
MELLA: Bridging Linguistic Capability and Cultural Groundedness for Low-Resource Language MLLMs

Yufei Gao^{1,2} Jiaying Fei¹ Nuo Chen³ Ruirui Chen⁴
Guohang Yan^{1*} Yunshi Lan^{2*} Botian Shi^{1*}

¹Shanghai Artificial Intelligence Laboratory

²East China Normal University

³ The Chinese University of Hong Kong, Shenzhen

⁴ Institute of High Performance Computing, A*STAR

yfgao.agmail.com

Abstract

Multimodal Large Language Models (MLLMs) have shown remarkable performance in high-resource languages. However, their effectiveness diminishes significantly in the contexts of low-resource languages. Current multilingual enhancement methods are often limited to text modality or rely solely on machine translation. While such approaches help models acquire basic linguistic capabilities and produce “thin descriptions”, they neglect the importance of multimodal informativeness and cultural groundedness — both of which are crucial for serving low-resource language users effectively. To bridge this gap, in this study, we identify two significant objectives for a truly effective MLLM in low-resource language settings, namely 1) linguistic capability and 2) cultural groundedness, placing special emphasis on cultural awareness. To achieve these dual objectives, we propose a dual-source strategy that guides the collection of data tailored to each goal—sourcing native web alt-text for culture and MLLM-generated captions for linguistics. As a concrete implementation, we introduce MELLA, a multimodal, multilingual dataset. Experiment results show that after fine-tuning on MELLA, there is a general performance improvement for the eight languages on various MLLM backbones, with models producing “thick descriptions”. We verify that the performance gains are from both cultural knowledge enhancement and linguistic capability enhancement. Our dataset can be found at <https://opendatalab.com/applyMultilingualCorpus>.

1 Introduction

Multimodal Large Language Models (MLLMs), such as Qwen2.5-VL [Bai et al., 2025] and InternVL2.5 [Chen et al., 2024a] have achieved great success, but their capabilities are predominantly confined to high-resource languages like English, as illustrated in Figure 1. This imbalance creates a significant “digital divide”, leaving speakers of low-resource languages behind.

Previous attempts, such as SDRRL [Zhang et al., 2024], LexC-Gen [Yong et al., 2024] and Amharic LLaVA [Andersland, 2024], to enhance multilingual capability primarily focus on **text modality** or rely on **machine translation(MT)**, see Table 1. However, these methods overlook a critical distinction. As Barthes [Barthes, 1985] suggests, images convey rich cultural narratives through “connotation”, a symbolic depth that translated text often fails to capture. Consequently, an MLLM trained on translation-based data is confined to performing what Geertz [Geertz, 1973] terms a “thin

*Corresponding author

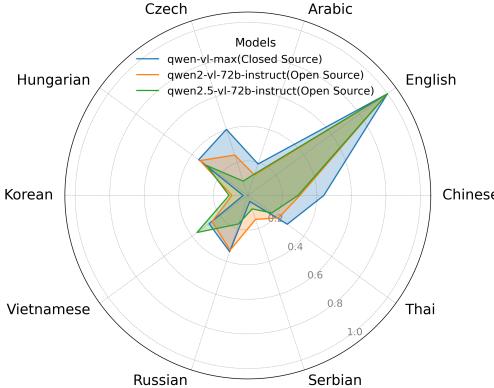


Figure 1: Image caption task performance on COCO dataset [Lin et al., 2015] across multiple languages. Compared to GPT-4o [OpenAI et al., 2024], most of the outstanding MLLMs get the highest BLEU [Papineni et al., 2002] score in English.



Figure 2: Standard MLLMs (e.g., InternVL2-8B, Qwen2-VL-7B) trained on generic datasets often fail to generate meaningful output due to limited visual-linguistic alignment. An MLLM with enhanced linguistic capability may produce detailed descriptions. However, only an MLLM enriched with cultural knowledge can accurately recognize the depicted celebrity. All conversations are expected to be in Arabic; “EN” provides translation for clarity.

description”: it recognizes surface-level content but fails to grasp the deeper, culturally embedded “webs of significance”. Taking Figure 2 as an example, without cultural grounding, the MLLM can describe visual content literally (e.g., “a man in traditional dress”) but fails to identify culturally significant entities (e.g., recognizing the man as a specific Arabic prince). For users speaking low-resource languages, this results in outputs that are factually correct but culturally irrelevant, which can harm user trust, usability, and inclusiveness. For a low-resource language MLLM to be truly effective, it cannot just speak a language; it must understand the culture it represents. This leaves a critical research gap: the lack of a methodology to jointly enhance both **linguistic capability** and **cultural groundedness** in a multimodal setting.

	Multimodal	Cultural Awareness	Linguistic Data Source	Cultural Data Source
SDRRL	✗	✗	LLM+Gen+MT	N/A
LexC-Gen	✗	✗	Lexicon Translation	N/A
Amharic LLaVA	✓	✗	MT Captions	MT Captions
Dual Source(Ours)	✓	✓	MLLM+Gen+MT	Native Web Alt-text

Table 1: Comparison of multilingual enhancement approaches. Unlike methods that ignore image informativeness and rely on machine translation, our method promotes cultural awareness by sourcing data from **Native Web Alt-text**—authentic web image descriptions authored by individuals within specific cultural contexts.

To address this issue, we decompose image meaning into two components: a literal, objective denotation and a symbolic, culturally-coded connotation. Prior approaches to multilingual enhancement have primarily focused on the former. To bridge this gap, we explicitly introduce a dual objective for low-resource language MLLMs: (1) **Linguistic Capability**, which ensures fluency and nuanced expression, and (2) **Cultural Groundedness**, which enables understanding of culturally specific knowledge. Recognizing that the gap largely stems from an imbalance of culturally-relevant multimodal data across languages [Romero et al., 2024], we further propose a high-level, dual-source framework that integrates both a data collection strategy and a training objective to achieve this dual goal.

To instantiate the dual-source framework, we construct **MELLA**, the first initiative to address the dual challenges jointly. As Table 2 shows, MELLA is unique in its motivation and data curation method. The construction and usage of MELLA follow the proposed dual-source data strategy. First, to instill cultural groundedness, we curate native web corpora, extracting images along with their original HTML alt-text to form a knowledge-rich dataset D_{know} . This alt-text provides invaluable, human-authored context about culturally specific people, places, and objects. Second, to foster linguistic

capability, we leverage a state-of-the-art MLLM to generate detailed English image descriptions, which are then translated into the target languages to create a linguistics-focused dataset D_{ling} . Experiments on two model backbones show clear improvements across both goals using our dataset, indicating the effectiveness of the dual-source framework.

Our main contributions are:

- We propose a dual objective for low-resource language MLLMs, placing special emphasis on cultural awareness. To support this, we also introduce a dual-source strategy that offers high-level guidance toward fulfilling the dual objective. (Section 2)
- As an instance of dual-source strategy, we present MELLA, a novel multimodal multilingual dataset with 6.8 million image-text pairs across eight low-resource languages. (Section 3.1)
- Extensive experiments across various model backbones demonstrate the effectiveness of our strategy, achieving significant improvements over existing methods. (Section 4)

Dataset	Primary Goal	Low-Resource Focus	Cultural Focus	Data Curation Method
WIT	Large-scale Pre-training	Incidental (100+ languages)	Incidental	Sourced from Wikipedia image-caption pairs across languages.
LAION-5B	Large-scale Pre-training & Finetuning	Incidental (English-centric)	Incidental	Filtered Common Crawl based on CLIP score; alt-texts are unverified.
MTV-QA	Multilingual Text-centric VQA Benchmarking	Targeted	Incidental	Filtered Common Crawl based on OCR API; manually collect.
EXA-MS	Multilingual Exam Benchmarking	Targeted	Specific	Sourced from multilingual high school exam papers.
CVQA	Cultural Benchmarking	Targeted	Specific	Local annotators manually collect images and create questions based on a guideline.
MELLA (Ours)	Fine-tuning for Cultural & Linguistic Skills	Targeted	Specific	Automated collection and annotation; Dual Source: 1) Native web alt-text for cultural Groundedness; 2) MLLM-generated descriptions for linguistic capability.

