# PALO: A Polyglot Large Multimodal Model for 5B People

**Anonymous ACL submission**

## Abstract

In pursuit of more inclusive Vision-Language Models (VLMs), this study introduces a Large Multilingual Multimodal Model called PALO. PALO offers visual reasoning capabilities in 10 major languages, including English, Chinese, Hindi, Spanish, French, Arabic, Bengali, Russian, Urdu, and Japanese, that span a total of ~5B people (65% of the world population). Our approach involves a semi-automated translation approach to adapt the multimodal instruction dataset from English to the target languages using a fine-tuned Large Language Model, thereby ensuring high linguistic fidelity while allowing scalability due to minimal manual effort. The incorporation of diverse instruction sets helps us boost overall performance across multiple languages especially those that are underrepresented like Hindi, Arabic, Bengali, and Urdu. The resulting models are trained across three scales (1.7B, 7B and 13B parameters) to show the generalization and scalability where we observe substantial improvements compared to strong baselines. We also propose the first multilingual multimodal benchmark for the forthcoming approaches to evaluate their vision-language reasoning capabilities across languages. Our codes, models and datasets will be publicly released.

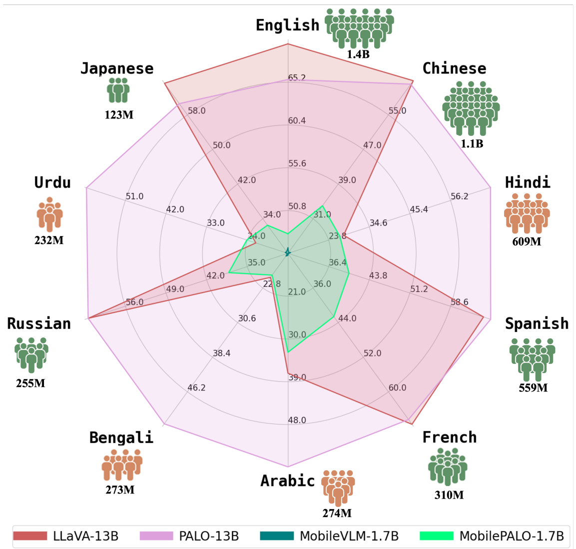


Figure 1: **PALO vs. English-VLMs.** The plot compares PALO with corresponding Vision-Language Models (VLMs) across 10 different languages. These languages include English, Chinese, Hindi, Spanish, French, Arabic, Bengali, Russian, Urdu, and Japanese, collectively covering approximately 5B people and 65% of the global population. English-trained VLMs, such as LLaVA and MobileVLM, exhibit poor performance on low-resource languages including Hindi, Arabic, Bengali, and Urdu, due to the under-representation of these languages during their training phases. PALO, in contrast, is a unified model that can hold conversations simultaneously in all the ten languages, demonstrating consistent performance across the board.

## 1 Introduction

Propelled by advancements in generative AI, Large Multimodal Models (LMMs) (Liu et al., 2023b; Zhu et al., 2023; Dai et al., 2023) have emerged as a pivotal advancement in the field, seamlessly bridging the gap between vision and language tasks. While initial efforts such as LLaVA (Liu et al., 2023b) and miniGPT4 (Zhu et al., 2023) have demonstrated intriguing performance in synthesizing effective textual responses based on visual inputs, they have predominantly focused on English, leaving a significant gap in multimodal understanding for non-English languages. As a result, the existing LMMs generally overlook the linguistic diversity of the global population, particularly languages spoken by large groups, such as Chinese, Hindi, Spanish, French, Arabic, Bengali, Russian, Urdu, and Japanese, which collectively account for billions of native speakers. Our work addresses this disparity by developing the first fully open-source multilingual LMM called PALO, which encompasses ten major languages covering 65% of the global population, with a special focus on languages underrepresented in the current multimodal models.

