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

Yufei Gao[1,2], **Jiaying Fei**[1], **Nuo Chen**[3*], **Ruirui Chen**[4], **Guohang Yan**[1*], **Yunshi Lan**[2*], **Botian Shi**[1,5*]
[1]Shanghai Artificial Intelligence Laboratory,
[2]East China Normal University,
[3]The Chinese University of Hong Kong, Shenzhen,
[4]Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A*STAR), Singapore,
[5]Shanghai Innovation Institute
yfgao.a@gmail.com

Abstract

Multimodal Large Language Models (MLLMs) perform strongly in high-resource languages, yet often produce fluent but culturally "thin" descriptions in low-resource settings. We argue that this failure is not merely a linguistic limitation: culture-specific visual knowledge depends on native visual–textual alignments that translation-centric pipelines rarely provide. We present MELLA, a multimodal dataset across eight low-resource languages, designed to support linguistic fluency and cultural groundedness. MELLA uses a dual-source strategy that combines native web image–alt-text pairs for culture-grounded supervision with generated-and-translated image descriptions for linguistically rich supervision, explicitly separating two learning signals often conflated in multilingual multimodal data. Through controlled diagnostic fine-tuning on multiple MLLM backbones, we show that MELLA mitigates cultural hallucination by helping models recognize and articulate culturally specific entities overlooked by translation-based adaptation. Our findings highlight data alignment, rather than model modification alone, as a path toward culturally grounded multimodal understanding in low-resource languages.

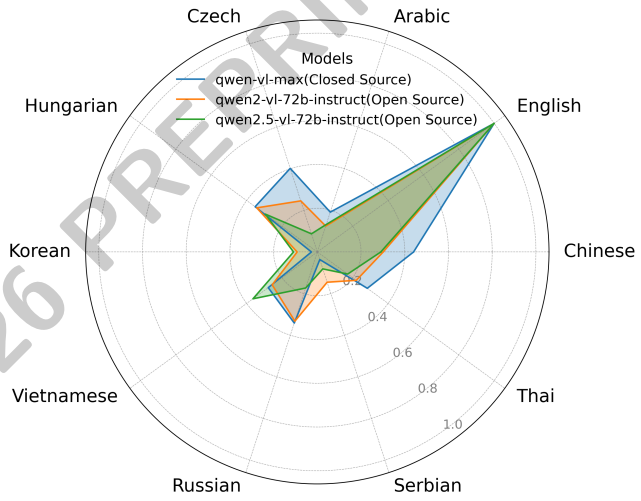

Figure 1: Image caption task performance on the COCO dataset [Lin *et al.*, 2015] across multiple languages. Compared to GPT-4o [Hurst *et al.*, 2024], most leading MLLMs achieve their highest BLEU [Papineni *et al.*, 2002] score in English.

1 Introduction

Multimodal Large Language Models (MLLMs) have achieved impressive performance in image understanding and generation [Bai *et al.*, 2025; Chen *et al.*, 2024a], particularly in high-resource languages like English, as illustrated in Figure 1. However, in low-resource languages, these models suffer from a systematic failure mode: **the generation of fluent but culturally superficial descriptions that fail to recognize salient local entities**. For instance, a model may accurately describe "a man in traditional clothing" while failing to identify a significant public figure or ceremony that is immediately recognizable to native speakers. We refer to this phenomenon as *cultural hallucination*: the production of culturally "thin" descriptions despite visually correct inputs.

This failure cannot be explained by limited linguistic fluency alone. As shown in Figure 2, even models that produce fluent and detailed descriptions in low-resource languages may still overlook culture-specific visual knowledge. This indicates that **linguistic adaptation and cultural understanding are not equivalent**: a model can learn to speak a language fluently without learning how visual concepts are grounded in that culture. This leads to outputs that are plausible but culturally uninformative, undermining usability and trust for low-resource language users.

A key commonality among existing multilingual MLLM adaptation methods is an implicit assumption: that cultural

*Corresponding authors.

