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

Anonymous ACL submission

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 and linguistic capability enhancement.

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.

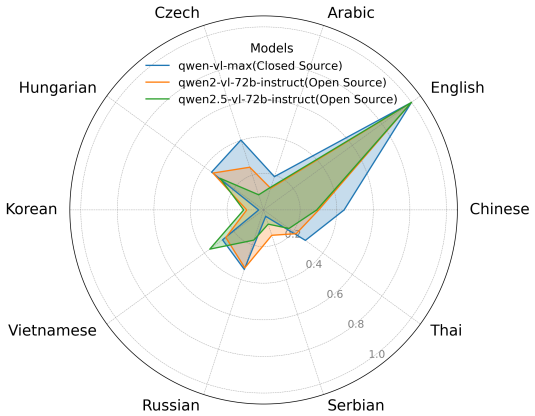

Figure 1: Image caption task performance on COCO dataset (Lin et al., 2015) across multiple languages. Compared to GPT-4o (OpenAI et al., 2024), most of the outstanding MLLMs get the highest BLEU (Papineni et al., 2002) score in English.

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 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

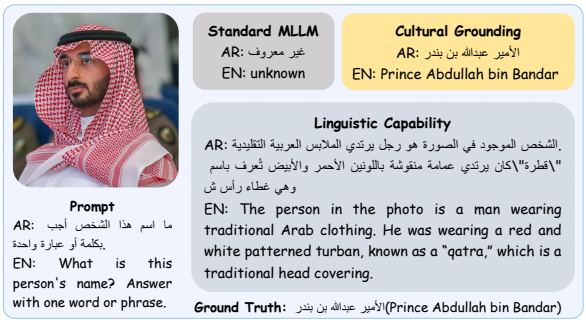

Figure 2: Standard MLLMs (e.g., InternVL2-8B, Qwen2-VL-7B) trained on generic datasets often fail to generate meaningful output due to limited visual-linguistic alignment. An MLLM with enhanced linguistic capability may produce detailed descriptions. However, only an MLLM enriched with cultural knowledge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

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 image-text pairs dataset 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)

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

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).

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 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:

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

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.

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 $D^L_{\text{ling}} = \{(I^L_i, T^L_{\text{den},i}), i = 1, \ldots, M\}$, 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 $T_{\text{den},i}^L$ — 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 $D_{\text{know}}^L = \{(I_j^L, T_{\text{con},j}^L), j = 1, \ldots, N\}$, 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: $D = D_{ling} \cup D_{know}$.

3 MELLA : Instantiating the Framework

MELLA (**M**ultilingual **E**nhancement for **L**ow-resource **LA**nguage 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 C.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 22 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 C.2, to ensure data quality. This process yields a final set of approximately 6.82 million (M) high-quality images.

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:

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

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

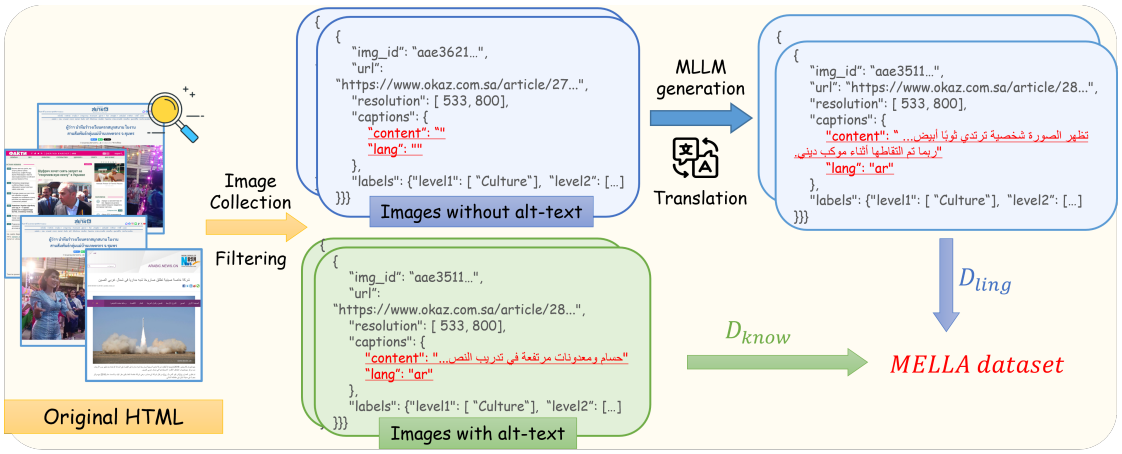

Figure 3: Data Collection Pipeline for MELLA . We first collect images with native alt-text from regional websites to form the cultural knowledge dataset (D_{know}). For images without alt-text, we use a powerful MLLM to generate descriptive captions, which are then translated into target low-resource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

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 C.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

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

where I_i denotes an image, and $T_{den,i}$ denotes the generated text paired with the image. The final dataset is $D^L = D^L_{know} \cup D^L_{ling}$, where language $L \in \{AR, CS, HU, KO, RU, SR, TH, VI\}$.