Table 2: Comparison of multimodal datasets: WIT [Srinivasan et al., 2021], LAION-5B [Schuhmann et al., 2022a], MTV-QA [Tang et al., 2024a], EXA-MS [Das et al., 2024], CVQA [Romero et al., 2024].

2 Bridging Linguistic Capability and Cultural Groundedness

We begin by elaborating on the motivation for bridging linguistic capability and cultural groundedness for low-resource language MLLMs. Building on this motivation, we define a dual objective and propose a framework that bridges the two.

2.1 Motivation

The meaning of an image is not monolithic. Drawing from semiotics [Barthes, 1985, Geertz, 1973], we posit that the total meaning μ of an image I can be decomposed into two fundamental components: a literal, objective denotation (μ_{den}) and a symbolic, culturally-coded connotation (μ_{con}). The denotation represents a “thin description”—what is explicitly visible—while the connotation carries the “thick description”—the culturally embedded “webs of significance” that give the image deeper cultural meaning:

$$\mu(I) = (\mu_{den}(I), \mu_{con}(I)). \quad (1)$$

Prevailing methods in Table 1 for low-resource language MLLM enhancement, which often rely on translating existing English-centric datasets, primarily address denotation (μ_{den}). Consequently, they train models that can describe a scene but fail to grasp its cultural context, such as identifying a local celebrity or understanding the significance of a traditional garment. Without cultural groundedness, MLLMs produce shallow, decontextualized outputs that fail to meet the needs of diverse global users.

2.2 Dual Objective

To bridge the “ $\mu_{den} - \mu_{con}$ ” performance gap, we formalize two core capabilities an MLLM must master to be truly effective in a low-resource setting — linguistic capability and cultural groundedness. We propose a dual-objective to jointly model these capabilities:

2.2.1 Objective 1: Linguistic Capability

We define linguistic capability f_{ling} as the model’s ability to generate a fluent and accurate text T_{den} in a target language L , effectively capturing the denotative meaning μ_{den} of an image I :

$$f_{ling} : (I, L) \rightarrow T_{den}, \quad (2)$$

where T_{den} is a textual representation of $\mu_{den}(I)$. This is the ability to produce a “thin description”. It requires mastery of vocabulary and grammar in language L .

2.2.2 Objective 2: Cultural Groundedness

We define cultural groundedness f_{cult} as the model’s ability that can infer and articulate the connotative, culturally-specific knowledge μ_{con} embedded in an image I :

$$f_{know} : (I, L) \rightarrow T_{con}, \quad (3)$$

where T_{con} is a textual representation of $\mu_{con}(I)$. This is the ability to produce a “thick description”. This function is difficult to learn through translation-based methods alone; we argue that it should instead be learned from authentic, culturally grounded data.

2.3 Dual-source Framework

To achieve the dual objective, we propose a framework that contains a dual-source data strategy and a unified training objective.

2.3.1 Dual-source Data Strategy

Previous methods struggle to address μ_{con} , primarily due to the profound scarcity of aligned, culturally relevant multimodal data for low-resource languages. To overcome this bottleneck, we propose a dual-source data strategy — constructing a dataset D from two distinct sources, each targeting one of the two objectives. One source is a linguistics-focused dataset, denoted as $D_{ling}^L = \{(I_i^L, T_{den,i}^L), i = 1, \dots, M\}$, where L represents the target language and M denotes the total number of image-text pairs in that language. Each pair contains an image I_i^L and a corresponding denotative description $T_{den,i}^L$ — a fluent and accurate caption originally generated in English and then translated into the target language L . This dataset provides the primary training signal for the linguistic capability function f_{ling} .

The other source is the cultural knowledge-focused dataset, denoted as $D_{know}^L = \{(I_j^L, T_{con,j}^L), j = 1, \dots, N\}$, where N is the number of culturally grounded samples in language L . D_{know} consists of image-text pairs sourced from authentic, in-culture contexts (e.g., native web corpora). This dataset provides the necessary signal for f_{cult} . Unlike D_{ling}^L , this dataset reflects culturally specific knowledge and expressions grounded in real-world usage, providing the essential training signal for f_{cult} . The final training corpus is the union of these two: $D = D_{ling} \cup D_{know}$.

2.3.2 Unified Training Objective

The ultimate goal is to train a unified model \mathcal{M} that approximates both functions. The model’s final output T_{output} for an image I should ideally integrate both denotative fluency and connotative awareness:

$$\mathcal{M}(I, L) \rightarrow T_{output} \approx T_{den} \oplus T_{con}, \quad (4)$$

where \oplus denotes the integration of both fluent description and cultural keywords, L denotes the expected language of T_{output} . The dual-source training on D is a direct operationalization of this principle, forcing the model to jointly optimize for both linguistic expression and cultural interpretation within a single framework.

3 MELLA : Instantiating the Framework

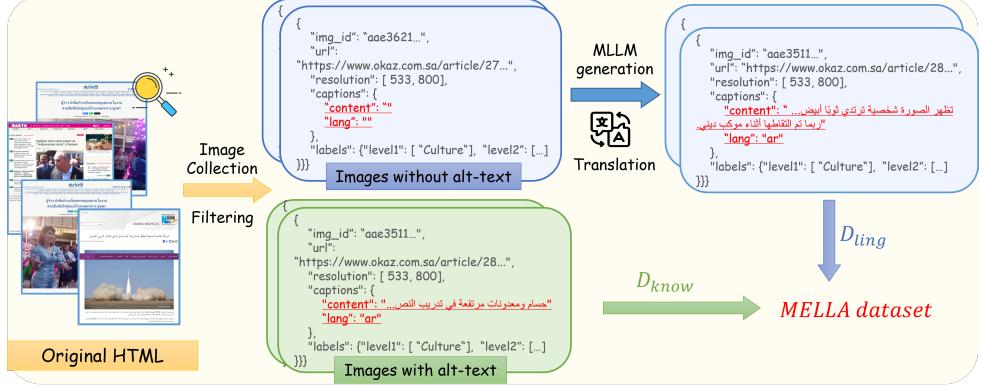


Figure 3: Data Collection Pipeline for MELLA . We first collect images with native alt-text from regional websites to form the cultural knowledge dataset (D_{know}), For images without alt-text, we use a powerful MLLM to generate descriptive captions, which are then translated into target low-resource languages to form the linguistic capability dataset (D_{ling}). The combination of these two sources creates our final MELLA dataset.

MELLA (Multilingual Enhancement for Low-resource LAnguage MLLM) is our dual-source, multimodal multilingual dataset created as a direct instantiation of the dual-source framework described in Section 2.3. We describe how the dataset was constructed, summarize key statistics, and the training procedure.

3.1 Dataset Construction

The construction process consists of 1) Image Collection and Filtering, 2) Text Generation for Alignment, 3) Translation for Low-resource Languages, as illustrated in Figure 3.

3.1.1 Image Collection

We focus on eight languages identified in prior work [Tang et al., 2024b, Srinivasan et al., 2021, Das et al., 2024]: Arabic (AR), Czech (CS), Hungarian (HU), Korean (KO), Russian (RU), Serbian (SR), Thai (TH), and Vietnamese (VI). These languages are selected based on their limited coverage in existing multimodal multilingual datasets and the increasing demand for inclusive language support in AI systems.