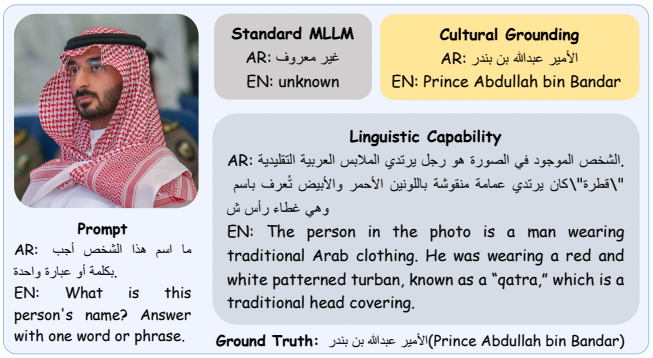

Figure 2: Standard MLLMs (e.g., InternVL2-8B, Qwen2-VL-7B) trained on generic datasets often fail to generate meaningful output due to limited visual-linguistic alignment. An MLLM with enhanced linguistic capability may produce detailed descriptions. However, only an MLLM enriched with cultural knowledge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

	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.

understanding can be induced indirectly through linguistic transfer. Methods such as SDRRL [Zhang *et al.*, 2024], LexC-Gen [Yong *et al.*, 2024], and Amharic LLaVA [Andersland, 2024] extend language coverage primarily via translation-based or text-centric supervision, see Table 1. While effective for improving surface-level fluency, these approaches **rarely introduce new native visual–textual alignments for culturally salient entities**. Consequently, they optimize linguistic expressiveness without addressing the underlying absence of culture-grounded multimodal data.

This reveals a fundamental limitation of translation-centric multilingual adaptation: in multimodal learning, cultural knowledge is **not a linguistic property** that can be transferred through language alone, **but a data alignment property** that depends on whether culturally salient entities ever appear as native visual–textual pairs during training. Consequently, this limitation is structural rather than algorithmic, and cannot be resolved by stronger translation models or increased model capacity. This perspective aligns with classic semiotic accounts of visual meaning: as noted by Barthes [Barthes, 1985] and Geertz [Geertz, 1973], images convey meaning not only through literal denotation but also through culturally coded connotation, a symbolic layer that translation alone cannot reliably recover. Without **native, culture-grounded supervision**, multilingual MLLMs are structurally biased toward producing *"thin descriptions"* that omit culturally salient meaning.

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 and 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` [1], 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. (Sec. 2)
- We present `MELLA` , a novel multimodal multilingual image-text dataset across eight low-resource languages as an instantiation of the dual-source strategy. (Sec. 3)
- Through controlled diagnostic experiments, we show that translation-centric multilingual adaptation improves linguistic fluency but fails to recover culture-grounded visual knowledge, while our dual-source framework explicitly addresses this gap. (Sec. 4)

2 Bridging Linguistic Capability and Cultural Groundedness

Motivated by the observation that low-resource multilingual MLLMs often produce fluent yet culturally "thin" descriptions, we formalize cultural hallucination as a failure to jointly acquire linguistic fluency and culture-grounded visual understanding. We then define a dual objective and propose a framework that explicitly bridges these two capabilities.

2.1 Motivation

Cultural hallucination arises when a model can describe what is visible in an image but fails to recognize what is cultur-

[1] https://opendatalab.com/applyMultilingualCorpus

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.*, 2022], MTV-QA [Tang *et al.*, 2024a], EXA-MS [Das *et al.*, 2024], CVQA [Romero and et al., 2024].

ally meaningful. To make this distinction precise, we draw from semiotics [Barthes, 1985; Geertz, 1973] and view image meaning as comprising two complementary components: a literal *denotation* and a symbolic, culturally coded *connotation*. Formally, we decompose the total meaning μ of an image I into:

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

where *denotation* (μ_{den}) corresponds to a "thin description" of what is explicitly visible, while *connotation* (μ_{con}) captures the "thick description", the culturally embedded knowledge, identities, and social significance associated with the image.

Prevailing multilingual MLLM adaptation methods (Table 1), which primarily rely on translating English-centric image–text datasets, target μ_{den}. As a result, such models may generate fluent descriptions of scenes while systematically failing to recover culturally salient entities (e.g., local public figures, ceremonies, or region-specific artifacts). This limitation reflects the core argument of this work: **without native visual–textual alignments, cultural knowledge cannot be recovered through linguistic transfer alone**.

2.2 Dual Objective

To bridge the μ_{den}–μ_{con} gap that translation-centric methods fail to address, we formalize two capabilities that an MLLM must acquire to work in low-resource settings.