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.

4 Experiments

We empirically evaluate the effectiveness of MELLA in enhancing MLLMs for low-resource languages, addressing the following research questions: (1) Does fine-tuning on MELLA improve performance across the dual objectives? (2) What are the contributions of each data source? (3) How to utilize the dual-source dataset better during training? (4) How does it compare to existing multilingual enhancement methods? By isolating the effects of linguistic and cultural training, our analysis reveals that dual-source training fosters "thick descriptions" with cultural depth, underscoring the need to treat multimodal enhancement as a balanced process of linguistic fluency and cultural grounding. This echoes the semiotic decomposition in Section 2.1 and paves the way for more inclusive MLLMs.

4.1 Experimental setup

For training, we use a random subset of 80K-140K samples for each of the eight languages from the collected datasets; detailed statistics are in Appendix D.5. We avoid mixing our data with external datasets to isolate the effect of our method

Statistic	Number	Size (GB)	Avg. Len.
Total of D	6816029	2153.924	-
D_{know}	2729891	1244.714	14
$D_{know} - AR$	317954	77.33	22
$D_{know} - CS$	364571	96.749	8
$D_{know} - HU$	266889	203.147	11
$D_{know} - KO$	367621	183.04	30
$D_{know} - RU$	623140	148.44	9
$D_{know} - SR$	271731	108.488	6
$D_{know} - TH$	214627	169.52	8
$D_{know} - VI$	303358	258.00	17
D_{ling}	4086138	909.21	258
$D_{ling} - AR$	336321	79.771	256
$D_{ling} - CS$	513795	92.013	260
$D_{ling} - HU$	548428	133.221	263
$D_{ling} - KO$	575581	135.951	251
$D_{ling} - RU$	497414	109.382	251
$D_{ling} - SR$	521856	88.811	261
$D_{ling} - TH$	542863	119.142	261
$D_{ling} - VI$	549880	150.919	261

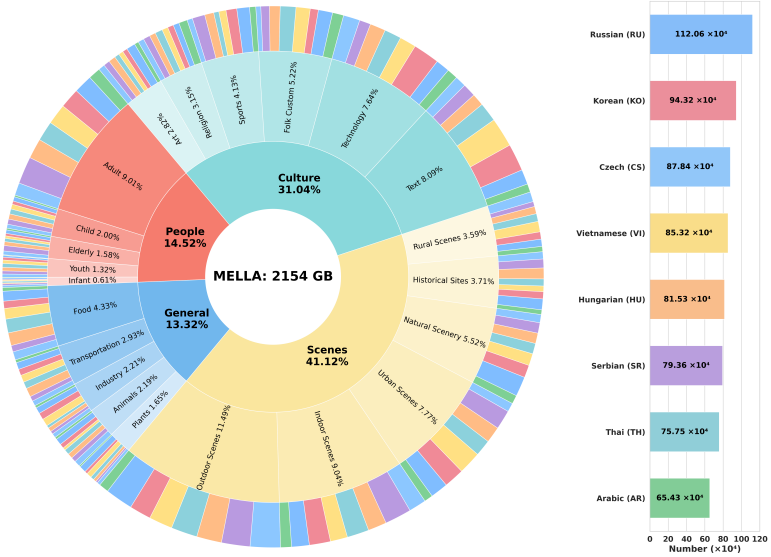

Figure 4: Statistical overview of the MELLA dataset. Left: Main statistics including total sample numbers, sizes, and average text lengths across different languages. Middle: Circular diagram of the category distribution. Right: Quantitative distribution showing the eight languages in the dataset with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

from confounding domain shift factors. The way of using the training sets is detailed in appendix D. Given the absence of existing benchmarks that simultaneously cover our target low-resource languages and align with our task formulation, we carefully construct held-out test sets by sampling 1,600 instances not used in training. Per language, this includes 100 samples from D^L_{know} (focusing on cultural knowledge) and 100 from D^L_{ling} (focusing on linguistic fluency), totaling 200 samples.

4.1.1 Evaluation Metrics and Details

Since D^L_{know} and D^L_{ling} are designed to investigate different understanding capabilities of an MLLM, we test them with different evaluation metrics. For cultural groundedness, following DeFactoNLP (Reddy et al., 2018), we employ keyword accuracy as the evaluation metric. We identify the keywords from the ground-truth alt-text using TF-IDF. We chose this method for its practicality and language-agnostic nature, allowing for a consistent evaluation protocol across all eight languages. While a dedicated Named Entity Recognition (NER) model for each language could potentially provide more precise entity extraction, the variable availability and performance of such tools for low-resource languages (Keraghel et al., 2024) led us to adopt TF-IDF as a robust and scalable proxy for cultural keyword identification. Regarding D^L_{ling}, 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). Appendix D.6 lists a uniform prompt leveraged for various low-resource languages for a fair evaluation.