Inspired by the methodology of Schuhmann et al. [2022b], we curated a diverse set of HTML web pages in these languages by crawling 24 high-traffic websites from regions where the target languages are primarily spoken. These sources span a broad range of domains—including news media, government services, commercial platforms, online forums, and encyclopedias—and cover diverse topics such as health, science, technology, and education. The full list of crawled websites is provided in the appendix B.1.

From the collected HTML files, we extract images that are culturally and linguistically relevant visual content. These images are automatically categorized using InternVL-1.5-25.5B [Chen et al., 2024b] into 4 major categories and 20 fine-grained subcategories, as illustrated in Figure 4. The extracted images are often embedded within the context of language-specific information. We then apply a rigorous series of filtering steps, detailed in the appendix B.2, to ensure data quality. This process yields a final set of approximately 6.82 million (M) high-quality images.

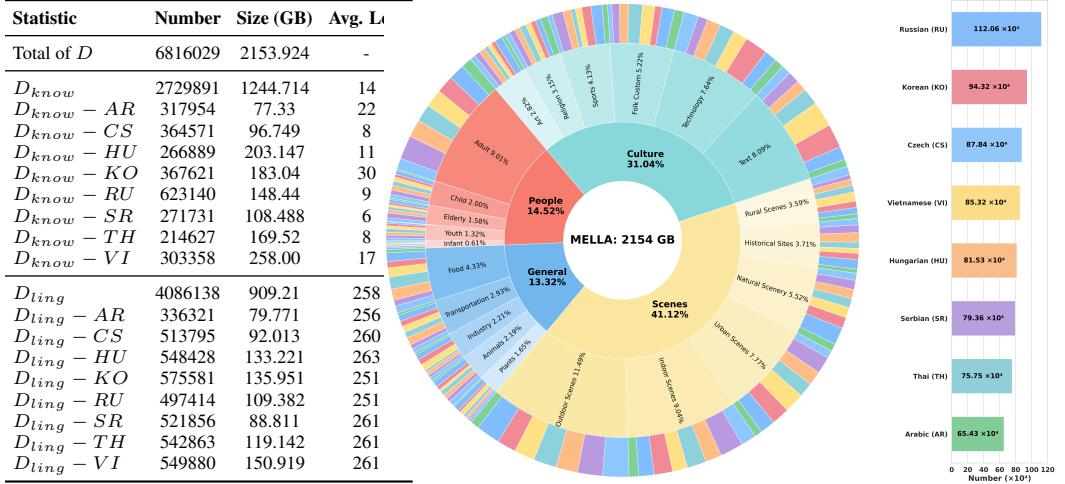


Figure 4: Statistical overview of the MELLA dataset. Left: Main statistics including total sample numbers, sizes, and average text lengths across different languages. Middle: Circular diagram of the category distribution visualization. Right: Quantitative distribution showing the eight languages in the dataset with consistent color coding across the diagram. As shown, the MELLA dataset exhibits both broad coverage and balanced representation across topics and languages.

3.1.2 Text Generation for Alignment

Before obtaining the full datasets D_{know} and D_{ling} , we collect T_{con} and T_{den} following the dual-source data strategy.

Alt-text collection for cultural groundedness. We use alt-text as T_{con} . Alt-text is a critical metadata from HTML files, providing semantic descriptions of web images, primarily aiding accessibility for visually impaired users [Sharma et al., 2018, Chintalapati et al., 2022]. The alt-text is authored by the web page creators and usually enriched with reliable knowledge, such as the name of a celebrity, the local dialect of an object, which is presented in low-resource languages. More importantly, the standard MLLMs or LLMs are short of such knowledge, so the alt-text can be deemed as the external knowledge curated from the raw corpus. To leverage this auxiliary signal, for each target low-resource language, we extract alt-texts as T_{con} , pairing them with corresponding images I to construct a set of aligned image-text pairs:

$$D_{know} = \{(I_i, T_{con,i}) | i = 1, \dots, N\}.$$

It is worth noting that the language of alt-text is decided by the language of the web page where the image is crawled. We also conduct language inspection using HTML metadata and a language detection tool [Joulin et al., 2016] to ensure alt-text in D_{know} is written in the target languages.

Text generation for linguistic capability. In the case of images lacking alt-text annotations, we generate textual descriptions for these images using an advanced MLLM and then translate them into different low-resource languages. This process yields T_{den} . T_{den} can effectively supply linguistic information with rich image descriptions in low-resource languages. Specifically, to facilitate an MLLM to generate more accurate, aligned text, we carefully design domain-specific prompting (see appendix B.3). To further ensure the utility and quality of the dataset, we conduct a manual review to verify the high relevance between each image and its text description.

While the generated texts are high-quality and standardized, they are presented in English. Hence, we employ the advanced machine translation systems to translate the texts in D_{ling} into the eight low-resource languages. For each target language, we translate the text via either *DeepL Translate* [DeepL, 2023] or *Google Translate* [Google, 2023] based on their supported languages. To ensure translation quality, outputs are reviewed by human experts with formal training or backgrounds in the target low-resource languages. Following SDRRL [Zhang et al., 2024], we use WMT22-cometkiwi-da [Rei et al., 2022] for evaluation, achieving an average score of 0.75. For each target low-resource language, this process yields a set of aligned image-text pairs

$$D_{ling} = \{(I_i, T_{den,i}) | i = 1, \dots, M\},$$

where I_i denotes an image, and $T_{den,i}$ denotes the generated text paired with the image. The final dataset is:

$$D^L = D_{know}^L \cup D_{ling}^L, L \in \{AR, CS, HU, KO, RU, SR, TH, VI\}. \quad (5)$$

3.2 Data Statistics

Figure 4 presents comprehensive statistics of the MELLA dataset. MELLA has a total number of 6.8M image-text pairs, evenly covering 8 low-resource languages, containing 4 major and 22 fine-grained semantic categories, highlighting the diversity and richness of the data.

3.3 Training Objectives

For the “unified training objective” in our proposed dual-source framework, we follow recent advances in low-resource language enhancement [Zhang et al., 2024], performing supervised fine-tuning (SFT) on an existing MLLM with the low-resource benchmark using the collected dataset D in a parameter-efficient manner.

We formally define the SFT task. To mitigate overfitting, we first manually crafted 20 prompts for each language L , constructing a prompt pool $P = \{x_i^L | i = 1, \dots, 20\}$ (refer to the appendix C.2 for details). Given an input image I and a corresponding prompt x randomly selected from P , the task is defined as generating a target text sequence T in a specific low-resource language L . For each target language, we fine-tune a model, parameterized by θ , using a standard cross-entropy objective:

$$\mathcal{L}_{CE} = -\mathbb{E}_{((I,x),T) \sim D^L} \left[\sum_{t=1}^{|T|} \log P_\theta(T_t | T_{<t}, I, x) \right], \quad (6)$$

where $T = \{T_1, \dots, T_n\}$ is the tokenized target text in language L , and P_θ is the probability of predicting the next token given the previous context and the multimodal input.

4 Experiments

4.1 Experimental setup

4.1.1 Dataset

For each language, we use a random subset whose size is about 80-140K from the collected datasets; a detailed training dataset statistic is listed in the appendix C.3. To construct our test sets, we randomly sample 1,600 instances from a held-out dataset that are not involved in any stage of training. For each target low-resource language L , we select 100 samples from D_{know}^L and 100 samples from D_{ling}^L , resulting in 200 test samples in total.