Objective 1: Linguistic Capability. We define linguistic capability as the model's ability to generate fluent and structurally accurate text in a target language. Formally, we denote this capability as:

$$f_{\text{ling}} : (I, L) \rightarrow T^{L}_{\text{den}}, \tag{2}$$

where L is the target language and T_{den} is a textual realization of the denotative meaning $\mu_{\text{den}}(I)$. This capability corresponds to producing a "thin description" and requires mastery of vocabulary and grammar in language L.

Objective 2: Cultural Groundedness. We define cultural groundedness as the model's ability to recognize and articulate culturally specific knowledge embedded in an image. This capability is formalized as:

$$f_{\text{cult}} : (I, L) \rightarrow T^{L}_{\text{con}}, \tag{3}$$

where T_{con} represents a textual expression of the connotative meaning $\mu_{\text{con}}(I)$. Unlike linguistic capability, this function is difficult to learn from translation-based supervision alone, as it depends on exposure to authentic, culture-grounded visual–textual associations.

Together, these two objectives highlight that linguistic fluency and cultural grounding constitute **distinct but complementary learning signals**, both of which are necessary to mitigate cultural hallucination in multilingual MLLMs.

2.3 Dual-source Framework

To operationalize the dual objective, we propose a framework based on a dual-source data strategy that explicitly targets each capability.

Dual-source Data Strategy. Existing methods struggle to address μ_{con} primarily due to the scarcity of aligned, culturally relevant multimodal data for low-resource languages. To overcome this bottleneck, we construct a training corpus D by combining two complementary data sources:

$$D = D_{\text{ling}} \cup D_{\text{know}}. \tag{4}$$

The first source, $D_{\text{ling}} = \{(I_i, T_{\text{den},i})\}_{i=1}^{M}$, is a linguistics-focused dataset designed to support f_{ling}. Each pair consists of an image and a fluent denotative description originally generated in English by a strong MLLM and subsequently translated into the target language L. This source provides rich syntactic and lexical supervision for low-resource language generation.

The second source, $D_{\text{know}} = \{(I_j, T_{\text{con},j})\}_{j=1}^{N}$, is a culture-grounded dataset designed to support f_{cult}. It is constructed from authentic in-culture sources (e.g., native web corpora), where images are paired with human-authored alt-text reflecting culturally specific knowledge and usage. Unlike translated captions, these descriptions encode native visual–textual alignments that are essential for learning μ_{con}.

By jointly training on D_{ling} and D_{know}, the model is encouraged to allocate representational capacity to both linguistic fluency and cultural grounding within a unified parameter space. This dual-source framework directly addresses the structural limitation identified in the introduction: **cultural grounding in multimodal models is fundamentally a data alignment problem rather than a purely linguistic one**.

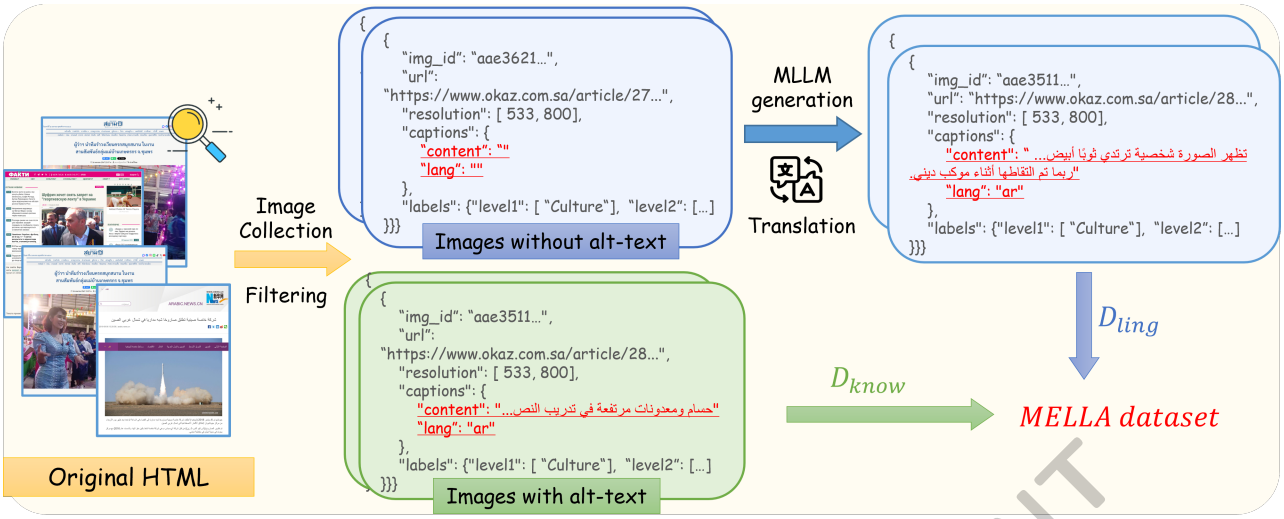

Figure 3: Data collection pipeline for MELLA . Native web images with alt-text are used to construct the culture-grounded dataset (D_{know}). For images without native descriptions, we generate denotative captions in English using a high-resource MLLM and translate them into target languages to form the linguistic dataset (D_{ling}).