4.1.2 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.

Additionally, mblip-mt0-xl (Geigle et al., 2023) is compared(see appendix E.2). **mBLIP** (Geigle et al., 2024) is a method that also incorporates translation for enhancing multilingual performance.

4.1.3 Implementation details

Our code is implemented using DeepSpeed (Rasley et al., 2020) on two NVIDIA A100-SXM4-80GB GPUs. The models are trained separately on each language. See appendix D for details.

4.2 Results

4.2.1 Main results

Table 9 presents 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.

4.2.2 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, suffering from catastrophic forgetting. Notably, it fails to improve METEOR score, which stems from the difficulty of maintaining a unified representation space across multi-stage LoRA training. During the Two Stage process, the models are first trained on D_{ling}, then the LoRA weights are combined before training on D_{know}. This sequential approach causes the model to disproportionately favor the latter dataset, resulting in degraded METEOR performance. One might consider reversing the training order; however, this would position the linguistic capability learning later in the sequence, which would further compromise performance on low-resource languages. This result strongly validates our choice of a joint training strategy(D.1), which compels the model to find a unified parameter space that balances both linguistic fluency and cultural knowledge, ultimately achieving both objectives simultaneously.

4.2.3 Human evaluation

Following the evaluation setup in ShareGPT4V (Chen et al., 2023), we conduct a qualitative human assessment to further validate the effectiveness of MELLA . Specifically, we conducted the evaluation over 100 samples from InternVL2-8B and InternVL2-8B-MELLA and 8 volunteers. Detailed evaluation procedure and result analysis are provided in Appendix D.7.

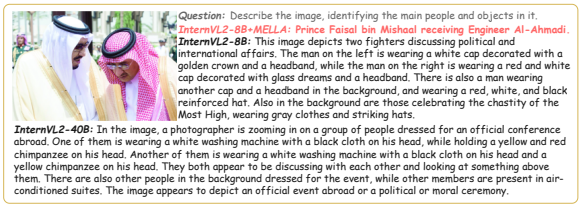

Figure 5: A case study on AR demonstrates the effectiveness of our model in enhancing cultural groundedness. Both the questions and answers were originally in Arabic; for ease of reading, translations are provided here.

4.2.4 Case study

Figure 5 presents a case study (with additional examples available in the appendix E.4). 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.

Backbones		AR	SR	RU	CS	KO	TH	VI	HU
					Keyword Accuracy				
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±0.69**	**3.07±0.72**	**8.37±1.50**	**15.56±0.47**	**5.06±0.82**	**4.50±0.51**	**2.50±0.75**	**5.57±0.60**
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±0.72**	**1.13±0.64**	**3.26±1.25**	**4.90±0.65**	**4.13±0.72**	**4.97±0.89**	**0.65±0.78**	**2.92±0.85**
					BLEU				
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±0.43**	**13.22±0.36**	4.41±0.51	**14.33±0.29**	**11.02±0.32**	**0.65±0.35**	**15.53±0.39**	**13.45±0.40**
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±0.45**	**16.33±0.40**	**6.26±0.38**	**14.80±0.62**	**11.48±0.44**	**1.00±0.25**	**30.18±0.68**	**13.39±0.49**

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*" is employed for evaluation on D_{ling}. "*Rouge-L*" "*Metor*" scores, which are also notable, are reported in the appendix.

Backbone		AR	SR	RU	CS	KO	TH	VI	HU
					Keyword Accuracy				
InternVL2-8B	D_{ling} only	3.20	0.56	2.80	1.80	0.72	**5.10**	1.10	3.50
	D_{know} only	7.00	**6.43**	10.62	17.66	6.90	2.21	2.78	5.81
	$ling-know$ Two Stage	**7.01**	5.46	**13.48**	**21.09**	**8.00**	2.29	**3.56**	**6.32**
	MELLA	6.26	3.07	8.37	15.56	5.06	4.50	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	1.86	3.09	2.67	**4.63**	1.84	1.46	2.29
	$ling-know$ Two Stage	2.20	**3.56**	**4.02**	4.57	4.53	4.44	**1.50**	**2.96**
	MELLA	**2.23**	1.13	3.26	**4.90**	4.13	4.97	0.65	2.92
					Meteor				
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
	$ling-know$ Two Stage	13.65	0.31	0.27	0.52	1.38	1.76	1.81	0.37
	MELLA	29.78	13.54	4.91	12.17	22.81	22.50	16.37	13.11
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
	$ling-know$ Two Stage	2.72	0.06	0.89	0.89	3.79	21.2	1.74	0.64
	MELLA	36.89	13.88	5.36	12.88	23.74	34.63	28.66	12.72

Table 4: Comparing to using D_{know} and D_{ling} seperately. "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.