4.1.2 Evaluation Metrics and Details

Since D_{know}^L and D_{ling}^L are designed to investigate different understanding capabilities of an MLLM, we test them with different evaluation metrics. Regarding D_{know}^L , following DeFactoNLP[Reddy et al., 2018], we employ keyword accuracy as the evaluation metric. We identify the keywords using TF-IDF, and accuracy is computed by comparing the presence of keywords between the prediction output and the ground truth annotations. Regarding D_{ling}^L , we require an MLLM to answer a question in fluent low-resource languages. Following Zhang et al. [2024], we use the metrics for text generation to compare the prediction with ground truth annotations: BLEU [Papineni et al., 2002], ROUGE-L [Lin, 2004] and METEOR [Denkowski and Lavie, 2014]. We use a uniform prompt that is leveraged for various low-resource languages for a fair evaluation. For data from D_{know}^L , our prompt is “*Describe the picture, point out the people and objects in it!*” for each language, and this prompt is translated to the corresponding language using *Google Translate*. For data from D_{ling}^L , the prompt is “*Describe this image.*” which is translated to the corresponding language. For instance, when evaluating on D_{know}^{HU} , the prompt is “*Ismertesse a képet, mutasson rá a rajta lévő személyekre és tárgyakra!*”

4.1.3 Comparable Methods

We choose both InternVL2-8B and QwenVL2-7B as our MLLMs backbones due to their wide usage in multimodal tasks [Wang et al., 2024, Zhang et al., 2025]. We compare with the following two baselines:

- - : This is the original MLLMs for evaluation. We do not do any fine-tuning and just prompt the MLLMs with the questions and evaluate their performance.
- **SDRRL** [Zhang et al., 2024]: This is an earlier method proposed to enhance large language models’ capabilities in low-resource languages. It constructs a cross-lingual transfer dataset and incorporates external parallel corpus. It also leverages “translate then SFT” paradigm with resource-rich languages. But SDRRL mainly focuses on the linguistic adaptation and does not involve knowledge of low-resource languages during training.

Implementation details

Our code is implemented using DeepSpeed [Rasley et al., 2020] on two NVIDIA A100-SXM4-80GB GPUs. Main training hyperparameters and experiment details can be found in the appendix C.1.

Backbones		AR	SR	RU	CS	KO	TH	VI	HU
<i>Keyword Accuracy</i>									
InternVL2-8B	-	2.46	0.56	1.24	1.10	0.50	3.72	0.78	4.39
	SDRRL	2.39	0.33	1.22	1.37	1.02	3.38	1.00	2.00
	MELLA	6.26	3.07	8.37	15.56	5.06	4.50	2.50	5.57
Qwen2-VL-7B-Instruct	-	1.56	0.80	3.12	2.89	2.00	4.55	0.32	2.16
	SDRRL	0.01	0.66	0.45	1.78	0.01	2.86	0.15	1.57
	MELLA	2.23	1.13	3.26	4.90	4.13	4.97	0.65	2.92
<i>Meteor</i>									
InternVL2-8B	-	26.07	2.70	7.71	3.37	14.54	19.95	18.19	0.11
	SDRRL	22.46	5.23	5.83	6.62	13.83	11.77	11.1	5.68
	MELLA	29.78	13.54	4.91	12.17	22.81	22.5	16.37	13.11
Qwen2-VL-7B-Instruct	-	15.49	2.33	6.54	6.03	12.93	17.14	16.77	6.37
	SDRRL	2.35	0.25	1.28	5.32	0.76	18.48	1.92	7.01
	MELLA	36.89	13.88	5.36	12.88	23.74	34.63	28.66	12.72
<i>BLEU</i>									
InternVL2-8B	-	1.79	1.05	5.56	1.31	2.56	0.15	6.91	0.05
	SDRRL	12.18	6.11	7.01	7.59	6.91	0.45	11.07	6.09
	MELLA	13.96	13.22	4.40	14.33	11.02	0.56	15.53	13.45
Qwen2-VL-7B-Instruct	-	2.45	0.60	3.24	2.37	1.48	0.32	8.17	3.40
	SDRRL	1.43	0.21	6.16	6.29	0.49	0.67	1.66	7.44
	MELLA	19.95	16.33	6.26	14.80	11.48	1.00	30.18	13.39
<i>Rouge-L</i>									
InternVL2-8B	-	5.23	6.41	12.73	6.25	6.25	0.50	12.39	0.22
	SDRRL	14.37	7.07	8.60	10.18	9.17	1.55	9.98	7.91
	MELLA	17.26	18.77	6.32	17.74	14.97	2.25	14.57	18.41
Qwen2-VL-7B-Instruct	-	11.30	5.50	12.86	1.11	7.85	1.31	16.84	11.30
	SDRRL	1.59	0.38	10.19	8.38	0.87	2.22	1.82	10.29
	MELLA	24.13	20.08	8.47	19.02	16.08	3.31	27.45	18.51

Table 3: Main results of evaluating the understanding capabilities of MLLMs in the contexts of low-resource languages. Please note that “*Keyword Accuracy*” is employed for evaluation on D_{know} . “*BLEU*”, “*Rouge-L*” and “*Metor*” is employed for evaluation on D_{ling} .

4.2 Results

Main results

As shown in Table 3, we present the performance comparison across different experimental settings. From the results, we have the following observations:

MELLA enhances MLLM’s cultural knowledge. Keyword accuracy is leveraged to evaluate on D_{know} , the extracted keywords of which include lots of key knowledge information such as the names and identities of celebrities. After fine-tuned on MELLA, MLLMs generally gain noticeable improvement for all low-resource languages, indicating the fine-tuned MLLMs can answer some cultural knowledge behind the image.

MELLA enhances MLLM’s linguistic skills. Meteor is leveraged to evaluate on D_{ling} , which has rich image captions in low-resource languages. After finetuning, MLLMs gain a huge improvement on nearly all of the languages, some even improve by two orders of magnitude (e.g., InternVL2-8B, HU), indicating MELLA is effective for MLLMs to learn linguistic skills.

Comparing with SDRRL. The original MLLMs struggle on test sets of both D_{know} and D_{ling} , indicating these low-resource languages are not well-trained for general MLLMs due to the scarcity of data. SDRRL, as another method focusing on the multilingual problems, shows moderate improvement compared with the original MLLMs. However, it sometimes decreases the performance of the original MLLMs on test sets of both D_{know} and D_{ling} . We investigate the instances and find that it often outputs cross-lingual content, which is not expected in our tasks.

Backbone		AR	SR	RU	CS	KO	TH	VI	HU
<i>Keyword Accuracy</i>									
InternVL2-8B	D_{ling} only	3.20	0.56	2.80	1.80	0.72	5.10	1.10	3.50
	D_{know} only	<u>7.00</u>	6.43	<u>10.62</u>	<u>17.66</u>	<u>6.90</u>	2.21	<u>2.78</u>	5.81
	<i>ling – know</i> Two Stage	7.01	<u>5.46</u>	13.48	21.09	8.00	2.29	3.56	6.32
	MELLA	6.26	3.07	8.37	15.56	5.06	<u>4.50</u>	2.50	5.57
Qwen2-VL-7B-Instruct	D_{ling} only	2.08	0.88	0.36	4.35	1.60	5.31	0.41	2.79
	D_{know} only	1.26	<u>1.86</u>	3.09	2.67	4.63	1.84	<u>1.46</u>	2.29
	<i>ling – know</i> Two Stage	<u>2.20</u>	3.56	4.02	<u>4.57</u>	<u>4.53</u>	4.44	1.50	2.96
	MELLA	2.23	1.13	<u>3.26</u>	4.90	4.13	<u>4.97</u>	0.65	<u>2.92</u>
<i>Meteor</i>									
InternVL2-8B	D_{ling} only	37.9	17.29	14.81	15.59	29.39	35.10	33.41	16.16
	D_{know} only	2.81	0.28	0.31	0.56	1.01	1.48	1.94	0.34
	<i>ling – know</i> Two Stage	13.65	0.31	0.27	0.52	1.38	1.76	1.81	0.37
	MELLA	<u>29.78</u>	<u>13.54</u>	<u>4.91</u>	<u>12.17</u>	<u>22.81</u>	<u>22.50</u>	<u>16.37</u>	<u>13.11</u>
Qwen2-VL-7B-Instruct	D_{ling} only	37.36	17.13	15.77	15.79	27.39	35.84	32.83	15.28
	D_{know} only	2.13	0.04	0.40	0.49	0.81	1.02	1.50	0.22
	<i>ling – know</i> Two Stage	2.72	0.06	0.89	0.89	3.79	21.2	1.74	0.64
	MELLA	<u>36.89</u>	<u>13.88</u>	<u>5.36</u>	<u>12.88</u>	<u>23.74</u>	<u>34.63</u>	<u>28.66</u>	<u>12.72</u>