3 MELLA : Instantiating the Framework

MELLA (**M**ultilingual **E**nhancement for **L**ow-resource **LA**nguage MLLMs) is a dual-source multimodal dataset that directly instantiates the framework introduced in Section 2.3. This section describes the data construction process, quality assurance principles, and dataset statistics. We follow standard supervised fine-tuning (SFT) with cross-entropy loss for all experiments, as our primary contribution lies in data design rather than model optimization.

3.1 Dataset Construction

The construction of MELLA follows three stages: (1) image collection from native web sources, (2) acquisition of culture-grounded and denotative textual supervision, and (3) quality assurance and filtering (Figure 3).

Image Collection We focus on eight low-resource languages: Arabic (AR), Czech (CS), Hungarian (HU), Korean (KO), Russian (RU), Serbian (SR), Thai (TH), and Vietnamese (VI), which are underrepresented in existing multimodal datasets [Tang *et al.*, 2024b; Srinivasan *et al.*, 2021]. To obtain culturally situated visual content, we crawl 24 high-traffic, language-native websites spanning multiple domains (e.g., news media, government portals, educational resources, forums, and commercial platforms). This multi-domain design mitigates domain bias and improves coverage of everyday cultural contexts.

Images are extracted from HTML pages and automatically categorized into 4 coarse and 22 fine-grained semantic categories using InternVL-1.5-25.5B [Chen *et al.*, 2024b]. We apply resolution filtering, deduplication, and content-based heuristics to remove low-quality or redundant images. The final corpus contains approximately 6.82 million images.

Alt-text for Cultural Groundedness (D_{know}) To support cultural groundedness (f_{cult}), we construct D_{know} from native web alt-text. Authored by users within the target culture, alt-text often contains culturally situated entities (e.g., local figures, events, and region-specific artifacts), serving as a valuable albeit noisy and linguistically sparse source of aligned visual–textual supervision [Sharma *et al.*, 2018; Chintalapati *et al.*, 2022]. For each target language L, we extract alt-texts written in the same language as the hosting web page and pair them with their corresponding images:

$$D^L_{know} = \{(I^L_i, T^L_{\text{con},i})\}^N_{i=1}. \quad (5)$$

Language consistency is verified using HTML metadata and fastText [Joulin *et al.*, 2016], ensuring 100% target language purity. We emphasize that D_{know} is not assumed to be noise-free; instead, it provides native visual–textual alignments that are largely absent from translation-centric pipelines, directly operationalizing our claim that cultural grounding is a data alignment property.

High-Resource Pivot for Linguistic Capability (D_{ling}) To support linguistic capability (f_{ling}), we construct D_{ling} using a **High-Resource Pivot** strategy. Rather than generating captions directly in low-resource languages, often resulting in poor fluency or hallucination, we first generate detailed, denotative descriptions in English using a strong MLLM (e.g., GPT-4o [Hurst *et al.*, 2024] or Qwen2.5-VL [Bai *et al.*, 2025]), where visual reasoning is most reliable.

The generation prompts are explicitly constrained to describe only visually observable content, avoiding identity inference or culturally specific speculation. The English captions are then translated into the target languages using high-quality machine translation systems (DeepL or Google Translate, depending on language support), followed by human spot-checking. Translation quality is evaluated using COMET-Kiwi [Rei and Treviso et al., 2022], achieving an average score of 0.75. The resulting dataset is:

$$D^L_{ling} = \{(I^L_i, T^L_{\text{den},i})\}^M_{i=1}. \quad (6)$$

Quality Assurance and Filtering We apply a rigorous two-stage quality assurance process. First, automated filters re-

Statistic	Number	Size (GB)	Avg. Length
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

Table 3: Statistics of the MELLA dataset across data sources and languages.

