

LRM-LLaVA: Overcoming the Modality Gap of Multilingual Large Language-Vision Model for Low-Resource Languages

Junchen Li^{1*}, Qing Yang^{1*}, Bojian Jiang¹, Shaolin Zhu^{2†}, Qingxuan Sun¹

¹Du Xiaoman Finance, Beijing, China

²Tianjin University, Tianjin, China

{lijunchen, yangqing, jiangbojian, sunqingxuan}@duxiaoman.com
zhushaolin@tju.edu.cn

Abstract

Multilingual large language-vision models (LVLMs), which understand and generate both text and images across multiple languages, have achieved remarkable performance on English-centric multimodal generation tasks. However, their performance on non-English tasks has been underwhelming. One major challenge with multilingual LVLMs is the modality gap between visual inputs and multilingual textual inputs/outputs due to the lack of high-quality multilingual training data. In this paper, we propose LRM-LLaVA, a multilingual large language-vision model designed for low-resource languages to overcome the modality gap. It is composed of four components: a visual encoder, a multilingual large language model, a vision-text representation projector, and a cross-modal regularizer. Both the projector and regularizer aim at reducing the modality gap and improving multilingual performance. To train LRM-LLaVA, we employ a two-stage training strategy including pre-training and instruction fine-tuning. Meanwhile, we construct a multilingual visual question answering dataset based on English open-source datasets and adopt multiple task instructions. To evaluate the performance of LVLMs across various languages, we construct four multilingual benchmarks for 10 languages, based on English open-source benchmarks. Experimental results show that LRM-LLaVA achieves competitive performance compared to other multilingual LVLMs of similar parameters.

Introduction

Multilingual large language-vision models (LVLMs), which receive images and multilingual text as input and then conduct multi-round dialogues or provide corresponding text reasoning (Zhou et al. 2024), have recently attracted significant interest (You et al. 2023; Ying et al. 2024). Recent mainstream multilingual LVLMs are mainly trained on English image-text pairs to achieve semantic alignment between text and images, e.g., BLIP-2 (Li et al. 2023b), LLaVA (Liu et al. 2024b), and LLaVAR (Zhang et al. 2023). Those multilingual LVLMs have achieved remarkable performance on English multimodal generation tasks, such as image captioning (Wang et al. 2024), visual question answering (Liu et al. 2024c), and image-text retrieval (Zhu

*These authors contributed equally.

†Corresponding authors.

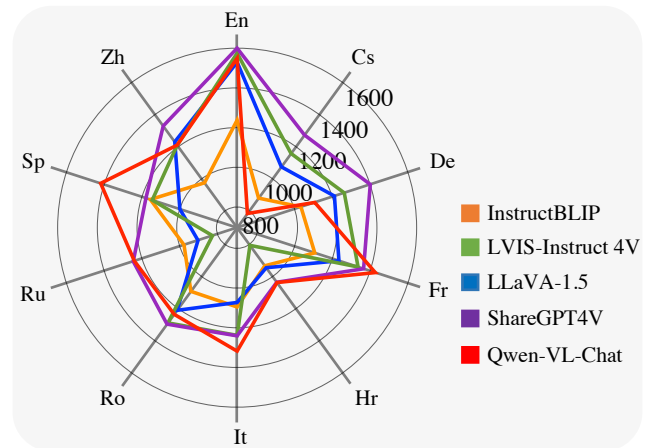


Figure 1: Performance of various large language-vision models on multilingual MME (Fu et al. 2023) benchmark.

et al. 2024). However, their performance on non-English tasks has been underwhelming as shown in Figure 1. This disparity is due to the limited availability of high-quality multilingual training data for low-resource languages.

As a cross-modal task, a major challenge of multilingual LVLMs for low-resource languages is the representation discrepancy across the textual and visual modality. This discrepancy arises because the modalities have distinct semantic spaces, and aligning them requires learning complex relationships between visual features and textual descriptions. The lack of sufficient text-image pairs for non-English languages makes it incredibly difficult for multilingual LVLMs to effectively bridge this gap (Zhu et al. 2023b).

Previous efforts for multilingual LVLMs for non-English languages, e.g., the method presented in Li et al. (2023d), use contrastive learning to align visual and textual representations by pulling positive pairs (text and image depicting the same concept) closer while pushing negative pairs (text and image depicting different concepts) further apart. Such methods may not be able to sufficiently leverage cross-modal knowledge as they require careful selection of training data and techniques to minimize the discrepancy across the textual and visual modalities and do not explicitly deal

with the modality gap issue for low-resource languages.

To mitigate this problem, we propose a multilingual LVLMs framework which can bridge the modality gap between non-English languages and images, and based on which we train LRM-LLaVA. LRM-LLaVA is composed of four components: a visual encoder ViT-L/14 (Radford et al. 2021), a backbone multilingual large language model Vicuna-13B (Chiang et al. 2023), a vision-text representation projector, and a cross-modal regularizer. We build a multilingual visual question answering dataset including 4.8M image-text pairs based on English open-source datasets and construct multiple monolingual or bilingual task instructions. We use a two-stage strategy to train LRM-LLaVA. The first pre-training stage aims to achieve low-cost alignment of multilingual features and visual features through a multilingual task instruction. The second fine-tuning stage aims to improve the multilingual instruction following ability through three multilingual task instructions. The regularizer is also optimized to force LRM-LLaVA to generate the same representations for the same input in different modalities. We construct four LVLMs multilingual benchmarks in 10 languages based on four English open-source benchmarks and evaluate their reliability. Experimental results show that LRM-LLaVA achieves competitive performance compared to other multilingual LVLMs of similar parameters and substantially improves multilingual image-text understanding without compromising English ability.