Table 4: Comparing to using D_{know} and D_{ling} separately. “ D_{ling} / D_{know} only” denotes we SFT using only D_{ling} or D_{know} , $ling – know$ Two Stage denotes training on D_{ling} first and merge LoRA blocks, than training on D_{know} . The size of the training dataset is the same as our main experiments.

Ablation study

Table 4 shows our ablation studies. From the results, it is clear that D_{ling} contributes to linguistic ability and D_{know} contributes to cultural knowledge. For instance, while training D_{ling}^{AR} only achieves 37.90 and 37.36 on Meteor, it falls to 3.20 and 2.08 on keyword accuracy. This reminds us to combine the two datasets. However, $ling – know$ Two Stage’s performance is also not satisfying, displaying a forgetting phenomenon. Perhaps this is because multi-stage LoRA training is hard for models to form a uniform representation space. Instead, our training paradigm combines two datasets and just trains once, displaying a balanced performance.

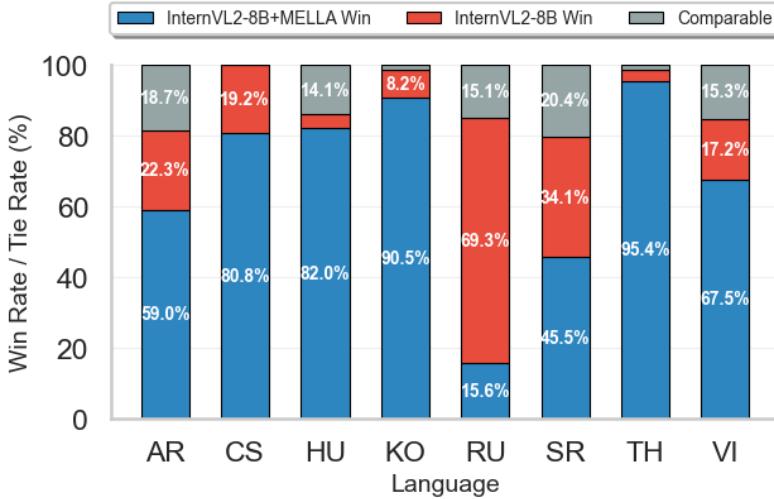


Figure 5: Human evaluation over 100 validation samples and 8 volunteers.

Backbone	AR	SR	RU	CS	KO	TH	VI	HU
InternVL2-8B	0.690	0.720	1.495	0.474	0.820	0.505	0.745	0.601
Qwen2-VL-7B-Instruct	0.724	0.642	1.252	0.646	0.721	0.891	0.779	0.845

Table 5: Standard deviation of keyword accuracy after 3 times of experiments with different random seeds.

Qualitative analysis

Following ShareGPT4V [Chen et al., 2023], we conduct a qualitative evaluation of MELLA by generating 100 samples with InternVL2-8B and InternVL2-8B-MELLA. The results, shown in Figure 5, strongly align with the findings reported in the main results.

Statistic analysis

Table 5 displays our standard deviation of keyword accuracy after 3 times of experiments. This quantitatively demonstrates that the experimental results have low randomness and high robustness.

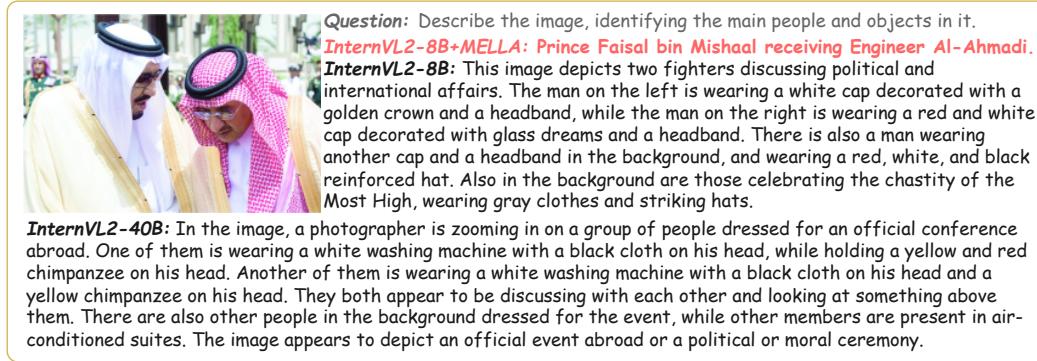


Figure 6: A case study on AR demonstrates the effectiveness of our model in enhancing cultural groundedness. Both the questions and answers were originally in Arabic; for ease of reading, translations are provided here.

Case study

Figure 6 presents a case study (with additional examples available in the appendix D.1). InternVL2-8B, even InternVL2-40B, provide only a thin description of the image, whereas InternVL2-8B-MELLA successfully identifies the prince depicted. This highlights the effectiveness of our dual-data strategy in achieving the dual objective.

4.3 Further Analysis

Performance variations analysis. We identify three primary sources for the performance variations observed across languages and models: 1) Linguistic differences affect learning difficulty; 2) Base models differ in architecture and pretraining coverage; 3) D_{ling} and D_{know} vary in quality and size across languages.

Alt-text as knowledge-rich but linguistically-weak data. As shown in Table 4, training solely on D_{know} (alt-text) further degrades language ability but combining D_{know} and D_{ling} successfully achieves dual objective.

MELLA is more effective at filling capability gaps. For low-performing languages like Hungarian (HU), it can raise performance to an acceptable level, while for partially learned languages like Russian (RU), standard training may introduce knowledge interference.

5 Conclusion

This study is motivated by the performance gap in MLLMs between linguistic capability and cultural groundedness in low-resource language contexts. To address this, we define a dual objective for low-resource language MLLMs and propose a framework to achieve it. Furthermore, we construct MELLA as an instantiation of our framework. Experimental results validate the effectiveness of our proposed approach. With the release of MELLA, we aim to foster cultural awareness and development in making multimodal AI more inclusive and representative of global linguistic diversity, and benefit speakers of multiple languages.

References

- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL <https://arxiv.org/abs/2502.13923>.
- Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024a.
- 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, 2015. URL <https://arxiv.org/abs/1405.0312>.
- OpenAI, :, Aaron Hurst, Adam Lerer, and Adam P. Goucher et al. Gpt-4o system card, 2024. URL <https://arxiv.org/abs/2410.21276>.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL <https://aclanthology.org/P02-1040/>.
- Yuanchi Zhang, Yile Wang, Zijun Liu, Shuo Wang, Xiaolong Wang, Peng Li, Maosong Sun, and Yang Liu. Enhancing multilingual capabilities of large language models through self-distillation from resource-rich languages. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings*

of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 11189–11204, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.603. URL <https://aclanthology.org/2024.acl-long.603/>.