Culture (31.04%)		**Scenes** (41.12%)	
Subcategory	**Ratio**	**Subcategory**	**Ratio**
Text	8.09%	Outdoor	11.49%
Technology	7.64%	Indoor	9.04%
Folk Custom	5.22%	Urban	7.77%
Sports	4.13%	Natural Scenery	5.52%
Religion	3.15%	Historical Sites	3.71%
Art	2.82%	Rural	3.59%
General (13.32%)		**People** (14.52%)	
Subcategory	**Ratio**	**Subcategory**	**Ratio**
Food	4.33%	Adult	9.01%
Transportation	2.93%	Child	2.00%
Industry	2.21%	Elderly	1.58%
Animals	2.19%	Youth	1.32%
Plants	1.65%	Infant	0.61%

Table 4: Category distribution of the MELLA dataset.

move samples with language mismatch, encoding errors, duplicated content, or weak image–text relevance. Second, native speakers manually review random subsets of each language to identify residual linguistic or cultural errors. If the rejection rate exceeds 5%, the filtering rules are further refined and the cleaning pipeline is reapplied.

3.2 Dataset Statistics and Release

Table 3 summarizes the dataset scale, and Table 4 shows the category distribution. The dataset contains 6.8M image–text pairs across eight languages, covering diverse topics and semantic categories. Following responsible data practices, we plan to release MELLA under the **CC BY 4.0** license. The released data will include image URLs, native alt-text, generated denotative captions, and metadata (e.g., resolution and category labels).

4 Experiments

We conduct a data-centric, diagnostic evaluation to examine whether MELLA mitigates the *cultural hallucination* gap in low-resource multilingual MLLMs. Rather than targeting state-of-the-art performance on public benchmarks, our goal is to isolate whether culturally grounded visual knowledge can be acquired when native visual–textual alignments are introduced. Accordingly, our evaluation emphasizes controlled comparisons over leaderboard-style benchmarking, and complements large-scale multilingual models [Dash *et al.*, 2025; Yue *et al.*, 2025] that prioritize breadth and scale. Concretely, our experiments address three questions: (1) Can MELLA reduce culturally "thin" descriptions while preserving fluent generation in target languages? (2) How do the two data sources (D_{know} and D_{ling}) contribute complementary learning signals for cultural grounding and linguistic fluency? (3) Under a controlled setting, how does joint dual-source training compare with staged or translation-centric adaptation pipelines? Together, these analyses support the view that multilingual multimodal enhancement requires balancing two distinct learning signals: structural fluency and culture-grounded visual knowledge (cf. Section 2.1).

4.1 Experimental setup

Training data. For each of the eight low-resource languages, we fine-tune models on a random subset of 80K–140K training pairs sampled from MELLA . We intentionally avoid mixing MELLA with external datasets to isolate the effect of our data design from confounding factors such as domain shift.

Diagnostic held-out evaluation. Because existing multilingual multimodal benchmarks do not simultaneously (i) cover our target low-resource languages and (ii) match our task formulation of producing detailed, culturally grounded descriptions, we construct diagnostic held-out test sets from MELLA . Specifically, we sample 1,600 instances not used in training. For each language L, we include 100 examples from D^L_{know} (culture-specific supervision) and 100 examples from D^L_{ling} (linguistically rich supervision), totaling 200 test examples per language. We emphasize that this held-out evaluation is designed to *diagnose cultural knowledge acquisition* within the same distribution as the training data, rather than to serve as a universal generalization benchmark. To further assess potential memorization effects, we additionally analyze train–test similarity and entity-frequency distributions. We observe low image-level nearest-neighbor overlap between training and test sets, and performance gains remain positive across entity-frequency buckets rather than concentrating only on high-frequency entities, suggesting that improvements are not solely explained by memorization of repeated samples or entity names.

Backbone	Training	AR	SR	RU	CS	KO	TH	VI	HU
Keyword Accuracy on D_{know} (Cultural Groundedness)									
InternVL2-8B	Base	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	Base	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 on D_{ling} (Linguistic Fluency Sanity Check)									
InternVL2-8B	Base	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	Base	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 5: **Diagnostic evaluation on MELLA held-out test sets.** Keyword Accuracy on D_{know} measures culture-specific entity identification, while BLEU on D_{ling} serves as a sanity check for linguistic fluency. Results highlight how different supervision signals affect complementary capabilities in multilingual MLLMs.