4.3 Further Analysis

Performance variations analysis. We observe performance variations across languages, which can be attributed to several factors: 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. For instance, the performance gains in Russian (RU) are sometimes more modest compared to languages like Czech (CS) or Hungarian (HU), with occasional dips in certain metrics (e.g., BLEU for InternVL2-8B). This phenomenon can be interpreted in two ways. Firstly, as a major world language, Russian likely had a stronger presence in the models' original pretraining data, resulting in a higher baseline and thus a smaller margin for improvement. Secondly, this pre-existing knowledge might lead to knowledge interference, where the specific cultural information from MELLA conflicts with the model's generalized priors, requiring a more complex reconciliation process during fine-tuning. This contrasts with languages like Hungarian, where the model starts from a lower baseline and is more readily shaped by the new data.

Alt-text as knowledge-rich but linguistically-weak data. As shown in Table 4, training solely on D_{know} (alt-text) results in a pronounced degradation of linguistic fluency, indicating that alt-text alone provides insufficient coverage of complex sentence structures and pragmatic cues. While alt-text offers dense factual grounding and object-level descriptions, its linguistic form tends to be short, elliptical, and stylistically homogeneous, which limits the model's ability to generalize to natural conversational or narrative contexts. By integrating D_{know} with D_{ling}, the model benefits from both factual precision and linguistic diversity, suggesting that alt-text should be leveraged as a complementary knowledge source rather than a primary training corpus.

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 dual-source data collection. Furthermore, we construct MELLA as an instantiation of our framework. Experimental results validate the effectiveness of our proposed approach. By releasing MELLA , we aim to foster a more linguistically and culturally inclusive multimodal AI for global communities.

Data Availability Statement

The dataset collected for this study is openly available under the Creative Commons Attribution 4.0 International License (CC BY 4.0). All collected images have undergone legal and safety review to ensure compliance with ethical and regulatory standards.

Limitations

While our work focuses on eight low-resource languages, this represents only a small fraction of the world's low-resource languages. The selection was constrained by data availability and computational resources.

As shown in our ablation study 4, two-stage training suffers from catastrophic forgetting. While our unified approach mitigates this, optimal integration strategies warrant further exploration, such as adding general datasets and developing new training methods.

Potential Risks (and Ethical Considerations)

Data Collection and Privacy. Our dataset is constructed from publicly available web content under CC BY 4.0 license. All images underwent rigorous filtering using an Image Moderation System to conduct privacy safeguards and remove harmful content, including violence, hate speech, and inappropriate material.

Bias and Cultural Representation. Web-crawled data may contain and perpetuate existing societal biases and stereotypes. Our cultural knowledge reflects web content, which may overrepresent dominant voices and not capture all perspectives within linguistic communities. We applied content filtering, but semantic biases may persist.

References

Michael Andersland. 2024. Amharic llama and llava: Multimodal llms for low resource languages. *Preprint*, arXiv:2403.06354.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *Preprint*, arXiv:2308.12966.

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, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

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

Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *Preprint*, arXiv:2110.01963.

Fredrik Carlsson, Philipp Eisen, Faton Rekathati, and Magnus Sahlgren. 2022. Cross-lingual and multilingual CLIP. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6848–6854, Marseille, France. European Language Resources Association.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023. Sharegpt4v: Improving large multimodal models with better captions. *Preprint*, arXiv:2311.12793.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, and 1 others. 2024a. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024b. 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.

Sanjana Shivani Chintalapati, Jonathan Bragg, and Lucy Lu Wang. 2022. 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.

Tencent Cloud. 2023. Image moderation system (ims). https://cloud.tencent.com/product/ims/.

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

DeepL. 2023. Deepl api documentation. https://developers.deepl.com/docs.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth*

Workshop on Statistical Machine Translation, pages 376–380, Baltimore, Maryland, USA. Association for Computational Linguistics.

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

Gregor Geigle, Abhay Jain, Radu Timofte, and Goran Glavaš. 2023. mblip: Efficient bootstrapping of multilingual vision-llms. *arXiv*, abs/2307.06930.

Gregor Geigle, Abhay Jain, Radu Timofte, and Goran Glavaš. 2024. mblip: Efficient bootstrapping of multilingual vision-llms. *Preprint*, arXiv:2307.06930.

Google. 2023. Google translate. https://translate.google.com/.

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

Imed Keraghel, Stanislas Morbieu, and Mohamed Nadif. 2024. Recent advances in named entity recognition: A comprehensive survey and comparative study. *Preprint*, arXiv:2401.10825.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Wen Lai, Mohsen Mesgar, and Alexander Fraser. 2024. Llms beyond english: Scaling the multilingual capability of llms with cross-lingual feedback. *Preprint*, arXiv:2406.01771.