The challenge lies in the scarcity of high-quality multilingual multimodal data compared to English. Addressing the challenge of limited high-quality data, especially for under-represented languages such as Hindi, Arabic, Bengali, and Urdu, our approach involves careful analysis and subsequent refinement of translations produced by a state-of-the-art Large Language Model (LLM) (Brown et al., 2020) for each target language. By identifying and correcting translation inaccuracies through human intervention, we generate a high-quality multilingual dataset. This curated dataset then serves as the foundation for refining the target language annotations, ensuring a more accurate and nuanced representation of the target language in training.

Leveraging our high-quality multilingual vision-language instruction dataset and the recent advances in large multimodal modeling, we develop PALO as a *unified* model that can simultaneously answer questions in ten different languages. Our training pipeline offers substantial gains in low-resource languages (underrepresented in the LLM training datasets) while maintaining (or further improving) performance on high-resource languages. The contributions of this work are as follows,

- We develop PALO: the first multilingual Large Multimodal Model (LMM) covering ten major languages, facilitating vision-language reasoning through a generic model capable of generating responses in any of the ten languages.
- We assemble an extensive multilingual (10 languages) instruction-tuning dataset, through a critical analysis and subsequent refinement of a state-of-the-art Large Language Model's target language translations. This dataset is pivotal in improving proficiency in processing and generating content that is linguistically precise across multiple languages.
- We enhance the multilingual performance of state-of-the-art LMMs (Liu et al., 2023b; Chu et al., 2023) across three distinct scales i.e., 1.7B, 7B, and 13B parameters to demonstrate the scalability of our training pipeline. The resulting polyglot LMMs demonstrate performance gains on diverse language tasks with substantial improvements in understanding and generating content for low-resource languages, e.g., Hindi, Arabic, Bengali, and Urdu, without compromising its high-performance on high-resource languages e.g., English, Chinese, French, and Spanish.

## 2 Related Works

The introduction of Large Language Models (LLMs) has significantly advanced the field of natural language processing. However, the development of multilingual LLMs has faced considerable challenges, primarily due to the skewed distribution of language data (Costa-jussà et al., 2022). English and European languages dominate existing datasets, leaving widely spoken languages such as Mandarin Chinese and Hindi underrepresented (Eberhard et al., 2015). Moreover, integrating multiple languages into LLMs often leads to a decline in English language performance (Scao et al., 2022), highlighting a major challenge in maintaining cross-lingual performance.

Recent efforts have aimed to address these challenges by developing multilingual LLMs with enhanced capabilities (Almazrouei et al., 2023; Touvron et al., 2023; Le Scao et al.; Wei et al., 2023). BLOOM (Le Scao et al.), trained on the ROOTS corpus (Laurençon et al., 2022) that comprises sources in 46 languages, marks a substantial step forward in making LLMs accessible across a wide range of languages, including those with fewer resources. PaLM (Chowdhery et al., 2023) showcases the advantages of scaling, achieving improved results in both monolingual and multilingual tasks through sophisticated training techniques and a novel pathways architecture.

Advancements in Large Multimodal Models (LMMs) have evolved from basic image-level interactions (Liu et al., 2023b; Chu et al., 2023) to offering flexibility by focusing on region-specific analysis (Rasheed et al., 2023) and spatio-temporal conversations (Maaz et al., 2023; Lin et al., 2023), highlighting the significant progress in this domain. However, the exploration of multilingual capabilities has been limited. Qwen (Bai et al., 2023) and mPLUG-Owl (Ye et al., 2023) extend LMM functionalities to process visual inputs in both English and Chinese, showcasing its adaptability in processing bilingual visual information. Ziya-Visual (Lu et al., 2023) demonstrates the translation of English image-text datasets into Chinese, employing in-context learning for instruction-response generation. However, these LMMs remain limited to two languages.