Evaluation metrics. Since D_{know} and D_{ling} probe different capabilities, we evaluate them using different metrics.

Cultural groundedness (on D_{know}). N-gram overlap metrics (e.g., BLEU/ROUGE) correlate poorly with cultural factuality: a model may generate a fluent caption while omitting or misidentifying the salient local entity. We employ **Keyword Accuracy** as a proxy for culture-specific entity recall, focusing on whether models recover salient entities grounded in the image. Following DeFactoNLP [Reddy *et al.*, 2018], we extract TF–IDF keywords from the ground-truth alt-text to verify their presence in model outputs. We adopt TF–IDF rather than language-specific NER tools due to the limited availability and reliability of NER systems for low-resource languages [Keraghel *et al.*, 2024].

Linguistic fluency (on D_{ling}). To quantify whether models can produce fluent, structurally complete descriptions in the target low-resource languages, we report standard generation metrics on D_{ling}: **BLEU-4** [Papineni *et al.*, 2002], **ROUGE-L** [Lin, 2004], and **METEOR** [Denkowski and Lavie, 2014]. We treat these metrics as *sanity checks* for linguistic quality rather than as measures of cultural correctness.

Baselines and comparable methods. We evaluate two widely used MLLM backbones, InternVL2-8B and Qwen2-VL-7B-Instruct, to ensure conclusions are not backbone-specific. We compare: (i) **Base** (no fine-tuning), which reflects the default multilingual and cultural coverage of the backbone; and (ii) **SDRRL** [Zhang *et al.*, 2024], a representative translate-then-SFT multilingual adaptation pipeline that primarily targets linguistic adaptation via cross-lingual transfer and parallel resources, without explicit culture-grounded native supervision. Rather than exhaustively comparing against all existing multilingual datasets or models, we focus on such translation-centric pipelines as representative baselines, enabling a controlled comparison aligned with our core hypothesis.

Implementation details. All experiments are implemented with DeepSpeed [Rasley *et al.*, 2020] on two NVIDIA A100-SXM4-80GB GPUs. Models are fine-tuned separately for each target language using the same training configuration and hyperparameter settings across all experiments.

4.2 Results

Diagnostic results on the dual objective

Table 5 provides a diagnostic view of how different training strategies shape two distinct capabilities in multilingual MLLMs: culture-grounded entity recognition and fluent generation in low-resource languages. Rather than interpreting these numbers as a leaderboard comparison, we analyze the magnitude and direction of capability changes induced by different forms of supervision.

Cultural grounding: recovering previously ignored entities. Across all languages, models trained without native cultural supervision exhibit *near-zero keyword accuracy* on D_{know}, reflecting a systematic bias toward culturally "thin" descriptions. These models tend to describe images at a generic level (for example, "a man in traditional clothing") while failing to recognize culturally significant entities such as local public figures, ceremonies, or region-specific artifacts. After fine-tuning on MELLA , keyword accuracy increases substantially, typically by several-fold, indicating that native alt-text supervision enables models to **recognize and articulate** culturally grounded entities that were systematically ignored under translation-centric training.

Linguistic fluency: preserving structurally complete descriptions. Importantly, introducing culture-grounded supervision does not come at the cost of linguistic fluency. On D_{ling}, standard generation metrics such as BLEU and METEOR serve as *sanity checks* for sentence completeness and surface-level fluency, rather than primary indicators of cultural understanding. Models fine-tuned on MELLA maintain comparable or improved scores relative to their base counterparts across languages, suggesting that culture-grounded supervision does not degrade the model's ability to generate fluent and well-formed descriptions in low-resource languages.

Limits of translation-centric adaptation. Translation-centric adaptation improves linguistic fluency but remains insufficient for cultural grounding. SDRRL primarily targets cross-lingual *linguistic* transfer and accordingly yields stronger fluency metrics on D_{ling}. However, on D_{know} its Keyword Accuracy remains low and often close to that of base models, indicating that translation-based supervision alone does not reliably recover culture-specific enti-