To summarize, our contributions are as follows:

- We propose a multilingual LVLMs framework for low-resource languages. We build a multilingual visual question answering dataset based on English open-source datasets and construct multiple task instructions to improve the multilingual ability of LVLMs.
- We build four multilingual benchmarks for 10 languages and evaluate their reliability, based on which we benchmark the mainstream LVLMs.
- Based on LVLMs multilingual framework, we train LRM-LLaVA on a 13B-parameter large language model and a 0.6B-parameter visual encoder, which significantly outperforms other LVLMs of similar parameters.

Related Work

The emergence of LVLMs has revolutionized multimodal understanding, enabling natural interactions between users and systems through images and text. Generally speaking, the architecture of LVLMs connects visual encoders such as CLIP (Radford et al. 2021) and SigLIP (Zhai et al. 2023) with large language models such as Vicuna (Chiang et al. 2023), LLaMA (Touvron et al. 2023), and Qwen (Bai et al. 2023a) through a cross-modal connection layer. Models like BLIP-2 (Li et al. 2023b), LLaVA (Liu et al. 2024b), and ShareGPT4V (Chen et al. 2023) have demonstrated the potential of LVLMs for a wide range of applications.

However, extending the capabilities of LVLMs to encompass multiple languages, especially for low-resource languages, presents significant challenges. The scarcity of high-quality, multilingual training data, coupled with the inherent difficulty of aligning visual and textual representa-

tions across different languages, has hindered the development of effective multilingual LVLMs. Existing research has explored various strategies to address this issue. MURAL (Cadene et al. 2019) and Pali (Chen et al. 2022) leverage dual encoders for parallel translation prediction and train large-scale multilingual models, respectively. M3P (Yu et al. 2024) and UC2 (Zhou et al. 2021) employ data augmentation techniques, such as translation and data construction, to expand the training data available for low-resource languages. While these methods show promise, they often face challenges in generating high-quality, contextually relevant data, especially for languages with limited resources. Some approaches (Bellagente et al. 2024) further prove the powerful language capabilities of multilingual large language models can be effectively migrated to various downstream tasks without relying on additional multilingual downstream training data. However, these methods may not be sufficient to bridge the modality gap effectively, particularly for low-resource languages, as they often require careful selection of training data and techniques to minimize the discrepancy across the textual and visual modalities.

Our work combines instruction data construction (Zhu et al. 2023a; Chen et al. 2024, 2023) and multilingual improvement using monolingual and bilingual data (Li et al. 2023d; Nguyen et al. 2024). We construct multilingual training data containing multiple monolingual and bilingual task instructions and use a cross-modal regularizer to make the model generate consistent representations for the same input across different modalities. This novel approach, which we term LRM-LLaVA, effectively addresses the key challenges of data scarcity and modality gap.

Methodology

This section introduces the model architecture of LRM-LLaVA and the two-stage pre-training strategy that leverages monolingual and bilingual task instructions on a constructed multilingual dataset to reduce the modality gap.

Model Architecture

The architecture of LRM-LLaVA, illustrated in Figure 2, is designed to effectively bridge the modality gap between visual and multilingual textual information. It comprises four key components: (1) **Visual Encoder**: pretrained ViT-L/14 (Radford et al. 2021) is employed to extract rich visual features from input images. (2) **Multilingual Large Language Model**: Vicuna-13B (Chiang et al. 2023) fine-tuned on LLaMA (Touvron et al. 2023), is responsible for processing textual information and generating coherent and relevant responses. (3) **Vision-Text Representation Projector**: A two-layer multi-layer perception (MLP) projector serves as a crucial bridge to align visual features and multilingual features. (4) **Cross-Modal Regularizer**: A cross-modal regularizer is incorporated into the architecture to further mitigate the modality gap between the multilingual large language model and the visual encoder. This regularizer operates by leveraging multiple task instructions to encourage the model to learn consistent representations for the same input across different modalities.

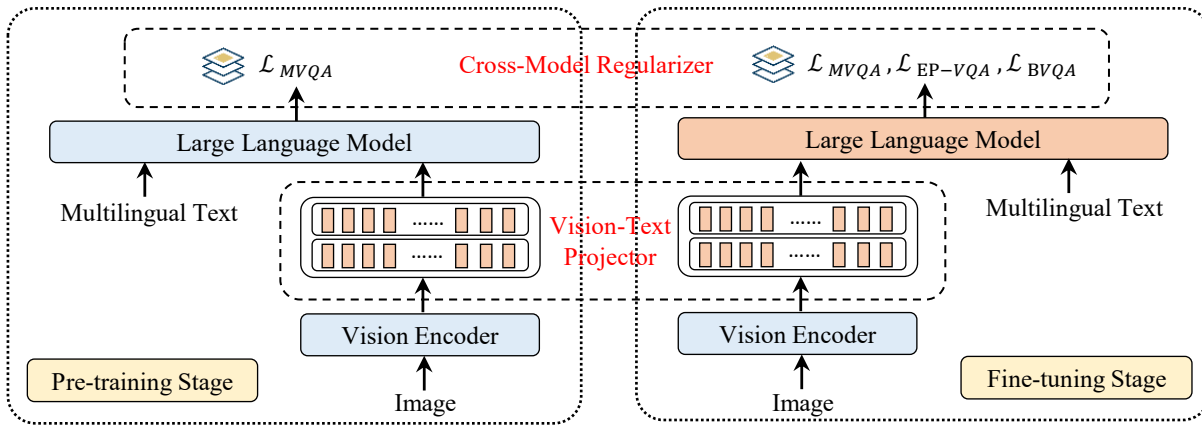


Figure 2: The two-stage training and model architecture of the LRM-LLaVA with the visual encoder, large language model, vision-text representation projector and cross-modal regularizer.