We introduce PALO, the first fully open-source LMM, offering visual reasoning capabilities across ten major languages, addressing the gap in multilingual LMMs. In contrast to GPT-4 (Achiam

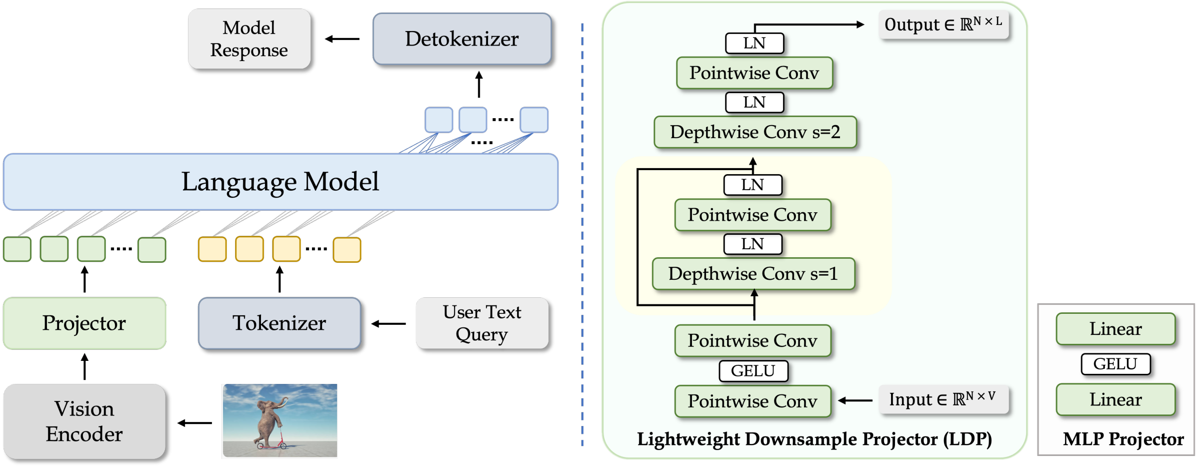


Figure 2: **Architecture overview of PALO.** (*left*) The model consists of a vision encoder that encodes the image, followed by a projector that projects the vision features into the input embedding space of the language model. The user's text query is tokenized, and the tokens are concatenated with the vision tokens before being input into the causal language model to generate the response. For the PALO 7B and 13B variants, Vicuna is used as the Large Language Model while MobileLLaMA (Chu et al., 2023) is used as the Small Language Model in our MobilePALO-1.7B variant. CLIP ViT-L/336px is used as the vision encoder in all variants. (*right*) Projectors used in different variants of PALO are shown. For the PALO 7B and 13B, following (Liu et al., 2023b), we use a two-layer MLP projector with GELU activation. For our mobile version of PALO (MobilePALO-1.7B), we use a Lightweight Downsample Projector (LDP) from (Chu et al., 2023). It utilizes depth-wise separable convolutions to downsample the image tokens, making it faster than a standard MLP projector.

et al., 2023) which is closed-source and only accessible via APIs, ours is the largest effort in the open-source domain to extend LMM capabilities to multiple languages.

## 3 PALO: A Polyglot LMM

Towards more globally accessible Vision-Language Models (VLMs), our model PALO (**P**olyglot **L**arge **Mu**ltimodal **Mo**del) is designed to comprehend and generate content in ten major languages, serving an audience that spans nearly two-thirds of the global population. The architecture of PALO is derived from LLaVA (Large Language and Vision Assistant) (Liu et al., 2023b,a) for our larger-scale models (7/13B), and from MobileVLM for our mobile-efficient model (1.7B), ensuring that PALO remains versatile across different computational settings.

The architecture seamlessly integrates a vision encoder with a language model (see Figure 2). Given an input image and user text query, the model generates an accurate natural language response.

PALO uses CLIP ViT-L/14 (Radford et al., 2021) as the vision encoder followed by a projector to transform vision tokens to the input embedding space of the language model. Following LLaVA (Liu et al., 2023b), we use a two-layer MLP with GELU activation as the projector for our 7/13B models. However, a lightweight downsample projector (LDP) (Chu et al., 2023) is used for MobilePALO-1.7B model. LDP utilizes depth-wise separable convolutions to downsample the vision tokens, largely reducing the input tokens to the language model and hence significantly reducing the training and inference time. Further, convolutions in LDP have fewer parameters as compared to MLP, making our mobile model both parameter and compute-efficient. The projector used in the different PALO versions are shown in Figure 2.