Xirong Li, Chaoxi Xu, Xiaoxu Wang, Weiyu Lan, Zhengxiong Jia, Gang Yang, and Jieping Xu. 2019. Coco-cn for cross-lingual image tagging, captioning and retrieval. *Preprint*, arXiv:1805.08661.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *Preprint*, arXiv:2304.08485.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023c. Visual instruction tuning.

Shudong Liu, Yiqiao Jin, Cheng Li, Derek F. Wong, Qingsong Wen, Lichao Sun, Haipeng Chen, Xing Xie, and Jindong Wang. 2025. Culturevlm: Characterizing and improving cultural understanding of vision-language models for over 100 countries. *Preprint*, arXiv:2501.01282.

OpenAI, :, Aaron Hurst, Adam Lerer, and Adam P. Goucher et al. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. 2025. Survey of cultural awareness in language models: Text and beyond. *Computational Linguistics*, pages 1–96.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. 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. Association for Computing Machinery.

Aniketh Janardhan Reddy, Gil Rocha, and Diego Esteves. 2018. Defactonlp: Fact verification using entity recognition, tfidf vector comparison and decomposable attention. *Preprint*, arXiv:1809.00509.

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. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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, and 57 others. 2024. Cvqa: Culturally-diverse multilingual visual question answering benchmark. *Preprint*, arXiv:2406.05967.

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

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

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

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. 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. Association for Computing Machinery.

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. 2024a. Mtvqa: Benchmarking multilingual text-centric visual question answering. *Preprint*, arXiv:2405.11985.

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. 2024b. Mtvqa: Benchmarking multilingual text-centric visual question answering. *Preprint*, arXiv:2405.11985.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

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

Zheng-Xin Yong, Cristina Menghini, and Stephen H. Bach. 2024. Lexc-gen: Generating data for extremely low-resource languages with large language models and bilingual lexicons. *Preprint*, arXiv:2402.14086.

Christoph Zauner. 2010. Implementation and benchmarking of perceptual image hash functions.

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

Yuanchi Zhang, Yile Wang, Zijun Liu, Shuo Wang, Xiaolong Wang, Peng Li, Maosong Sun, and Yang Liu. 2024. Enhancing multilingual capabilities of large language models through self-distillation from resource-rich languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11189–11204, Bangkok, Thailand. Association for Computational Linguistics.

A Appendix

B 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., 2023b). 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, 2023a,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.

C Data collection details

C.1 Image Resource Website List

Table 5 lists the websites we crawled from.

C.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). Our data are collected from publicly available sources that permit research use, ensuring compliance with their terms of service and copyright laws.

C.3 Image description prompts

Figure 6 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.

D Experiment Details

D.1 Training objective

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 $D = D_{ling} \cup D_{know}$ in a parameter-efficient manner.

ar	ru	ko	vi
https://alghad.com	https://russian.rt.com	https://www.clien.net/service/	https://vietnambiz.vn/
https://albiladdaily.com	https://fakty.ua	https://www.healthfocus.co.kr/	https://daidoanket.vn/
https://www.okaz.com	https://old.day.kyiv.ua/uk	https://news.kbs.co.kr	https://www.qdnd.vn/
th	**hu**	**sr**	**cs**
https://siamrath.co.th/	https://www.blikk.hu/	https://www.kurir.rs/	https://www.patro.cz/
https://www.khaosod.co.th/	https://hvg.hu/	https://www.danas.rs/	https://cs.wikipedia.org/
https://buriram.mol.go.th/	https://prohardver.hu/	https://www.blic.rs/	https://www.euro.cz/

Table 5: Crawled websites.

We formally define the SFT task. To mitigate overfitting, we first manually crafted 20 prompts for each language L, constructing a prompt pool $P = \{x_i^L | i = 1, \ldots, 20\}$(refer to the appendix D.4 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:

$$\mathcal{L}_{\text{CE}} = -\mathbb{E}_{((I,x),T)\sim D^L}\left[\sum_{t=1}^{|T|} \log P_\theta(T_t \mid T_{<t}, I, x)\right], \tag{4}$$

where $T = \{T_1, \ldots, T_n\}$ 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.

D.2 Hyperparameters

Table 6 lists the main hyperparameters used in the fine-tuning process. For training Qwen2-VL-7B-Instruct, we use Huggingface Trainer.

D.3 Tokenization

If tokenization is needed, we use these Python packages: *pyarabic* for AR, *pythainlp* for TH, *stanza* for other languages.

D.4 Prompt pool

Figure 7 illustrates a subset of the manually designed prompt pool.

D.5 Data statistics for training

Table 7 shows specific training data statistics.

D.6 Evaluation details

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!*"

D.7 Human evaluation details

To further assess the quality of model-generated outputs beyond automatic metrics, we conducted a human evaluation. Specifically, we randomly sampled 100 pairs of ground-truth and model-generated responses from the test set. The evaluators were not informed which response in each pair was the ground truth to ensure unbiased judgments.