Backbone	Method	AR	SR	RU	CS	KO	TH	VI	HU
Keyword Accuracy (on D_{know})									
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
	Two-stage	7.01	5.46	13.48	21.09	8.00	2.29	3.56	6.32
	Joint (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
	Two-stage	2.20	3.56	4.02	4.57	4.53	4.44	1.50	2.96
	Joint (MELLA)	2.23	1.13	3.26	4.90	4.13	4.97	0.65	2.92
METEOR (on D_{ling})									
InternVL2-8B	D_{ling} only	37.90	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
	Two-stage	13.65	0.31	0.27	0.52	1.38	1.76	1.81	0.37
	Joint (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
	Two-stage	2.72	0.06	0.89	0.89	3.79	21.20	1.74	0.64
	Joint (MELLA)	36.89	13.88	5.36	12.88	23.74	34.63	28.66	12.72

Table 6: **Ablation: disentangling learning signals.** We compare fine-tuning with only D_{ling} (linguistic supervision) or only D_{know} (culture-grounded supervision), a two-stage pipeline ($ling \rightarrow know$), and joint training (MELLA).

	AR	CS	HU	KO	RU	SR	TH	VI
MELLA Win (%)	59.0	80.8	82.0	90.5	15.6	45.5	95.4	67.5
Baseline Win (%)	22.3	0.0	3.9	1.3	69.3	34.1	0.0	17.2
Comparable (%)	18.7	19.2	14.1	8.2	15.1	20.4	4.6	15.3

Table 7: Human evaluation results over 100 validation samples and 8 volunteers.

ties grounded in the image. This contrast directly supports our central claim that **linguistic adaptation and cultural grounding are distinct, complementary learning signals**, and motivates incorporating native, culture-grounded multimodal supervision.

Ablation study

Table 6 validates the complementarity of our dual-source design. Training on D_{ling} alone yields strong METEOR but weak Keyword Accuracy, indicating fluent yet culturally "thin" descriptions. Conversely, D_{know} alone improves keyword recall but severely degrades linguistic fluency, reflecting the stylistic sparsity of native alt-text. A two-stage approach partially recovers cultural recall but often harms linguistic metrics due to catastrophic forgetting: sequential LoRA training biases the model toward the later stage and makes it difficult to maintain a unified representation that supports both objectives. Joint training compels the model to allocate capacity for both learning signals within a shared parameter space, resulting in a more balanced outcome.

Human evaluation

Following the qualitative evaluation protocol in ShareGPT4V [Chen *et al.*, 2023], we conduct human assessment to validate whether MELLA reduces culturally "thin" descriptions in practice. We evaluate 100 samples comparing InternVL2-8B (Base) against InternVL2-8B fine-tuned on MELLA , with 8 volunteers. The overall human evaluation results, reported in Table 7, show a strong alignment with our quantitative findings.

4.3 Further Analysis

Low-resource fine-tuning does not harm high-resource performance. A common concern is that adapting MLLMs to low-resource languages may degrade performance in high-resource languages due to negative transfer or catastrophic forgetting. As shown in Table 8, fine-tuning on individual low-resource languages consistently preserves English (EN) and Chinese (ZH) performance, with occasional small gains. This suggests that learning culture-grounded multimodal representations in low-resource settings does not interfere with, and may even support, general visual–language capabilities.

Training Language	EN Meteor	EN BLEU-4	ZH Meteor	ZH BLEU-4
Baseline (No FT)	17.37	2.01	16.23	0.00
FT on Arabic (AR)	17.28	2.25	17.21	0.06
FT on Thai (TH)	17.92	2.78	17.31	0.02
FT on Korean (KO)	17.59	2.54	16.80	0.02
FT on Hungarian (HU)	17.21	2.20	17.64	0.10

Table 8: Cross-lingual transfer. Fine-tuning on one low-resource language maintains EN/ZH, indicating limited negative transfer.

Cultural hallucination manifests as fluent but entity-deficient descriptions. Error analysis reveals a consistent failure pattern we term *cultural hallucination*: models generate fluent and visually plausible descriptions while omitting culturally salient entities that are obvious to native speakers. These errors are not random, but fall into three recurring categories: (i) **entity omission**, where specific local identities are ignored; (ii) **entity substitution**, where a culturally specific entity is replaced by a generic or foreign counterpart; and (iii) **over-generalization**, where culturally meaningful attributes are flattened into broad categories. These patterns explain why translation-centric adaptation yields linguistically fluent yet culturally "thin" outputs.