Two-Stage Training

To effectively train LRM-LLaVA and achieve robust cross-modal alignment, we employ a two-stage training strategy, inspired by recent advancements (Liu et al. 2024b).

For pre-training, the emphasis is on aligning the visual features and the multilingual features to establish a shared semantic space. We freeze the parameters of both the visual encoder and the large language model, and use a multilingual task instruction to train the parameters of the projector. The training data consists of images, multilingual questions, and image descriptions. This approach can achieve low-cost alignment of visual and multilingual features. Specifically, for an image X_v and a text sequence X_q of length L , we compute the probability of the target answers X_a^p by:

$$p(X_a^p | X_v, X_q) = \prod_{i=1}^L p_{\theta 1}(x_i | X_v, X_{q, < i}, X_{a, < i}^p) \quad (1)$$

where $\theta 1$ are the trainable parameters in the pre-training stage. $X_{q, < i}$, $X_{a, < i}^p$ are the input sequence and answer tokens before the current prediction token x_i .

For instruction fine-tuning, the emphasis is on improving the instruction following capabilities. We only freeze the parameters of the visual encoder and train the large language model and the projector with three multilingual task instructions based on various types of training data to improve the instruction following ability. This diverse training data exposes the model to a wide range of instructions and prompts, enhancing its ability to generalize to unseen scenarios. Specifically, for an image X_v and a text sequence X_q of length L , we compute the probability of the target answers X_a^f by:

$$p(X_a^f | X_v, X_q) = \prod_{i=1}^L p_{\theta 2}(x_i | X_v, X_{q, < i}, X_{a, < i}^f) \quad (2)$$

where $\theta 2$ are the trainable parameters in the fine-tuning stage. $X_{q, < i}$ and $X_{a, < i}^f$ are the input sequence and answer tokens before the current prediction token x_i .

Vision-Text Representation Projector

The visual-text representation projector plays a critical role by bridging the modality gap between visual and multilingual information. This projector, implemented as a two-layer multi-layer perceptron (MLP), converts the visual features extracted by the visual encoder into semantic features of the multilingual language model. Its concise architecture allows low-cost training and provides an efficient and effective mechanism for aligning the visual features and the multilingual features. Specifically, given an image X_v , the projector first extracts visual features using the visual encoder, denoted as Z_v :

$$Z_v = g_{\phi}(X_v) \quad (3)$$

where g_{ϕ} represents the visual encoder with parameters ϕ . These visual features are then passed through the two-layer MLP projector, parameterized by weight matrices W_1 and W_2 and bias vectors b_1 and b_2 , to produce the aligned multilingual features M_v :

$$H_v = \sigma(W_1 \cdot Z_v + b_1) \quad (4)$$

$$M_v = W_2 \cdot H_v + b_2 \quad (5)$$

where σ represents a non-linear activation function. During the two-stage training process, the projector's parameters (W_1 , W_2 , b_1 , b_2) are continuously updated to optimize the alignment between visual and multilingual features.

Cross-Modal Regularizer

The core of successful LRM-LLaVA training lies in effectively regularizing the interaction between visual encoder and multilingual language model. The regularizer is crucial to minimizing the modality gap between visual and multilingual textual information. In LRM-LLaVA, we achieve this through carefully designed three task instructions including monolingual vision question answering (MVQA), non-English vision question answering with English as the pivot (EP-VQA), and bilingual vision question answering

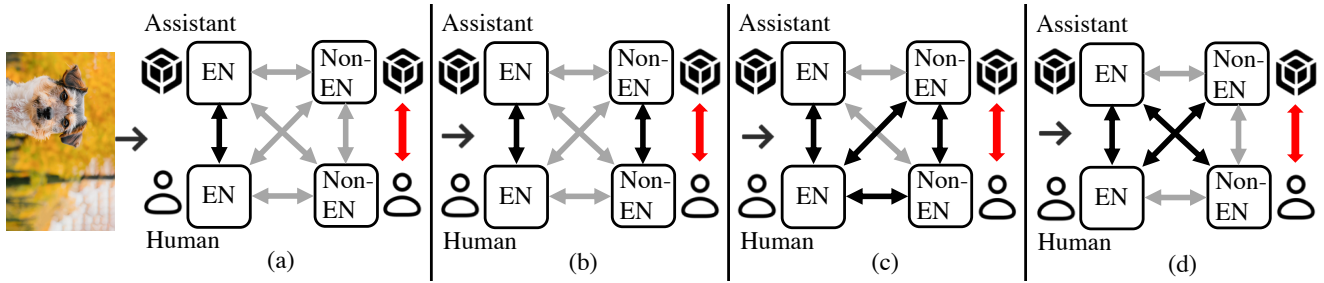


Figure 3: Illustration of how the multilingual task instructions enhance the alignment of image, non-English human question and assistant answer. (a) (b), (c), and (d) respectively represent the alignment for ordinary LLMs, MVQA instruction, EP-VQA instruction, and BVQA instruction. Among them, black arrows indicate strong alignment, while gray arrows indicate weak alignment. It can be observed that these three multilingual training tasks enhance alignment ability through different paths.

(BVQA). Among them, only MVQA is used in the pre-training stage, while all instructions are used in the fine-tuning stage. These instructions, individually and collectively, contribute to enhancing the multilingual capabilities by promoting robust cross-modal alignment.