Evaluators. We recruited native speakers or individuals who had lived in the corresponding language region for more than three years, as well as participants majoring in the respective language. This ensured sufficient linguistic proficiency and cultural familiarity among evaluators.

Evaluation criteria. The evaluation followed a three-stage judgment protocol:

1. **Grammatical correctness:** If one response contained severe grammatical errors or garbled text, the other was preferred.
2. **Content relevance:** If one response was completely irrelevant to the input or topic, the other was favored.
3. **Content accuracy:** When both responses were grammatically correct and relevant, the one providing more accurate and appropriate information was selected as the winner.

Procedure. For each pair, evaluators were asked to choose which response was better according to the above criteria, or to mark them as *comparable* if both were of similar quality. The final human preference rate was calculated as the proportion of samples in which the model-generated response was preferred over the ground truth.

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

	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 7: Statistics of D_know and D_ling used in training phase.

	average length
mblip-mt0-xl	91.26
annotation	1221.02

Table 8: Length of mblip-mt0-xl generation and MELLA annotation.

E Experiment results

E.1 Full main results

Table 9 shows the full table of results.

E.2 Comparison with mBLIP

As mBLIP employs a similar machine translation-based approach to ours, we conduct a comparative evaluation against mblip-mt0-xl on our test sets. Table 10 presents the results, revealing that mblip-mt0-xl exhibits limited performance on our evaluation benchmark. We attribute this performance gap to the characteristics of mBLIP's training data, which consists of considerably shorter captions with less detailed descriptions, as evidenced in able 8. This suggests that training data complexity and caption length play crucial roles in model performance on our tasks.

E.3 Human evaluation results

The overall human evaluation results, illustrated in Figure 8, show a strong alignment with our quantitative findings.

For the Russian (RU) subset in particular, human evaluators tended to prefer the pre-trained model over the fine-tuned one. This trend is consistent with the automatic evaluation results: as shown in Table 9, the BLEU score of InternVL2-8B slightly decreased after fine-tuning on RU data. Given that grammatical correctness was given the highest priority in our human evaluation criteria, this preference likely reflects the pre-trained model's stronger grammatical fluency, even though the fine-tuned model may have improved in content relevance or factual accuracy.

E.4 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 hallu-

Backbones		AR	SR	RU	CS	KO	TH	VI
				Meteor				
InternVL2-8B	-	26.07	2.70	**7.71**	3.37±0.66	14.54	19.95	**18.19**
	SDRRL	22.46	5.23	5.83	6.62	13.83	11.77	11.1
	MELLA	**29.78±0.49**	**13.54±0.42**	4.91±0.50	**12.17**	**22.81±0.31**	**22.5±0.72**	16.37±0.55
Qwen2-VL-7B-Instruct	-	15.49	2.33	**6.54**	6.03	12.93	17.14	16.77
	SDRRL	2.35	0.25	1.28	5.32	0.76	18.48	1.92
	MELLA	**36.89±0.35**	**13.88±0.57**	5.36±0.77	**12.88±0.40**	**23.74±0.55**	**34.63±0.67**	**28.66±0.48**
				Rouge-L				
InternVL2-8B	-	5.23	6.41	**12.73**	6.25	6.25	0.50	12.39
	SDRRL	14.37	7.07	8.60	10.18	9.17	1.55	9.98
	MELLA	**17.26±0.33**	**18.77±0.39**	6.32±0.81	**17.74±0.60**	**14.97±0.71**	**2.25±0.32**	**14.57±0.72**
Qwen2-VL-7B-Instruct	-	11.30	5.50	**12.86**	1.11	7.85	1.31	16.84
	SDRRL	1.59	0.38	10.19	8.38	0.87	2.22	1.82
	MELLA	**24.13±0.30**	**20.08±0.57**	8.47±0.63	**19.02±0.58**	**16.08±0.89**	**3.31±0.38**	**27.45±0.61**

Table 9: 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*" are employed for evaluation on D_{ling}.

Backbones	RU	KO	TH	VI	AR	CS	SR	HU
				METEOR				
mblip-mt0-xl	0.96	1.73	1.81	7.61	10.53	0.75	1.49	1.54
InternVl2-8B +MELLA	4.91 (+3.95)	22.81 (+21.08)	22.50 (+20.69)	16.37 (+8.76)	29.78 (+19.25)	12.17 (+11.42)	13.54 (+12.05)	13.11 (+11.57)
Qwen2-VL-7B-Instruct +MELLA	5.36 (+4.40)	23.74 (+22.01)	34.63 (+32.82)	28.66 (+21.05)	36.89 (+26.36)	12.88 (+12.13)	13.88 (+12.39)	12.72 (+11.18)
				Rouge-L				
mblip-mt0-xl	2.92	1.85	0.10	9.89	8.08	2.23	4.13	3.61
InternVl2-8B +MELLA	6.32 (+3.40)	14.97 (+13.12)	2.25 (+0.44)	14.57 (+4.68)	17.26 (+9.18)	17.74 (+15.51)	18.77 (+14.64)	18.41 (+14.80)
Qwen2-VL-7B-Instruct +MELLA	8.47 (+5.55)	16.08 (+14.23)	3.31 (+1.50)	27.45 (+17.56)	24.13 (+16.05)	19.02 (+16.79)	20.08 (+15.95)	18.51 (+14.90)

Table 10: Compare with mblip-mt0-xl.

cinations 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 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.