Zheng-Xin Yong, Cristina Menghini, and Stephen H. Bach. Lexc-gen: Generating data for extremely low-resource languages with large language models and bilingual lexicons, 2024. URL <https://arxiv.org/abs/2402.14086>.

Michael Andersland. Amharic llama and llava: Multimodal llms for low resource languages, 2024. URL <https://arxiv.org/abs/2403.06354>.

Roland Barthes. Rhetoric of the image. *Semiotics: An introductory anthology*, pages 192–205, 1985.

Clifford Geertz. Chapter 1/thick description: Toward an interpretive theory of culture. *The interpretation of cultures: Selected essays*, pages 3–30, 1973.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, Bontu Fufa Balcha, Chenxi Whitehouse, Christian Salamea, Dan John Velasco, David Ifeoluwa Adelani, David Le Meur, Emilio Villa-Cueva, Fajri Koto, Fauzan Farooqui, Frederico Belcavello, Ganzorig Batnasan, Gisela Vallejo, Grainne Caulfield, Guido Ivetta, Haiyue Song, Henok Biadgignan Ademtew, Hernán Maina, Holy Lovenia, Israel Abebe Azime, Jan Christian Blaise Cruz, Jay Gala, Jiahui Geng, Jesus-German Ortiz-Barajas, Jinheon Baek, Jocelyn Dunstan, Laura Alonso Alemany, Kumaranage Ravindu Yasas Nagasinghe, Luciana Benotti, Luis Fernando D’Haro, Marcelo Viridiano, Marcos Estechea-Garitagoitia, Maria Camila Buitrago Cabrera, Mario Rodríguez-Cantelar, Mélanie Jouitteau, Mihail Mihaylov, Mohamed Fazli Mohamed Imam, Muhammad Farid Adilazuarda, Munkhjargal Gochoo, Munkh-Erdene Otgonbold, Naome Etori, Olivier Niyomugisha, Paula Mónica Silva, Pranjal Chitale, Raj Dabre, Rendi Chevi, Ruochen Zhang, Ryandito Diandaru, Samuel Cahyawijaya, Santiago Góngora, Soyeong Jeong, Sukanya Purkayastha, Tatsuki Kurabayashi, Teresa Clifford, Thanmay Jayakumar, Tiago Timponi Torrent, Toqueer Ehsan, Vladimir Araujo, Yova Kementchedzhieva, Zara Burzo, Zheng Wei Lim, Zheng Xin Yong, Oana Ignat, Joan Nwatu, Rada Mihalcea, Thamar Solorio, and Alham Fikri Aji. Cvqa: Culturally-diverse multilingual visual question answering benchmark, 2024. URL <https://arxiv.org/abs/2406.05967>.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR ’21, page 2443–2449, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380379. doi: 10.1145/3404835.3463257. URL <https://doi.org/10.1145/3404835.3463257>.

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, 2022a. URL <https://arxiv.org/abs/2210.08402>.

Jingqun Tang, Qi Liu, Yongjie Ye, Jinghui Lu, Shu Wei, Chunhui Lin, Wanqing Li, Mohamad Fitri Faiz Bin Mahmood, Hao Feng, Zhen Zhao, Yanjie Wang, Yuliang Liu, Hao Liu, Xiang Bai, and Can Huang. Mtvqa: Benchmarking multilingual text-centric visual question answering, 2024a. URL <https://arxiv.org/abs/2405.11985>.

Rocktim Jyoti Das, Simeon Emilov Hristov, Haonan Li, Dimitar Iliyanov Dimitrov, Ivan Koychev, and Preslav Nakov. Exams-v: A multi-discipline multilingual multimodal exam benchmark for evaluating vision language models, 2024. URL <https://arxiv.org/abs/2403.10378>.

Jingqun Tang, Qi Liu, Yongjie Ye, Jinghui Lu, Shu Wei, Chunhui Lin, Wanqing Li, Mohamad Fitri Faiz Bin Mahmood, Hao Feng, Zhen Zhao, Yanjie Wang, Yuliang Liu, Hao Liu, Xiang Bai, and Can Huang. Mtvqa: Benchmarking multilingual text-centric visual question answering, 2024b. URL <https://arxiv.org/abs/2405.11985>.

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, 2022b. URL <https://arxiv.org/abs/2210.08402>.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024b.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*, 2018.

Sanjana Shivani Chintalapati, Jonathan Bragg, and Lucy Lu Wang. A dataset of alt texts from hci publications: Analyses and uses towards producing more descriptive alt texts of data visualizations in scientific papers. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility*, page 1–12. ACM, October 2022. doi: 10.1145/3517428.3544796. URL <http://dx.doi.org/10.1145/3517428.3544796>.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016.

DeepL. DeepL api documentation. <https://developers.deepl.com/docs>, 2023.

Google. Google translate. <https://translate.google.com/>, 2023.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri, editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. URL <https://aclanthology.org/2022.wmt-1.60/>.

Aniketh Janardhan Reddy, Gil Rocha, and Diego Esteves. Defactonlp: Fact verification using entity recognition, tfidf vector comparison and decomposable attention, 2018. URL <https://arxiv.org/abs/1809.00509>.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL <https://aclanthology.org/W04-1013/>.

Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, and Lucia Specia, editors, *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-3348. URL <https://aclanthology.org/W14-3348/>.

Ziyue Wang, Chi Chen, Fuwen Luo, Yurui Dong, Yuanchi Zhang, Yuzhuang Xu, Xiaolong Wang, Peng Li, and Yang Liu. Activview: Evaluating active perception ability for multimodal large language models. *arXiv preprint arXiv:2410.04659*, 2024.

Kejia Zhang, Keda Tao, Jiasheng Tang, and Huan Wang. Poison as cure: Visual noise for mitigating object hallucinations in lvms. *arXiv preprint arXiv:2501.19164*, 2025.

- Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 3505–3506, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379984. doi: 10.1145/3394486.3406703. URL <https://doi.org/10.1145/3394486.3406703>.
- Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions, 2023. URL <https://arxiv.org/abs/2311.12793>.
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.
- Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. URL <https://arxiv.org/abs/2308.12966>.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023a. URL <https://arxiv.org/abs/2304.08485>.
- 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. URL <https://llava-vl.github.io/blog/2024-01-30-llava-next/>.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023b.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023c.
- Xirong Li, Chaoxi Xu, Xiaoxu Wang, Weiyu Lan, Zhengxiong Jia, Gang Yang, and Jieping Xu. Coco-cn for cross-lingual image tagging, captioning and retrieval, 2019. URL <https://arxiv.org/abs/1805.08661>.
- Wen Lai, Mohsen Mesgar, and Alexander Fraser. Llms beyond english: Scaling the multilingual capability of llms with cross-lingual feedback, 2024. URL <https://arxiv.org/abs/2406.01771>.
- Fredrik Carlsson, Philipp Eisen, Faton Rekathati, and Magnus Sahlgren. Cross-lingual and multilingual CLIP. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6848–6854, Marseille, France, June 2022. European Language Resources Association. URL <https://aclanthology.org/2022.lrec-1.739/>.
- Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. Survey of cultural awareness in language models: Text and beyond. *Computational Linguistics*, pages 1–96, 2025.
- Shudong Liu, Yiqiao Jin, Cheng Li, Derek F. Wong, Qingsong Wen, Lichao Sun, Haipeng Chen, Xing Xie, and Jindong Wang. Culturevlm: Characterizing and improving cultural understanding of vision-language models for over 100 countries, 2025. URL <https://arxiv.org/abs/2501.01282>.
- Christoph Zauner. Implementation and benchmarking of perceptual image hash functions, 2010.
- Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.
- Tencent Cloud. Image moderation system (ims). <https://cloud.tencent.com/product/ims/>, 2023.
- Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes, 2021. URL <https://arxiv.org/abs/2110.01963>.