MVQA: MVQA task instruction focuses on solidifying the ability to understand and respond to visual questions within individual languages. The model is trained on image-question-answer triplets in the same language, learning to map visual information to language-specific question-answer pairs. This process strengthens the connection between visual information and its corresponding linguistic expressions within each language. For an image X_v , the user question X_q^{la} , and assistant answer X_a^{la} in language la , the loss function for MVQA can be expressed as:

$$\mathcal{L}_{MVQA} = - \sum_{i=1}^{|X_a^{la}|} p(x_{a_i} | X_v, X_q^{la}; \theta) \log q(x_{a_i}) \quad (6)$$

where $p(x_{a_i} | X_v, X_q^{la}; \theta)$ represents the one-hot encoding of the ground truth answer token, $q(x_{a_i})$ is the predicted probability distribution over the answer vocabulary for token x_{a_i} , and θ represents the model parameters.

EP-VQA: EP-VQA task instruction leverages better English proficiency to promote cross-lingual understanding. By introducing English as a pivot language, we bridge the gap between non-English languages and English. This instruction is also part of the text input, which first asks the model to translate non-English questions into English questions, and then combines the non-English and translated questions with visual input to generate non-English answers. This multi-step process prompts the model in a step-by-step guided manner, improving translation skills and cross-lingual reasoning because the model must synthesize information from both languages in the context of the visual scene. For an image X_v , the user question X_q^{la} in non-English language la , the EP-VQA instruction X_e , and assistant answer X_a , the loss function for MVQA can be expressed as:

$$\mathcal{L}_{EP-VQA} = - \sum_{i=1}^{|X_a|} p(x_{a_i} | X_v, X_q^{la}, X_e; \theta) \log q(x_{a_i}) \quad (7)$$

where X_a includes non-English answer and English translation of non-English question.

BVQA: BVQA task instruction causes the model to have a bilingual question and answer conversation based on visual information. This instruction requires the model to answer questions in one language with answers in another language, and ensures that one language is English. This forces the model to learn direct mappings between visual representations and linguistic expressions in two languages simultaneously. The objective function of BVQA is similar to MVQA, but operates on cross-lingual question-answer pairs, ensuring that the model learns accurate translation while maintaining visual context. For an image X_v , the user question X_q^{la1} in language $la1$, and assistant answer X_a^{la2} in language $la2$, the loss function for MVQA can be expressed as:

$$\mathcal{L}_{BVQA} = - \sum_{i=1}^{|X_a^{la2}|} p(x_{a_i} | X_v, X_q^{la1}; \theta) \log q(x_{a_i}) \quad (8)$$

LVLMs use the ability of the large language model to align images, questions, and answers in various languages. However, since the training dataset is mainly in English, the model forms a strong alignment relationship between images, English questions, and English answers, while it forms a weak alignment relationship for images, questions, and answers in other non-English languages. Figure 3 shows how MVQA, EP-VQA, and BVQA improve the alignment relationship between images and non-English through different strategic paths, either directly or indirectly with English as the pivot. Each task instruction contributes a unique facet to the overall learning process, encouraging the model to develop a rich understanding of cross-lingual relationships, grounded in visual information. This multi task instruction strategy, coupled with LRM-LLaVA’s carefully designed architecture, significantly enhances its multilingual ability.

Experiments

Datasets and Benchmarks

Due to the lack of multilingual LVLMs datasets, the two-stage training data and benchmarks are constructed from

Model	Languages									
	En	Cs	De	Fr	Hr	It	Ro	Ru	Sp	Zh
Multilingual MME (Fu et al. 2023)										
InstructBLIP	1237.5	879.8	1034.0	1107.1	933.0	1095.2	1092.2	977.1	1155.3	975.3
LLaVA-v1.5	1529.4	1074.7	1211.2	1235.0	945.1	1072.5	1210.4	902.9	999.1	1230.1
LVIS-4V	1574.9	1157.8	1263.2	1336.5	805.8	1238.2	1288.6	826.1	1145.5	1209.3
ShareGPT4V	1599.9	1272.9	1401.0	1365.6	1035.6	1240.6	1296.4	1246.0	1178.7	1329.0
Qwen-VL-Chat	1549.0	784.1	1107.0	1427.7	1037.3	1317.7	1235.2	1239.9	1418.5	1209.9
LRM-LLaVA	1534.3	1338.9	1386.7	1437.4	1422.3	1406.3	1441.5	1305.6	1454.2	1358.3
Multilingual POPE (Li et al. 2023c)										
InstructBLIP	84.8	67.2	81.8	81.5	40.6	81.1	78.1	79.4	81.2	84.3
LLaVA-v1.5	86.0	76.1	83.9	77.9	44.8	76.2	77.5	83.6	75.1	83.7
LVIS-4V	85.6	74.2	84.4	80.6	40.6	71.2	78.0	83.6	72.9	84.3
ShareGPT4V	86.4	64.6	81.4	77.4	43.1	69.2	79.1	84.4	71.9	85.0
LRM-LLaVA	86.6	86.8	87.0	86.9	86.0	87.4	83.3	87.4	84.6	85.1
Multilingual SEED-Bench (Image) (Li et al. 2023a)										
LLaVA-v1.5	68.2	63.6	66.8	67.1	62.5	66.6	64.1	66.7	67.0	66.6
LVIS-4V	69.0	64.3	67.4	67.5	63.0	67.2	64.2	67.4	67.9	67.0
ShareGPT4V	70.8	66.1	69.6	68.8	65.0	68.4	66.4	69.2	70.3	68.1
LRM-LLaVA	70.0	68.7	69.1	69.1	68.6	69.1	68.9	69.4	69.1	68.5
Multilingual MMBench (Liu et al. 2023)										
LLaVA-v1.5	69.5	59.8	62.5	62.1	58.2	59.9	62.4	59.6	61.9	65.3
LVIS-Instruct4V	68.4	60.3	62.8	64.8	57.5	62.8	62.1	59.9	62.4	64.7
ShareGPT4V	69.8	61.0	64.6	63.3	59.1	63.1	62.7	60.4	62.6	65.1
LRM-LLaVA	70.1	66.1	66.3	66.1	65.1	65.8	65.5	66.1	66.4	67.2

Table 1: Evaluation results on our proposed 4 multilingual benchmarks among similar parameters LVLMS. Model names are InstructBLIP (Dai et al. 2023), LLaVA-1.5 (Liu et al. 2024a), LVIS-Instruct4V (Wang et al. 2023), ShareGPT4V (Chen et al. 2023), and Qwen-VL (Bai et al. 2023b). Bold numbers are the best results.