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 6: An example of image description prompt.

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	Nahraďte 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	แทนที่ข้อความในภาพด้วย
	คำอธิบายของภาพคืออะไร?
VI	Thay thế văn bản trong hình ảnh bằng
	Mô tả văn bản của hình ảnh là gì?

Figure 7: A subset of our prompt pool for training.

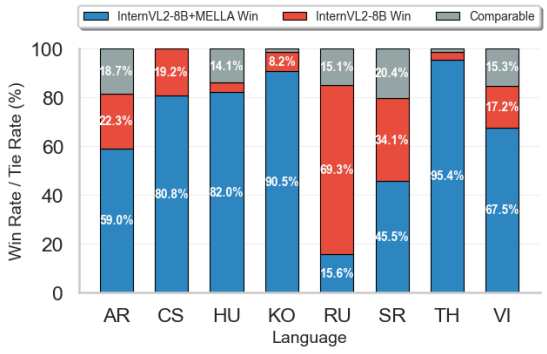

Figure 8: Human evaluation over 100 validation samples and 8 volunteers.

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 elnökjelöltje a társasági ellenzék egyik vezetője
InternVL2-26B: Nem tudom, hogy az ember a képben látható, miközben úgy tűnik, hogy egy nyilvános beszédést tart. A képen egy férfi áll egy éppen a beszédés során, egy megnyitott csillagás építés hátterén. Egy olyan helyen, ahol a férfi éppen beszél egy mikrofonnál, azt érdemes lenne, hogy kiképzeljük a beszéd tartalmát. Ha valaki szeretne megismerni a férfi identitását, akkor valamilyen ismertetője szükséges lenne.

Question: 그림에 있는 사람과 사물을 가리키며 그림을 설명해 보세요.
InternVL2-26B-MELLA:식사 토론하는 남북 양자
InternVL2-26B:0/ 사진은 한국의 정치 계급자들이 식사를 하는 장면을 보여줍니다. 사진의 좌측에서 우측으로 순차로 다음과 같은 사물들을 볼 수 있습니다:\n\n1. **사람들**: 금발을 가진 남성들이 식탁 주변에 앉아 있습니다. 모두 적극적으로 대화하거나 식사를 설레고 있습니다.\n2. **식탁**: 중앙에 있는 식탁에는 다양한 식사 액세서리들이 있습니다. 예를 들어, 반찬 있는 접시, 술잔, 접힌 음식 등이 있습니다.\n3. **식당 장식**: 배경에는 식당의 장식이 보입니다. 식당의 벽에는 식물이 식당에 장식되어 있습니다. 식물의 생명력과 정적인 호박 배경이 있습니다.\n4. **식사 테이블**: 각 좌석에는 식사 테이블 설명서가 두 장 있습니다. 이는 음식의 이름과 설명을 제공하는 데 사용됩니다.\n5. **음식**: 몇 명의 사람들이 음식을 먹고 있는 것을 볼 수 있습니다. 접힌 음식이나 반찬을 볼 수 있습니다.\n\n이 사진은 한국의 정치 계급자들이 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: อธิบายภาพนี้
InternVL2-8B-MELLA: ภาพนี้แสดงจากประเพณีในรูปแบบไทย ซึ่งอาจเป็นโอกาสทางการหรือกิจกรรมทางศาสนา ตรงกลางภาพมีภาพวาดขนาดใหญ่ของผู้หญิงสวมชุดไทยแบบดั้งเดิม เธอสวมเครื่องแต่งกายสีทองและมีเลื่อมสีม่วงประดับอยู่บนศีรษะ สีหน้าของเธอดูเคร่งขรึมและเคร่งขรึม และดูเหมือนเธอจะถูกเฉลิมฉลองหรือรำลึกถึงบางอย่าง ด้านหลังภาพวาดมีป้ายประปามีอักษรไทย บ่งบอกว่าเหตุการณ์นี้อาจเกี่ยวข้องกับวัฒนธรรมหรือประวัติศาสตร์ไทย เบื้องหน้าของภาพมีโต๊ะสองโบประดับด้วยเครื่องประดับและดอกไม้ โต๊ะแต่ละโต๊ะจัดไว้ด้วยดอกไม้และเครื่องประดับ รวมถึงภาชนะทองคำและของประดับตกแต่ง ซึ่งเป็นสัญลักษณ์ของความเคารพและการเฉลิมฉลองในวัฒนธรรมไทย ผ้าม่านสีขาวและผ้าม่านสีแดงโดดเด่นตัดกับความสวยงามของโต๊ะและภาพวาด จากทั้งหมดสื่อถึงบรรยากาศแห่งความเคร่งขรึมและความเคารพ อาจเป็นการรำลึกถึงบุคคลหรือเหตุการณ์สำคัญ จากภูมิหลังทางวัฒนธรรม ภาพนี้แสดงให้เห็นถึงความซับซ้อนและความงดงามของวัฒนธรรมไทย ตัวเลือกเครื่องแต่งกายและเครื่องประดับของผ
InternVL2-8B: ภาพนี้แสดงเสริมสร้องและการเข้าของผู้คนที่ถูกความหมิ่นหมอบและมีความสำคัญในชุมชน โดยจะใช้เปล่งและการรับความหมิ่นหมอบในการประกอบการของรัฐบาลและองค์กรทางการเพื่อให้ความสำคัญในการหมิ่นหมอบและการเปล่งให้เป็นประจำในขณะที่ประกอบการของรัฐบาลและองค์กรทางการ โดยมีการแสดงภาพของผู้คนที่ถูกความหมิ่นหมอบและอาธิบาลของเมืองในพิญาณและโรงเรียนของเมืองในพิญาณ โดยจะใช้เปล่งและการรับความหมิ่นหมอบในการประกอบการของรัฐบาลและองค์กรทางการเพื่อให้ความสำคัญในการหมิ่นหมอบและการเปล่งให้เป็นประจำในขณะที่ประกอบการของรัฐบาลและองค์กรทางการ โดยมีการแสดงภาพของผู้คนที่ถูกความหมิ่นหมอบและอาธิบาลของเมืองในพิญาณและโรงเรียนของเมืองในพิญาณ