The projected vision tokens are then concatenated with the tokenized user text query and passed to the language model for generating the response. As PALO trains on ten languages using an extensive multi-modal instruction tuning dataset, this not only enables more effective utilization of the tokenizer's capacity but also expands the search space, providing a richer context and more challenging examples for training. the language model. This approach significantly enhances the ability of the model to understand and generate responses across a diverse set of languages.

We use Vicuna (Zheng et al., 2023) as the large language model (LLM) in our 7/13B models and MobileLLaMA (Chu et al., 2023) as the small language model (SLM) in MobilePALO-1.7B model. Vicuna fine-tunes LLaMA-2 on user-shared conversations collected from ShareGPT, while LLaMA-2 is pre-trained on 2T tokens collected from different public sources (Touvron et al., 2023). On the other hand, MobileLLaMA performs pretraining on 1.3T tokens from RedPajama-v1 (Computer, 2023) followed by fine-tuning on a publicly available version of ShareGPT data (Huggingface).

### 3.1 Dataset

The primary contribution of our work lies in the meticulous preparation of a comprehensive multilingual vision-language instruction-tuning dataset. We begin by selecting a state-of-the-art LMM model (Liu et al., 2023b) for our focus. To tailor the instruction-tuning dataset more effectively for multiple languages in a scalable way, we leverage an LLM model (Brown et al., 2020) to develop a semi-automated translation pipeline. This approach involves translating the English dataset into the target languages, thereby creating a robust multilingual dataset, which significantly broadens the linguistic scope and applicability of the model.

**Translation Process and Challenges**: A naive translation approach from English to the target languages using an LLM model (Brown et al., 2020) effectively conveys the basic meanings but introduces several linguistic challenges specific to each language. Issues such as punctuation, grammatical nuances, translation consistencies, and gender usage errors are observed via a direct LLM-based translation (refer Figure.3). These challenges vary greatly due to the linguistic diversity of the languages involved, from the tonal complexities of Chinese to the script variances in Hindi and the gender-specific intricacies of languages like Spanish, Arabic and Russian. For instance, in the case of Arabic, common punctuation mistakes involve incorrect spacing around commas and periods. Nunnation, vital in Arabic grammar, is sometimes omitted or wrongly applied. Additionally, certain English words remain untranslated in the translated text, and there are instances where verbs are incorrectly converted to nouns alongside incorrect gender alignment in translations that pose significant concerns, given the gender-specific nature of grammar in some target languages.

**Addressing the Challenges**: To improve the quality of the translated dataset, we employ a combination of automated and manual verification steps. In this semi-automated pipeline, a team of native speakers for each language provides detailed review and correction of a small subset from initial translations, addressing language-specific issues, gender accuracy, and overall linguistic integrity.

| | |
|---|---|
| In the image, two zebras are walking together in their zoo. | في الصورة، زبران يسيران معًا في موطنهما بحديقة الحيوان. |
| In the image, two zebras are walking together in their zoo. | في الصورة، يمشي إثنان من الحمر الوحشية معا في حديقتهما الحيوانية. |

Improvements in lexical precision: The vanilla model inaccurately translates zebras as زبران and uses a less common term for zoo. Post-tuning, the model employs the correct term حمار الوحش for zebras and the widely accepted حديقة الحيوان for zoo, enhancing the naturalness of the translation.

| | |
|---|---|
| The clock is situated on an ornate, yellow and tan building. | الساعة يقع على مبنى مزخرف باللون الأصفر والبيج. |
| The clock is situated on an ornate, yellow and tan building. | الساعة موجودة على مبنى مُزخرف باللون الأصفر والبُنيّ. |

Improvements in grammatical concordance: Initially, the model misaligns the verb gender with the feminine noun الساعة (the clock), using يقع instead of the appropriate تقع. The fine-tuned model corrects this by using, موجودة which not only matches the noun's gender but also provides a more colloquial translation.

