong, Zhen Wang, Daisy Zhe Wang, and Zhiting Hu. Reasoning with language model is planning with world model. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

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

Cheng-Yu Hsieh, Chun-Liang Li, Chih kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! Outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL*, pages 8003–8017, Toronto, Canada, July 2023. Association for Computational Linguistics.

Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Zhengping Jiang, Yining Lu, Hanjie Chen, Daniel Khashabi, Benjamin Van Durme, and Anqi Liu. RORA: Robust free-text rationale evaluation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1070–1087, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

Kushal Kafle and Christopher Kanan. An analysis of visual question answering algorithms. In *International Conference on Computer Vision (ICCV)*, pages 1983–1991, 2017.

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

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. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123:32–73, 2017.

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

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

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

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 158–167, Vancouver, Canada, July 2017. Association for Computational Linguistics.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. Visually grounded reasoning across languages and cultures. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 10467–10485, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge, January 2024.

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

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

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

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *IEEE conference on computer vision and pattern recognition (CVPR)*, pages 11–20, 2016.

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

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

Hellina Hailu Nigatu, Atnafu Lambebo Tonja, Benjamin Rosman, Thamar Solorio, and Monojit Choudhury. The Zeno's paradox of 'low-resource' languages. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 17753–17774, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

OpenAI. GPT-4o system card. https://arxiv.org/abs/2410.21276, 2024a. Accessed: 2025-05-09.

OpenAI. GPT-4o Mini: Advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/, 2024b. Accessed: 2025-05-09.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

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

Viorica Pătrăucean, Lucas Smaira, Ankush Gupta, Adrià Recasens Continente, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Joseph Heyward, Mateusz Malinowski, Yi Yang, Carl Doersch, Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alex Frechette, Hanna Klimczak, Raphael Koster, Junlin Zhang, Stephanie Winkler, Yusuf Aytar, Simon Osindero, Dima Damen, Andrew Zisserman, and João Carreira. Perception test: A diagnostic benchmark for multimodal video models. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

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. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

Hanoona Rasheed, Muhammad Maaz, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M. Anwer, Tim Baldwin, Michael Felsberg, and Fahad S. Khan. PALO: A polyglot large multimodal model for 5B people. In *Winter Conference on Applications of Computer Vision (WACV)*, pages 1745–1754, 2025.

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. LAION-5B: An open large-scale dataset for training next generation image-text models, 2022.

Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. Towards VQA models that can read. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8317–8326, 2019.

Ashish V. Thapliyal, Jordi Pont Tuset, Xi Chen, and Radu Soricut. Crossmodal-3600: A massively multilingual multimodal evaluation dataset. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 715–729, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution, 2024a.

Zhaowei Wang, Wei Fan, Qing Zong, Hongming Zhang, Sehyun Choi, Tianqing Fang, Xin Liu, Yangqiu Song, Ginny Wong, and Simon See. AbsInstruct: Eliciting abstraction ability from LLMs through explanation tuning with plausibility estimation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 973–994, Bangkok, Thailand, August 2024b. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP)*, pages 38–45, Online, October 2020. Association for Computational Linguistics.

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

Zhiyang Xu, Ying Shen, and Lifu Huang. MultiInstruct: Improving multi-modal zero-shot learning via instruction tuning. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 11445–11465, Toronto, Canada, July 2023. Association for Computational Linguistics.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik R. Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

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

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, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI, 2023.

Omar Zaidan, Jason Eisner, and Christine Piatko. Using "Annotator Rationales" to improve machine learning for text categorization. In *North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL-HLT)*, pages 260–267, Rochester, New York, April 2007. Association for Computational Linguistics.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-Language models for vision tasks: A survey, 2024a.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models, 2024b.

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. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena, 2023.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V. Le, and Ed H. Chi. Least-to-most prompting enables complex reasoning in large language models. In *International Conference on Learning Representations (ICLR)*, 2023.