(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.
InternVL2-8B-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 ảnh 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 trông 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 đỏ và quần đùi đen, trông giống như huấn luyện viên hoặc người dẫn chương trình. Anh ta đang cầm một tập tài liệu hoặc giấy tờ, có thể đang giải thích hoặc bầu bối cảnh của sự kiện cho trẻ em. Ở phía bên phải của bức tranh, có một người phụ nữ mặc áo tím và váy đen, có vẻ là người chị em hoặc người lãnh đạo sự kiện. Tư thế và biểu cảm của cô ấy truyền tải không khí chào đón và hỗ trợ. Ở phía sau, có thể nhìn thấy một số bảng quảng cáo với dòng chữ và đồ họa bằng tiếng Việt, cho biết sự kiện này có thể được tổ chức tại Việt Nam. Các biển quảng cáo đề cập đến các thương hiệu như \"LG Twin Wash\" và \"Yen Viet\", có thể là nhà tài trợ hoặc đối tác của sự kiện. Nhìn chung, hình ảnh này truyền tải một bầu không khí tích cực và ăn mừng, có lẽ là một sự kiện bóng đá được thiết kế để khuyến khích tham gia cộng đồng, giáo dục đồng thời thể thao. Trẻ em trong ảnh trông rất vui vẻ, điều này cho thấy sự kiện này có thể mang ý nghĩa đặc biệt đối với trẻ em
InternVL2-8B: Trong bức tranh, một nhóm các trẻ em và giáo viên đang ảnh hình trên một cảnh trận cổ điểm. Trẻ em và giáo viên đang một số chúng đang trong độ đội trẻ, trẻ em đang đồi một số chúng đang trong độ đồi. Các trẻ em đang đồi một số chúng đang trong độ đồi. Các trẻ em đang đồi một số chúng đang trong độ đồi. Các trẻ em đang đồi một số chúng đang trong độ đồi. Các trẻ em đang đồi một số chúng đang trong độ đồi. Các trẻ em đang đồi một số chúng đang trong độ đồi. Các trẻ em đang đồi một số chúng đang trong độ đồi. Các trẻ em đang đồi một số chúng đang trong độ đồi. Các trẻ em đang đồi một số chúng đang trong độ đồi. Các trẻ em đang đồi một số chúng đang trong độ đồi. Các trẻ em đang đồi một số chúng đang trong độ đồi. Các trẻ em đang đồi một số chúng đang trong độ đồi. Các trẻ em đang đồi một số chúng đang trong độ đồi. Các trẻ em đang đồi một số chúng đang trong độ đồi. Các trẻ em đang đồi một số chúng đang trong độ đồi. Các trẻ em đang đồi một số chúng đang trong độ đồi. Các trẻ em đang đồi một số chúng đang trong độ đồi. Các trẻ em đang đồi một số chúng đang trong độ đồi. Các trẻ em đang đồi mộ

(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.