Figure 3: **Qualitative results showing the impact of fine-tuning**. Comparative visualization of English to Arabic translations before and after fine-tuning the LLM. The figure shows improvements in language-specific issues such as accurate vocabulary usage, gender agreement, and grammatical correctness, highlighting the enhanced performance of the fine-tuned model.

Automated scripts are tailored for each language to correct common punctuation mistakes and optimize the verification process.

**Fine-tuning of the LLM**: Acknowledging the limitations of the LLM for multilingual translations, we leverage manually verified and corrected translations (1K conversations per language) as a high-quality dataset for fine-tuning the LLM. This fine-tuning is focused not only on improving translation accuracy but also on aligning the outputs with the specific attributes of each language, such as tone and orthography. The enhanced and fine-tuned LLM is then employed to translate the extensive VLM instruction tuning dataset (Liu et al., 2023b) comprising approximately 150K instructions (i.e. LLaVA-Instruct-150K from (Liu et al., 2023b)) from English into the respective languages. We use GPT3.5-Turbo as the translation model and finetune it using OpenAI finetuning platform.

**Impact of the Refined Dataset**: This process results in a comprehensive and high-quality multilingual dataset, crucial for the effective fine-tuning of PALO. The improved dataset not only addresses specific aspects of each language but also markedly improves the ability of the model to process and generate contextually relevant and grammatically accurate content in all included languages. For instance, Figure 3 highlights two key improvements in English to Arabic translation, the first example shows enhanced lexical precision, and the second shows improved grammatical concordance. Integrating this dataset into the LMM's training process is the key to expanding its capabilities to include both English and nine other languages effectively.

| Model | Eng. | Chinese | French | Spanish | Russ. | Japan. | Arabic | Hindi | Bengali | Urdu | Avg.H | Avg.L | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| LLaVA-7B | **67.9** | **55.7** | **62.4** | **64.5** | 55.3 | **59.2** | 38.9 | 29.4 | 13.9 | 21.8 | **60.8** | 26.0 | 46.9 |
| PALO-7B | 64.2 | **55.7** | 58.3 | 61.0 | **57.4** | 57.5 | **57.8** | **57.6** | **51.7** | **55.3** | 59.0 | **55.6** | **57.7** |
| | -3.7 | 0.0 | -4.1 | -3.5 | +2.1 | -1.7 | +18.9 | +28.2 | +37.8 | +33.5 | -1.8 | +29.6 | +10.8 |
| LLaVA-13B | **69.5** | **62.9** | **67.5** | 64.6 | 62.3 | **65.3** | 37.2 | 27.8 | 20.4 | 22.1 | **65.4** | 26.9 | 49.9 |
| PALO-13B | 65.5 | 62.1 | 66.4 | **65.9** | **62.4** | 60.6 | **56.9** | **66.8** | **53.5** | **59.6** | 63.8 | **59.2** | **61.9** |
| | -4.0 | -0.8 | -1.1 | +1.3 | +0.1 | -4.7 | +19.7 | +39.0 | +33.1 | +37.5 | -1.5 | +32.3 | +12.0 |
| MobileVLM-1.7B | 46.6 | 23.2 | 28.1 | 29.1 | 28.1 | 26.4 | 12.4 | 13.7 | 15.6 | 15.6 | 30.3 | 14.3 | 23.9 |
| MobilePALO-1.7B | **48.2** | **34.0** | **42.6** | **40.1** | **38.2** | **32.5** | **32.8** | **26.8** | **19.9** | **24.1** | **39.3** | **25.9** | **33.9** |
| | +1.6 | +10.8 | +14.5 | +11.0 | +10.1 | +6.1 | +20.4 | +13.1 | +4.3 | +8.5 | +9.0 | +11.6 | +10.0 |

Table 1: **Standard VLMs vs PALO on multi-lingual multimodal evaluation.** The table shows the comparison of LLaVA and MobileVLM with PALO on ten languages on the specially adapted multilingual version of LLaVA-Bench (In-the-Wild). LLaVA 7/13B and MobileVLM-1.7B are fine-tuned on LLaVA-Instruct-665K, and PALO is fine-tuned on LLaVA-Instruct-665K plus the LLaVA-Instruct-150K translated in all ten languages. All models are pretrained on CC-595K (Liu et al., 2023b) dataset. Avg.H and Avg.L represent the average over high-resource (English, Chinese, French, Spanish, Russian and Japanese) and low-resource (Arabic, Hindi, Bengali and Urdu) languages respectively. Avg. represents the average over all the languages.