Joint supervision is necessary to balance linguistic fluency and cultural grounding. The ablation results in Table 6 show that training on D_{ling} alone improves linguistic fluency but fails to recover cultural entities, while training on D_{know} alone increases entity recall at the expense of grammaticality. A staged training strategy ($ling \rightarrow know$) partially recovers cultural information but often harms linguistic metrics due to catastrophic forgetting under sequential optimization. In contrast, joint training consistently yields more balanced representations, supporting both fluent generation and culture-grounded entity recognition. This finding supports our design choice to treat linguistic and cultural supervision as a *coupled* learning signal rather than independent stages.

5 Conclusion

This work addresses a persistent gap in low-resource multilingual MLLMs between linguistic fluency and cultural groundedness. We argue that this gap stems primarily from missing native, culture-grounded visual–textual alignments, rather than from model limitations. To this end, we propose a dual objective and a dual-source data strategy, instantiated as MELLA , which explicitly separates linguistic supervision from culture-grounded knowledge. Through diagnostic experiments, we show that while translation-centric adaptation improves fluency, it remains insufficient for cultural grounding, whereas MELLA enables models to recover culturally salient visual entities.

Ethical 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.

Our dataset is constructed from publicly available web content, and we plan to release the processed annotations and metadata under the CC BY 4.0 license. All images underwent rigorous filtering using an Image Moderation System to support privacy safeguards and remove harmful content, including violence, hate speech, and inappropriate material.

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.

Acknowledgements

The research was supported by Shanghai Artificial Intelligence Laboratory. This work was partially supported by the Natural Science Foundation of China (Project No. U23A20298).

References

[Andersland, 2024] Michael Andersland. Amharic llama and llava: Multimodal llms for low resource languages, 2024.

[Bai *et al.*, 2025] 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.

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

[Chen *et al.*, 2023] 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.

[Chen *et al.*, 2024a] 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*, 2024.

[Chen *et al.*, 2024b] 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, 2024.

[Chintalapati *et al.*, 2022] 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.

[Das *et al.*, 2024] 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.

[Dash *et al.*, 2025] Saurabh Dash, Yiyang Nan, John Dang, Arash Ahmadian, Shivalika Singh, Madeline Smith, Bharat Venkitesh, Vlad Shmyhlo, Viraat Aryabumi, Walter Beller-Morales, Jeremy Pekmez, Jason Ozuzu, Pierre Richemond, Acyr Locatelli, Nick Frosst, Phil Blunsom, Aidan Gomez, Ivan Zhang, Marzieh Fadaee, Manoj Govindassamy, Sudip Roy, Matthias Gallé, Beyza Ermis, Ahmet Üstün, and Sara Hooker. Aya vision: Advancing the frontier of multilingual multimodality, 2025.

[Denkowski and Lavie, 2014] 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.

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

[Hurst *et al.*, 2024] OpenAI: Aaron Hurst, Adam Lerer, and Adam P. Goucherc. Gpt-4o system card, 2024.

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

[Keraghel *et al.*, 2024] Imed Keraghel, Stanislas Morbieu, and Mohamed Nadif. Recent advances in named entity recognition: A comprehensive survey and comparative study, 2024.

[Lin *et al.*, 2015] 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.

[Lin, 2004] 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.

[Papineni *et al.*, 2002] 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.

[Rasley *et al.*, 2020] 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.

[Reddy *et al.*, 2018] Aniketh Janardhan Reddy, Gil Rocha, and Diego Esteves. Defactonlp: Fact verification using entity recognition, tfidf vector comparison and decomposable attention, 2018.

[Rei and Treviso et al., 2022] Ricardo Rei and Marcos Treviso et al. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In Philipp Koehn and Loïc Barrault et al., 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.

[Romero and et al., 2024] David Romero and Chenyang Lyu et al. Cvqa: Culturally-diverse multilingual visual question answering benchmark, 2024.

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

[Sharma *et al.*, 2018] 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.

[Srinivasan *et al.*, 2021] 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.

[Tang *et al.*, 2024a] 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, 2024.

[Tang *et al.*, 2024b] 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, 2024.

[Yong *et al.*, 2024] 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.

[Yue *et al.*, 2025] Xiang Yue, Yueqi Song, Akari Asai, Seungone Kim, Jean de Dieu Nyandwi, Simran Khanuja, Anjali Kantharuban, Lintang Sutawika, Sathyanarayanan Ramamoorthy, and Graham Neubig. Pangea: A fully open multilingual multimodal LLM for 39 languages. In *The Thirteenth International Conference on Learning Representations*, 2025.

[Zhang *et al.*, 2024] 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.