Benchmarks	MME	MMB	SEED	POPE
Avg BLEU	57.34	54.23	58.55	56.19

Table 2: Average BLEU scores of Google Translate and GPT-4o on non-English languages for different benchmarks.

open-source English datasets. For training data, we introduce more than 120M English pre-trained image-text alignment dataset ShareGPT4V-PT (Chen et al. 2023) in pre-training and more than 120M English instruction fine-tuning datasets including LLaVA-1.5 mixtures (Liu et al. 2024a), ChartQA (Masry et al. 2022), GeoQA (Chen et al. 2021), DocVQA (Mathew, Karatzas, and Jawahar 2021), DVQA (Kafle et al. 2018), and AI2D (Kembhavi et al. 2016) in fine-tuning to ensure diversity and comprehensiveness. Furthermore, based on these English training data, we construct the final multilingual training data containing three task instructions by translating from English to non-English languages. For benchmarks, we select four English benchmarks including MME (Fu et al. 2023), MMBench (Liu et al. 2023), POPE (Li et al. 2023c), and SEED-Bench (Li et al. 2023a). Their questions and answers about the images are relatively short, which effectively reduces the information errors or distortions caused by translation.

We select Google Translate, which has the best performance in BLEU scores (Goyal et al. 2022) based on the analysis in Zhu et al. (2023c). In particular, we train di-

rectly on English data for tasks such as mathematical computations due to the unreliability of translating these tasks. For the training languages, we consider the diversity of language families and language sizes under limited resources, and finally select 10 languages, including English (En), Chinese (Zh), Italian (It), French (Fr), German (De), Spanish (Sp), Russian (Ru), Romanian (Ro), Czech (Cs), and Croatian (Hr). We sample data from these languages and integrate them into the final training data. The final data contains more than 240M pre-training data and more than 240M instruction fine-tuning data, of which English accounts for 50% and the 9 non-English languages each account for 5.6%.

Implementation Details

LRM-LLaVA follows a two-stage training process on the 13B-parameter large language model Vicuna-13B (Chiang et al. 2023) and the 0.6B-parameter visual encoder ViT-L/14 (Radford et al. 2021), and it takes 144 hours to complete all the training on 8*A800 GPUs. We pre-train LRM-LLaVA on the proposed 240M multilingual pre-training image-text pairs for 1 epoch with a learning rate of 1e-3 and a batch size of 24 per GPU. Also, we fine-tune LRM-LLaVA on the proposed 240M multilingual instruction fine-tuning image-text pairs for 1 epoch with a learning rate of 2e-5 and a batch size of 32 per GPU. In both stages of training, we use a max context length of 2048 and adopt the same response formatting prompts as LLaVA-1.5.


<p>User: Popište prosím obsah tohoto obrázku v českém jazyce. (English translation: Please describe the content of this image in Czech.)</p> 	<p>LRM-LLaVA (English translation of Czech answer) : There are two elephants in the picture. The first elephant, which occupies the upper right, is only partially visible, mainly its rear, and appears to be standing off-camera. The second elephant, which is the main focus of the image, is positioned in the middle and appears to be standing up or moving in shallow water.</p>	<p>LLaVA-1.5 (English translation of Czech answer) : In the picture there is a large elephant that is in the mud-eating water. The elephant has its trunk extended above the water and its head under water. There are also several eggs around the elephant that are scattered in the mud eater water. The image is created digitally, which is evident in its quality and detail.</p>
--	---	--

Figure 4: An example of LRM-LLaVA and LLaVA-1.5 using Czech questions and answers on the same image. Due to readability and limited space, we only show the English translations of the Czech answers. Compared with LLaVA-1.5, LRM-LLaVA’s description of the image is more accurate, including the number, location and behavior of the elephants.

Method	Languages	
	En	Avg
w/o All Task Instructions	70.2	61.6
w/o MVQA Instruction (Pre-training)	69.8	65.0
w/o MVQA Instruction (Fine-tuning)	69.6	63.8
w/o EP-VQA Instruction	70.0	63.9
w/o BVQA Instruction	69.6	65.2
LRM-LLaVA	70.1	66.1

Table 3: Analysis results of our proposed method on multilingual MMBench. Avg represents the average score of 9 non-English languages. Bold numbers are the best results.

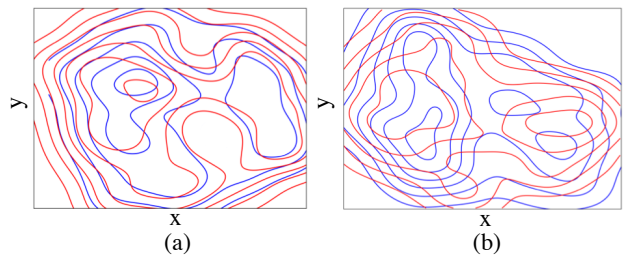


Figure 5: The red line represents the distribution of visual features through projector, while the blue line represents the distribution of multilingual features. (a) corresponds to LRM-LLaVA, (b) corresponds to LLaVA.