# 4 Experiments

## 4.1 Implementation Details

Similar to the LLaVA and MobileVLM baselines, we pretrain our models on a subset of CC3M dataset called CC-595K (Liu et al., 2023b). During pretraining, only the projector is learned and the rest of the model components are kept frozen. We train the model for 1 epoch with an overall batch size of 256 with 32 batch size per GPU on eight A-100 40GB GPUs. The model is optimized using Adam optimizer and cosine LR scheduler with a learning rate of 2e-3. The pertaining takes around 1.5 hours for 1.7B, 5 hours for 7B and almost 9 hours for the 13B model.

We fine-tune our model on a diverse instruction dataset comprising conversations from ten languages. Specifically, 665K instructions from LLaVA-Instruct-665K (Liu et al., 2023a) are used for English, and approximately 150K conversations from LLaVA-Instruct-150K (Liu et al., 2023b) for Chinese, French, Spanish, Russian, Japanese, Arabic, Hindi, Bengali and Urdu, summing up to almost 2.1M instructions in total. During fine-tuning, only the vision encoder is kept froze and the rest of the model is trained. Projector is fully trained while language model is LORA (Hu et al., 2022) fine-tuned with $\alpha = 128$. We train the model for 1 epoch with an overall batch size of 128 with 16 batch size per GPU on eight A-100 GPUs. We use 40GB A-100 GPUs for 1.7/7B variants and 80GB A-100 GPUs for 13B variants. The model is optimized using Adam optimizer and cosine LR scheduler with 2e-5 base learning rate for the projector and 2e-4 for the language model. The finetuning takes around 12 hours for 1.7B, 42 hours for 7B and almost 76 hours for the 13B model.

## 4.2 High-resource vs Low-resource Languages

Our work trains and evaluates on ten languages divided into two groups, high-resource and low-resource languages. English, Chinese, French, Spanish, Russian and Japanese are considered high-resource languages as the language model training data contains a reasonable number of samples from these languages. On the other hand, Arabic, Hindi, Bengali and Urdu are categorized as low-resource languages as they are under-represented in the language model training data.

For example, LLaMA-2 (Touvron et al., 2023) pretraining data contains almost 2 trillion tokens, out of which 89.7% are of English and almost 1.92% is for Chinese, French, Spanish, Russian, Japanese, and 21 more similar languages. While the representation of Arabic, Hindi, Bengali and Urdu is negligible. Similarly, MobileLLaMA (Chu et al., 2023) pretrains on RedPajama-v1 (Computer, 2023) dataset which consist of almost 1.3 trillion tokens, predominantly English tokens.

## 4.3 Results

In evaluating the multilingual capabilities of VLMs, we conduct a comprehensive evaluation across various languages, utilizing a high-quality evaluation set. This set is constructed by translating the LLaVA-Bench (In-the-Wild) (Liu et al., 2023b) into all target languages using GPT-4-Turbo (Achiam et al., 2023), with particular attention to preserving linguistic authenticity and