A Related Work

Multimodal Large Language Models

While closed-source large models, such as Gemini [Team et al., 2023], demonstrate stronger multilingual capabilities, open-source multimodal large language models (MLLMs) still offer limited support for multilingual understanding, particularly in low-resource languages. Many existing models lack dedicated components for handling low-resource languages, such as Qwen-VL [Bai et al., 2023] and LLaVA [Liu et al., 2023a]. Others support only a limited set of languages or provide inadequate multilingual performance; for example, InternVL2 [Chen et al., 2024b] supports only Chinese and English, while LLaVA-1.5 [Liu et al., 2024, 2023b,c] primarily learns to follow Chinese instructions through multilingual instruction tuning without corresponding image inputs. Motivated by these limitations, this paper aims to equip open-source MLLMs with broader multilingual capabilities—especially for low-resource languages.

Multilingual Multimodal Datasets

Multimodal datasets play a crucial role in training large-scale vision-language models, enabling them to capture richer semantic representations across modalities. Existing datasets such as MSCOCO [Lin et al., 2015], COCO-CN [Li et al., 2019], and WIT [Srinivasan et al., 2021] focus predominantly on high-resource languages like English and Chinese, with limited coverage of low-resource languages. Recent efforts such as MTVQA [Tang et al., 2024b] and EXAMS-V [Das et al., 2024] begin to address this gap but remain restricted in scale and diversity. Our dataset offers a large-scale, culturally diverse collection of multimodal data across eight low-resource languages, supporting both pretraining and finetuning.

Cross-Lingual Transfer

With the flourishing development of natural language processing technology, how to transfer the capabilities of models to low-resource languages has garnered attention from researchers [Andersland, 2024, Yong et al., 2024, Zhang et al., 2024, Lai et al., 2024, Carlsson et al., 2022]. To address the scarcity of data, researchers have proposed various efficient methods for generating high-quality data. The majority of these generation methods are related to machine translation, such as LexC-Gen [Yong et al., 2024], which uses a bilingual lexicon for word-to-word translation; Lai et al. [2024] and Andersland [2024] have translated existing datasets. Our approach differs in that we advocate for placing greater emphasis on cultural awareness.

Culture Awareness of MLLMs

The cultural awareness of LLMs and MLLMs in low-resource language contexts has been largely overlooked in the Western-centric development of AI [Pawar et al., 2025]. Recently, however, there has been a growing interest in addressing this gap. For example, CVQA [Romero et al., 2024] is a multilingual multiple-choice benchmark designed to evaluate the extent of culturally relevant knowledge in MLLMs. CultureVLM [Liu et al., 2025] aims to enhance the cultural understanding of VLLMs but primarily focuses on English. In contrast, our work targets low-resource languages by enhancing cultural groundedness—an objective we formally define—to improve cultural awareness in these underrepresented contexts.

B Data collection details

B.1 Image Resource Website List

ar	ru	ko	vi	th	hn	tr	es
https://aighadi.com	https://russian.rt.com	https://www.clien.net/service/	https://vietnambiz.vn/	https://siamrath.co.th/	https://www.blikk.hu/	https://www.kuris.rs/	https://www.patreo.cz/
https://albiladdaily.com	https://fakty.ua	https://www.healthfocus.co.kr/	https://daldaonket.vn/	https://www.khaosod.co.th/	https://hvg.hu/	https://www.danas.rs/	https://cs.wikipedia.org/
https://www.okaz.com	https://old.day.kyiv.ua/uk	https://news.kbs.co.kr	https://www.qndt.vn/	https://buriram.mol.go.th/	https://prohardver.hu/	https://www.blic.rs/	https://www.euro.cz/

Table 6: Crawled websites.

Table 6 lists the websites we crawled from.

Model	Setting	Hyperparameter	Value	Setting	Hyperparameter	Value
InternVL2-8B	Main	Epoch	1	Ablation	Epoch	1
		Batch Size per GPU	2, 4		Batch Size per GPU	4
		Learning Rate	4e-5		Learning Rate	4e-5
		Warmup Ratio	0.03		Warmup Ratio	0.03
		LR Scheduler Type	cosine		LR Scheduler Type	cosine
		Weight Decay	0.01		Weight Decay	0.01
		Max Seq Length	4096		Max Seq Length	4096
		Gradient Accumulation Steps	2		Gradient Accumulation Steps	2
Qwen2-VL-7B-Instruct	Main	Epoch	1	Ablation	Epoch	1
		Batch Size per GPU	4		Batch Size per GPU	4
		Learning Rate	4e-5		Learning Rate	4e-5
		Warmup Ratio	default		Warmup Ratio	default
		LR Scheduler Type	default		LR Scheduler Type	default
		Weight Decay	default		Weight Decay	default
		Max Seq Length	4096		Max Seq Length	4096
		Gradient Accumulation Steps	default		Gradient Accumulation Steps	default

Table 7: Hyperparameters used in the experiments for InternVL2-8B and Qwen2-VL-7B-Instruct. Default values refer to those in Huggingface Trainer.

B.2 Image filtering

We notice there are a number of images with low resolution, irrelevant contexts, and ethical issues. Hence, we conduct a series of filters that consider both the qualities and contents of the images.

- **Resolution:** To ensure that the images convey clear semantics, we retain only high-resolution images. Specifically, an image (and its associated text, if any) is included only if both its width and height exceed 256 pixels.
- **Conciseness:** Since the collected images may include duplicate content, such as the same person in different backgrounds, which can introduce redundant information, we apply a hierarchical deduplication strategy to ensure dataset conciseness. First, we remove duplicate images with identical pixel-level content. Then, we employ pHash(Perceptual hash) [Zauner, 2010], which is an algorithm robustly generating a hash value for image features and calculating Hamming distance for coarse-grained deduplication, eliminating near-identical images. At last, we apply a convolutional neural network (CNN) [Krizhevsky et al., 2012] for fine-grained removal of semantically similar images.
- **Ethics:** To avoid the toxic and harmful information, we filter out the images with sensitive or inappropriate material, including violence, hate speech, and advertisements via an Image Moderation System (IMS) API[Cloud, 2023].

B.3 Image description prompts

Figure 7 is an example of an image description prompt. For each domain (e.g., natural images, technical diagrams), we design specialized prompt templates to maximize description quality. We randomly select 200 images and design about 15 aspects for the reviewers to inspect. If any issues are reported, we adjust our prompt and regenerate the output till no more issues can be raised.

C Training Details

C.1 Hyperparameters

Table 7 lists the main hyperparameters used in the fine-tuning process. For training Qwen2-VL-7B-Instruct, we use Huggingface Trainer.

C.2 Prompt pool

Figure 8 illustrates a subset of the manually designed prompt pool.

Please carefully observe the image from a specific lesser-known language country. Based on the main elements and scene in the image, generate a detailed and precise Chinese description. Your description should focus on the following aspects:

1. Clearly describe the main subject of the image, such as people, specific objects, or key locations/scenes.
2. Describe the activity the subject is engaged in, along with its characteristics and condition, as well as how the subject is presented in a specific time and space.
3. Describe other elements in the image, and the spatial relationships and interactions between them and the main subject.
4. Incorporate background knowledge to elaborate on relevant cultural features of the country using the lesser-known language, such as traditional clothing, language, festivals, or customs.
5. Discuss how this culture is represented and symbolized in the image, and deepen understanding by connecting with text information.
6. You may extend your expression by drawing on personal experience, knowledge, or associations, covering artistic style, aesthetic preferences, cultural perceptions, etc., while respecting the culture itself.
7. Express your understanding of how the culture is visually conveyed in the image, highlighting its distinctive characteristics, enhancing visual impression, and helping the reader form a clear and vivid perception.