Main Results

We compare the performance of multiple LVLMs with 7B parameters and 13B parameters on our proposed four multilingual benchmarks and present the results in Table 1. Compared with previous LVLMs, LRM-LLaVA outperforms competitors especially in non-English languages.

Notably, previous LVLMs are far less capable of understanding non-English, especially in low-resource languages such as Croatian and Czech. LRM-LLaVA uses the same large language model as some of the LVLMs such as LLaVA-1.5 (Liu et al. 2024a) and LVIS-Instruct4V (Wang et al. 2023). However, with our proposed multilingual training framework, the performance of LRM-LLaVA in non-English languages significantly improves and is close to that in English. In particular, although the training data of LRM-LLaVA includes non-English languages, LRM-LLaVA performs well on English and even achieves the best results on multilingual MMBench and multilingual POPE. For qualitative analysis, Figure 4 shows an example of LRM-LLaVA and LLaVA-1.5 using Czech questions and answers on the same image. It can be seen that LRM-LLaVA’s description of the image is more detailed and accurate. The comprehensive results indicate that our proposed method can significantly improve the multilingual ability of LVLMs.

Analysis and Discussion

Analysis on Benchmark Translation Results

To verify the reliability of the translated multilingual benchmarks, we use Google Translate and GPT-4o (Achiam et al. 2023) to translate the English benchmarks including MMBench, MME, POPE, and SEED-Bench into 9 non-English languages. Based on the two translation results, we calculate the average BLEU scores of the 9 non-English languages to quantify the similarity of the two translation results. Table 2 shows the specific results. The experimental results show that Google Translate and GPT-4o demonstrate good synergy in the translation results of the four benchmarks, and their BLEU scores have good performance. This is because the questions and answers of the multiple-choice questions and true-or-false questions in the four benchmarks are relatively short, making the translation results more reliable.

Ablation Study on Task Instructions

To better understand the contribution of the three multilingual task instructions, we conduct an ablation study on LRM-LLaVA as shown in Table 3.

Firstly, removing all task instructions reduces the average score of non-English languages by 4.5, but the English score remains the same as LRM-LLaVA. This shows

Language Model	Method	Languages	
		En	Avg
Vicuna-13B	w/o All Instructions	70.2	61.6
	w/ All Instructions	70.1	66.1
Vicuna-7B	w/o All Instructions	69.4	60.9
	w/ All Instructions	69.2	64.7
LLaMA2-13B	w/o All Instructions	65.2	58.7
	w/ All Instructions	65.5	62.1
LLaMA3-8B	w/o All Instructions	69.7	62.3
	w/ All Instructions	69.4	66.0

Table 4: Multilingual performance of different LLMs on multilingual MMBench with and without our approach. Avg represents the average score of 9 non-English languages.

our proposed method can significantly improve multilingual ability without reducing English ability. Secondly, removing MVQA in the pre-training stage, and removing BVQA in the fine-tuning stage will reduce 1.1 and 0.9 points respectively, while removing MVQA and EP-VQA in the fine-tuning stage will reduce 2.3 and 2.2 points respectively. This is because MVQA and EP-VQA in the fine-tuning stage change the parameters of the large language model and their alignment paths are more brief. Nevertheless, MVQA in the pre-training stage and BVQA in the fine-tuning stage are also necessary, and play roles in different alignment paths.

Analysis on Vision-Text Modality Alignment

As shown in Figure 5, we visualize the distribution of the multilingual features and visual features of LRM-LLaVA and LLaVA to examine the aligning effectiveness of our method. Specifically, we average the sequential representations of the image features through the projector and the multilingual features in 5,000 data points, and then apply the T-SNE dimensionality reduction algorithm to reduce the hidden dimensions to 2 dimensions. We plot the kernel density estimation distribution based on the reduced 2-dimensional representations of the two features. Our method significantly enhances the modality alignment of multilingual and visual features in the semantic space.

Analysis on the Large Language Models

Table 4 shows the effect of our proposed multilingual training framework on different large language models. Firstly, to observe the impact of parameter scale, we use the same series of large language models, Vicuna-7B and Vicuna-13B, to train LVLMs with English training data and multilingual training data in the same process. With our method, the 13B-parameter model achieves higher multilingual performance than the 7B model. Moreover, compared with the model trained with English data, the improvement in multilingual performance of the 13B-parameter model is also higher than that of the 7B-parameter model. This shows that larger-scale models have better multilingual understanding, can more easily achieve alignment between multilingual question answering and visual modalities, and achieve better returns.

Secondly, to verify our method does not depend on a specific large language model, we select LLaMA2-13B (Tou-

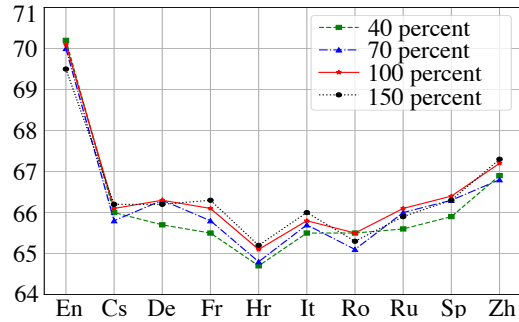


Figure 6: We mix 40%, 70%, 100%, and 150% of non-English data into the English data respectively for training, and score them in 10 languages on multilingual MMBench.

vron et al. 2023) and LLaMA3-8B (Dubey et al. 2024) as language models to train multilingual LVLMs. Experimental results show that our method achieves improvements on different large language models and exhibits the same trend.