| Data | English | Chinese | French | Spanish | Russian | Japanese | Arabic | Hindi | Bengali | Urdu | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 665K-English | **67.9** | **55.7** | **62.4** | **64.5** | 55.3 | **59.2** | 38.9 | 29.4 | 13.9 | 21.8 | 46.9 |
| 150K-Chinese | 59.3 | 55.0 | 60.0 | 57.0 | 32.9 | 40.5 | 21.2 | 20.3 | 21.7 | 19.3 | 38.7 |
| 150K-French | 51.0 | 41.0 | 57.8 | 54.4 | 35.4 | 54.6 | 17.6 | 23.2 | 13.1 | 16.7 | 36.5 |
| 150K-Spanish | 61.1 | 52.2 | 54.8 | 61.6 | 50.1 | 51.7 | 27.8 | 24.4 | 15.4 | 18.5 | 41.8 |
| 150K-Russian | 55.2 | 51.1 | 62.2 | 60.6 | **57.8** | 50.9 | 25.3 | 28.2 | 13.6 | 16.7 | 42.2 |
| 150K-Japanese | 54.5 | 41.1 | 59.2 | 57.6 | 36.1 | 57.6 | 18.0 | 23.6 | 13.3 | 18.4 | 37.9 |
| 150K-Arabic | 67.8 | 42.9 | 56.4 | 54.7 | 38.4 | 44.7 | 56.0 | 25.7 | 19.4 | 33.4 | 43.9 |
| 150K-Hindi | 52.2 | 39.1 | 56.8 | 54.0 | 35.0 | 33.4 | 18.4 | 54.1 | 12.8 | 23.8 | 37.9 |
| 150K-Bengali | 26.4 | 40.2 | 56.0 | 54.5 | 37.3 | 26.0 | 12.8 | 16.3 | 34.8 | 14.0 | 31.8 |
| 150K-Urdu | 28.9 | 30.6 | 44.6 | 50.1 | 22.5 | 16.0 | 22.1 | 25.5 | 20.9 | 47.7 | 30.9 |
| Combined | 64.2 | **55.7** | 58.3 | 61.0 | 57.4 | 57.5 | **57.8** | **57.6** | **51.7** | **55.3** | **57.7** |

Table 2: **Ablation on multi-lingual fine-tuning dataset.** The table shows an effect of performance on ten languages when using fine-tuning data from different languages. Models with 7B parameters are used for this ablation.

mitigating common issues of automated translations through careful human correction. The benchmark comprises 24 diverse and challenging images from different domains, such as indoor and outdoor scenes, memes, and artwork, each with detailed descriptions and a set of 60 questions designed to test the understanding and generalization abilities of the model.

The results in Table 1 show that PALO obtains robust performance in high-resource languages, as shown by the 7/13B models scoring an average of 59.0 and 63.8 respectively across these languages. This demonstrates that our multilingual extension has been effectively integrated without compromising the original capabilities of the model. Further, the model shows good performance improvements in low-resource languages, with average scores rising from 26.0 and 26.9 to 55.6 and 59.2 points, for the 7B and 13B models, respectively.

The overall performance across all ten languages also improves, with the 7B model achieving an average score of 57.65, and the 13B model reaching 61.97. The data reflects that our approach successfully creates a more inclusive, diverse, and high-performing VLM, capable of handling the complex landscape of global languages in vision-language tasks (see Figures 4 and 5 for qualitative results).

Our mobile model demonstrates consistent improvements across both high-resource and low-resource languages, with an overall average gain of 33.9 points compared to the MobileVLM baseline of 23.9 points. Contrary to the trend observed in the 7/13B model, our mobile version also shows improvements in high-resource languages such as English and Chinese. This performance difference is attributed to the language model pretraining data. LLaMA-2 is trained on 2 trillion tokens with a better representation of high-resource languages compared to MobileLLaMA, which is predominantly trained on 1.3 trillion English tokens.

### 4.4 Ablations

Table 2 shows an ablation where we trained our 7B model on 150K translated instructions from each language and evaluated all models across all languages. The results show that the baseline performs better than the language-specific fine-tuned models for high-resource languages, including Chinese, French, Spanish, and Japanese. This is because these languages have less multi-modal data compared to the baseline (i.e., the English model is trained on 665K instructions, while language-specific models are trained on 150K instructions), and due to the noisy semi-automatic translation process. Conversely, the language-specific fine-tuned models perform better in the case of Arabic, Hindi, Bengali, and Urdu, as these languages are under-represented in the LLM pretraining data. Lastly, combined training further improves performance on low-resource languages. Further, we found that increasing the quantity of translated multi-modal training data enhances performance. For instance, translating an additional 72K instructions from the GQA dataset (Hudson and Manning, 2019) into Bengali and training with a total of 222K instructions improves Bengali results from 34.8 to 38.3. This study is limited to 150K instructions for each language due to resource constraints.