General

- What details in the image draw attention? What might these details signify?
- Based on the context of the lesser-known language country, what emotions or messages does the image convey?

- If the action or situation in the image were to continue, what might happen next?

Comprehensive

- What are the main objects or scenes shown in the image?
- What role do these elements play? How are they positioned or spatially related to each other?

People

- Who are the people in the image? What might their relationships be? What are they doing?
- Consider their clothing, facial expressions, gestures, and postures. What do these convey?

Culture

- What cultural elements are reflected in the image?
- What aesthetic or symbolic meanings might these cultural elements carry?
- How does the image reflect the unique traditional or modern aspects of the local culture of the lesser-known language country?

History

- Does the image reference any historical events, symbols, or figures?
- How is the image related to the historical development of the lesser-known language country?

Figure 7: An example of image description prompt.

	AR	SR	RU	CS	KO	TH	VI	HU
InternVL2-8B D_know	80K	84.32K	81K	90K	90K	90K	90K	90K
InternVL2-8B D_ling	36K	40.5K	17.92K	40.5K	34.74K	22.5K	38.25K	37.26K
Qwen2VL-7B D_know	80K	105.4K	63K	90K	90K	90K	90K	90K
Qwen2VL-7B D_ling	36K	40.5K	26.25K	40.5K	34.74K	22.5K	38.25K	37.26K

Table 8: Statistics of D_know and D_ling used in training phase.

C.3 Data statistics for training

Table 8 shows specific training data statistics.

D Experiment results

D.1 Case study

Figure 9 presents additional case studies. These examples clearly demonstrate how our training process enhances the MLLM’s linguistic capabilities and cultural understanding. Although some hallucinations are observed—an inherent limitation of alt-text data [Birhane et al., 2021]—our method serves as a strong example of the effectiveness of the proposed dual-source data strategy. Figure 9a

EN	Replace the text in the image with; What is the text description of this image?
AR	استبدل النص في الصورة بـ ما هو وصف الصورة؟
CS	Nahradte text na obrázku za; Jaký je popis obrázku?
HU	Cserélje le a képen lévő szöveget erre Mi a kép szöveges leírása?
KO	이미지의 텍스트를 다음으로 바꾸세요 이 이미지의 텍스트 설명은 무엇인가요?
RU	Замените текст на изображении на Какое описание изображения?
SR	Замените текст на слици са Какав је опис слике?
TH	ແກ່ນທີ່ຂໍຄວາມໃນກາພ້ວມ ກໍ່າອື່ນຫຍ່ອງກາຫົວໜ້າ?
VI	Thay thế văn bản trong hình ảnh bằng Mô tả văn bản của hình ảnh là gì?

Figure 8: A subset of our prompt pool for training.

shows our attempt at a larger model, which demonstrates a certain degree of cultural grounding. The examples in Figures 9b and 9c illustrate how the model incorporates culturally relevant knowledge when generating image descriptions.



Question: Ismertesse a képet, mutasson rá a rajta lévő személyekre és tárgyakra!

InternVL2-26B-MELLA: Az MSZP és a DK közötti közös bejelentés alatt Zelenyak Zsolt, a DK elnökönéljölte a térsági ellenzék egyik vezetője

InternVL2-206: Nem tudom, hogy az ember a képen látható, miközben úgy tűnik, hogy egy nyílvános beszélést tart. A képen egy férfi áll így éppen a beszélés során, egy megnagyított csillagás építés háttérén. Egy olyan helyen, ahol a férfi éppen beszél egy mikrofonnal, azt érdemnés lenne, hogy kiképzéjük a beszél tartalmát. Ha valaki szeretné megismerni a férfi identitását, akkor valamelyen ismertetője szükséges lenne.



Question: 그림에 있는 사람과 사물을 가리키며 그림을 설명해 보세요

InternVL2-26B-MELLA: 식사 토론하는 남북 양자

InternVL2-26B: 이 사진은 한국의 정치 계급자들이 식사를 하는 장면을 보여줍니다. 사진의 좌측에서 무소속으로 차를 같은 사람들을 볼 수 있습니다.\n\n**사람들은**: 금발을 가진 남성들이 식탁 주변에 앉아 있습니다. 모두 적극적으로 대화하거나 식사를 설레고 있습니다.\n\n**식탁**: 중앙에 있는 식탁에는 다양한 식사 액세서리들이 있습니다. 예를 들어, 반찬 있는 접시, 술잔, 접시 음식 등이 있습니다.\n\n**식당 장점**: 배경에는 식당의 정식이 보입니다. 식당의 벽에는 식물이 식당에 장식되어 있습니다. 식물의 생생함과 정적인 향기로써

배경이 있습니다.\n4. **식사 테이블**: 각 좌석에는 식사 테이블은 설명서가 두 장 있습니다. 이는 음식의 이름과 설명을 제공하는 데 사용됩니다.\n5. **음식점**: 몇 명의 사람들이 음식을 먹고 있는 것을 볼 수 있습니다. 접힌 음식이나 반찬을 볼 수 있습니다.\n6. **사진**: 한국의 정체계급자들이 식사를 하는 순간을 보여줍니다. 배경의 식물과 식사 장식은 식당의 특성을 나타내며, 식사 테이블에 있는 접시와 술잔은 정적인 장면을 구성하고 있습니다.

(a) Cases on HU and KO. The cultural connotation has been recognized.



Question: ຖືນາຫາກວາ

(b) A case on TH. The issue of repeated outputs has been resolved, and Thai cultural elements have been incorporated into the descriptions.



Question: Hãy mô tả bức tranh này

Question: *Play the role of the teacher today.*

InternalVL2-BB-MELLA: Hình ảnh này cho thấy một nhóm trẻ em trên sân bóng đá, có vẻ như họ đang tham dự một sự kiện hoặc cuộc thi bóng đá. Trong cảnh có 11 người, bao gồm 10 trẻ em và 1 người lớn. Trẻ em mặc đồng phục bóng đá màu trắng có logo màu đỏ và vàng, có thể cho biết họ thuộc về một đội hoặc câu lạc bộ. Họ đang cầm bóng và运转 rất hào hứng và vui vẻ, có thể đang chờ đợi một sự kiện hoặc buổi lễ nào đó. Ở phía bên trái của bức ảnh, có một người đàn ông mặc áo dài và quần dài, trông giống như huấn luyện viên hoặc người dẫn chương trình. Anh ta đang cầm một tệp tài liệu hoặc giấy tờ, có thể đang giải thích hoặc bối cảnh của sự kiện cho trẻ em. Ở phía bên phải có bức tranh, có một người phụ nữ mặc áo tắm và váy ngắn lanh da suoni. Tên và biểu cảm của cô ấy truyền tải không khí thân thiện và hỗ trợ. Ở bảng quảng cáo gần đó có dòng chữ và đồ họa bằng tiếng Việt, cho biết sự kiện có thể được tổ chức tại sân vận động nào đó. **C.Truyền thông**, **VTV** và **VTVcab** là các tên và logo của các đài truyền hình Việt Nam.

(c) A case on VI. The issue of repeated outputs has been resolved.

Figure 9: Case studies on HU, KO, TH, and VI showing improved cultural understanding and resolution of repeated output issues.