Analysis on the Size of Synthesized Data

To explore the reasonable size of synthetic non-English data, we add different proportions of non-English data to the English fine-tuning data and evaluate their scores in 10 languages on multilingual MMBench. As shown in Figure 6, we mix 40%, 70%, 100%, and 150% of non-English data respectively. Experimental results show that mixing 100% non-English data significantly improves the multilingual performance compared with 40% and 70% mixtures. However, when the proportion increases to 150%, the non-English performance does not improve significantly, but the English performance decrease significantly. For the final training data, we choose to use a 100% non-English data integration ratio in both pre-training and instruction fine-tuning stages to improve multilingual capabilities as much as possible without affecting English capabilities.

Conclusion

This paper proposes a multilingual LVLMs training framework for low-resource languages. Based on this, we train LRM-LLaVA on 10 languages, which mainly solves the modality gap between visual features and multilingual features. The multilingual training framework consists of four components, including a visual encoder, a backbone multilingual large language model, a vision-text representation projector, and a cross-modal regularizer. LRM-LLaVA uses a two-stage training approach for model training and vision-text alignment. Due to the lack of non-English data, we translate English training data into non-English data and construct three monolingual or bilingual instructions. To evaluate the multilingual capabilities of LVLMs, we translate four English benchmarks to obtain multilingual benchmarks and verify their reliability. Experimental results show that LRM-LLaVA achieves competitive multilingual performance compared to other multilingual LVLMs of similar parameters.

Acknowledgments

The present research was supported by the National Natural Science Foundation of China Youth Foud (Grant No. 62306210) and Natural Science Foundation of Xinjiang Uygur Autonomous Region (Grant No. 2022D01D43). We would like to thank the anonymous reviewers for their insightful comments.

References

- Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; Hui, B.; Ji, L.; Li, M.; Lin, J.; Lin, R.; Liu, D.; Liu, G.; Lu, C.; Lu, K.; Ma, J.; Men, R.; Ren, X.; Ren, X.; Tan, C.; Tan, S.; Tu, J.; Wang, P.; Wang, S.; Wang, W.; Wu, S.; Xu, B.; Xu, J.; Yang, A.; Yang, H.; Yang, J.; Yang, S.; Yao, Y.; Yu, B.; Yuan, H.; Yuan, Z.; Zhang, J.; Zhang, X.; Zhang, Y.; Zhang, Z.; Zhou, C.; Zhou, J.; Zhou, X.; and Zhu, T. 2023a. Qwen Technical Report. *arXiv preprint arXiv:2309.16609*.
- Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023b. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.
- Bellagente, M.; Brack, M.; Teufel, H.; Friedrich, F.; Deiseroth, B.; Eichenberg, C.; Dai, A. M.; Baldock, R.; Nanda, S.; Oostermeijer, K.; et al. 2024. Multifusion: Fusing pre-trained models for multi-lingual, multi-modal image generation. *Advances in Neural Information Processing Systems*, 36.
- Cadene, R.; Ben-Younes, H.; Cord, M.; and Thome, N. 2019. Murel: Multimodal relational reasoning for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1989–1998.
- Chen, G. H.; Chen, S.; Zhang, R.; Chen, J.; Wu, X.; Zhang, Z.; Chen, Z.; Li, J.; Wan, X.; and Wang, B. 2024. Allava: Harnessing gpt4v-synthesized data for a lite vision-language model. *arXiv preprint arXiv:2402.11684*.
- Chen, J.; Tang, J.; Qin, J.; Liang, X.; Liu, L.; Xing, E. P.; and Lin, L. 2021. GeoQA: A geometric question answering benchmark towards multimodal numerical reasoning. *arXiv preprint arXiv:2105.14517*.
- Chen, L.; Li, J.; Dong, X.; Zhang, P.; He, C.; Wang, J.; Zhao, F.; and Lin, D. 2023. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*.
- Chen, X.; Wang, X.; Changpinyo, S.; Piergiovanni, A.; Padlewski, P.; Salz, D.; Goodman, S.; Grycner, A.; Mustafa, B.; Beyer, L.; et al. 2022. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*.
- Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. See <https://vicuna.lmsys.org> (accessed 14 April 2023), 2(3): 6.
- Dai, W.; Li, J.; Li, D.; Tiong, A. M. H.; Zhao, J.; Wang, W.; Li, B.; Fung, P.; and Hoi, S. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. *arXiv:2305.06500*.
- Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*.
- Fu, C.; Chen, P.; Shen, Y.; Qin, Y.; Zhang, M.; Lin, X.; Yang, J.; Zheng, X.; Li, K.; Sun, X.; et al. 2023. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. *arXiv preprint arXiv:2306.13394*.
- Goyal, N.; Gao, C.; Chaudhary, V.; Chen, P.-J.; Wenzek, G.; Ju, D.; Krishnan, S.; Ranzato, M.; Guzmán, F.; and Fan, A. 2022. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10: 522–538.
- Kafle, K.; Price, B.; Cohen, S.; and Kanan, C. 2018. Dvqa: Understanding data visualizations via question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5648–5656.
- Kembhavi, A.; Salvato, M.; Kolve, E.; Seo, M.; Hajishirzi, H.; and Farhadi, A. 2016. A diagram is worth a dozen images. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, 235–251. Springer.
- Li, B.; Wang, R.; Wang, G.; Ge, Y.; Ge, Y.; and Shan, Y. 2023a. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.
- Li, J.; Li, D.; Savarese, S.; and Hoi, S. C. H. 2023b. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In Krause, A.; Brunskill, E.; Cho, K.; Engelhardt, B.; Sabato, S.; and Scarlett, J., eds., *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, 19730–19742. PMLR.
- Li, Y.; Du, Y.; Zhou, K.; Wang, J.; Zhao, W. X.; and Wen, J.-R. 2023c. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.
- Li, Z.; Fan, Z.; Chen, J.; Zhang, Q.; Huang, X.; and Wei, Z. 2023d. Unifying Cross-Lingual and Cross-Modal Modeling Towards Weakly Supervised Multilingual Vision-Language Pre-training. In Rogers, A.; Boyd-Graber, J. L.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9–14, 2023*, 5939–5958. Association for Computational Linguistics.
- Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26296–26306.

- Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024b. Visual instruction tuning. *Advances in neural information processing systems*, 36.
- Liu, Y.; Duan, H.; Zhang, Y.; Li, B.; Zhang, S.; Zhao, W.; Yuan, Y.; Wang, J.; He, C.; Liu, Z.; et al. 2023. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.
- Liu, Y.; Wang, Y.; Sun, L.; and Yu, P. S. 2024c. Rec-GPT4V: Multimodal Recommendation with Large Vision-Language Models. *CoRR*, abs/2402.08670.
- Masry, A.; Long, D. X.; Tan, J. Q.; Joty, S.; and Hoque, E. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*.
- Mathew, M.; Karatzas, D.; and Jawahar, C. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2200–2209.
- Nguyen, T.; Wallingford, M.; Santy, S.; Ma, W.; Oh, S.; Schmidt, L.; Koh, P. W.; and Krishna, R. 2024. Multilingual Diversity Improves Vision-Language Representations. *CoRR*, abs/2405.16915.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.
- Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Wang, J.; Meng, L.; Weng, Z.; He, B.; Wu, Z.; and Jiang, Y.-G. 2023. To see is to believe: Prompting gpt-4v for better visual instruction tuning. *arXiv preprint arXiv:2311.07574*.
- Wang, L.; He, J.; Li, S.; Liu, N.; and Lim, E. 2024. Mitigating Fine-Grained Hallucination by Fine-Tuning Large Vision-Language Models with Caption Rewrites. In Rudinac, S.; Hanjalic, A.; Liem, C. C. S.; Worring, M.; Jonsson, B.; Liu, B.; and Yamakata, Y., eds., *MultiMedia Modeling - 30th International Conference, MMM 2024, Amsterdam, The Netherlands, January 29 - February 2, 2024, Proceedings, Part IV*, volume 14557 of *Lecture Notes in Computer Science*, 32–45. Springer.
- Ying, K.; Meng, F.; Wang, J.; Li, Z.; Lin, H.; Yang, Y.; Zhang, H.; Zhang, W.; Lin, Y.; Liu, S.; Lei, J.; Lu, Q.; Chen, R.; Xu, P.; Zhang, R.; Zhang, H.; Gao, P.; Wang, Y.; Qiao, Y.; Luo, P.; Zhang, K.; and Shao, W. 2024. MMT-Bench: A Comprehensive Multimodal Benchmark for Evaluating Large Vision-Language Models Towards Multitask AGI. *CoRR*, abs/2404.16006.
- You, H.; Sun, R.; Wang, Z.; Chen, L.; Wang, G.; Ayyubi, H. A.; Chang, K.; and Chang, S. 2023. IdealGPT: Iteratively Decomposing Vision and Language Reasoning via Large Language Models. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, 11289–11303. Association for Computational Linguistics.
- Yu, T.; Yao, Y.; Zhang, H.; He, T.; Han, Y.; Cui, G.; Hu, J.; Liu, Z.; Zheng, H.-T.; Sun, M.; et al. 2024. RLhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13807–13816.
- Zhai, X.; Mustafa, B.; Kolesnikov, A.; and Beyer, L. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11975–11986.
- Zhang, Y.; Zhang, R.; Gu, J.; Zhou, Y.; Lipka, N.; Yang, D.; and Sun, T. 2023. LLaVAR: Enhanced Visual Instruction Tuning for Text-Rich Image Understanding. *CoRR*, abs/2306.17107.
- Zhou, C.; Zhang, M.; Chen, P.; Fu, C.; Shen, Y.; Zheng, X.; Sun, X.; and Ji, R. 2024. VEGA: Learning Interleaved Image-Text Comprehension in Vision-Language Large Models. *CoRR*, abs/2406.10228.
- Zhou, M.; Zhou, L.; Wang, S.; Cheng, Y.; Li, L.; Yu, Z.; and Liu, J. 2021. Uc2: Universal cross-lingual cross-modal vision-and-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4155–4165.
- Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023a. Minigt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.
- Zhu, H.; Huang, J.; Rudinac, S.; and Kanoulas, E. 2024. Enhancing Interactive Image Retrieval With Query Rewriting Using Large Language Models and Vision Language Models. In Gurrin, C.; Kongkachandra, R.; Schoeffmann, K.; Dang-Nguyen, D.; Rossetto, L.; Satoh, S.; and Zhou, L., eds., *Proceedings of the 2024 International Conference on Multimedia Retrieval, ICMR 2024, Phuket, Thailand, June 10-14, 2024*, 978–987. ACM.
- Zhu, S.; Li, S.; Lei, Y.; and Xiong, D. 2023b. PEIT: Bridging the Modality Gap with Pre-trained Models for End-to-End Image Translation. In Rogers, A.; Boyd-Graber, J. L.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, 13433–13447. Association for Computational Linguistics.
- Zhu, W.; Liu, H.; Dong, Q.; Xu, J.; Huang, S.; Kong, L.; Chen, J.; and Li, L. 2023c. Multilingual machine translation with large language models: Empirical results and analysis. *arXiv preprint arXiv:2304.04675*.