## 5 Conclusion

We introduce PALO, a polyglot LLM for 5B people, covering almost two-thirds of the world's population. It takes image and user text query as input and effectively converse in both high-resource languages such as English, Chinese, French, Spanish, Russian and Japanese, and low-resource lan-

Qué tiene de extraño esta imagen?
[Spanish-English Trans: What is strange about this image?]

La imagen es extraña porque muestra a dos personas vestidas con trajes medievales, una escena inusual. La visión de un hombre tirado en el suelo amenazado o a punta de cuchillo no es una escena típica en una tienda de comestibles.
[Spanish-English Trans: The image is strange because it shows two people dressed in medieval costumes, an unusual scene. The sight of a man lying on the ground being threatened or pointed at a knife by two men is not a typical scene in a grocery store.]

如果这张图片是故事书的一部分，这些角色会是谁？
[Chinese-English Trans: If this image were a part of a storybook, who could these characters be?]

如果这个图像是故事书的一部分，那么穿着中世纪服装的人物可能是叙述中的中心人物。 他们的服装表明他们可能是一位伟大的国王和王后。
[Chinese-English Trans: If this image were part of a storybook, the figure in medieval clothing might be the central figure in the narrative. Their clothing suggests they may have been a great king and queen.]

उस परिदृश्य का वर्णन करें जिसके कारण एक मध्ययुगीन दंपति खुद को एक समकालीन किराने की दुकान के गलियारे में पाता है?
[Hindi-English Trans: Describe a scenario that leads to a medieval couple finding themselves in the aisle of a contemporary grocery store?]

एक संभावित कहानी जिसमें अलग-अलग समय अवधि के पात्र आधुनिक समय की सेटिंग्स के साथ बातचीत करते हैं, एक समय-यात्रा साहसिक कार्य हो सकता है। छवि में, मध्ययुगीन कपड़े पहने एक पुरुष और एक महिला किराने की दुकान में जमीन पर लेटे हुए एक आदमी के साथ बातचीत करते हुए दिखाई दे रहे हैं। मध्ययुगीन पोशाक पहने हुए जोड़े समय यात्री हो सकते हैं जो गलती से वर्तमान किराने की दुकान में पहुंच गए, और वे आधुनिक दुनिया को नेविगेट करने और समझने की कोशिश कर रहे हैं। जमीन पर लेटा हुआ आदमी आधुनिक समय का व्यक्ति हो सकता है, जिसने मध्ययुगीन कपड़े पहने जोड़े का सामना किया और असामान्य स्थिति को समझने की कोशिश कर रहा है
[Hindi-English Trans: A possible story involving characters from different time periods interacting with modern-day settings could be a time-traveling adventure. In the image, a man and a woman dressed in medieval clothing are seen interacting with a man lying on the ground in a grocery store. The medieval-dressed couple could be time travelers who accidentally landed in the present-day grocery store, and they are trying to navigate and understand the modern world. The man lying on the ground could be a modern-day person who encountered the medieval-dressed couple and is trying to make sense of the unusual situation.]

هل يمكنك أن تتخيل شيئًا مضحكًا قد يقوله الملك في هذا المشهد؟
[Arabic-English Trans: Can you imagine something funny that the king might say in this scene?]

قد يقول شيء مثل: "أنا ملك في عصر ماضٍ، ولكن في هذا العصر الحديث، يبدو أن الملكية تتبع قواعد مختلفة. لذلك، سيفي لا يكون مفيدًا فقط في القتال، بل أيضًا في تنظيف الأرضيات والتعامل مع المشتتات في السوبرماركت. من المهم أن نتكيف"
[Arabic-English Trans: He might say something like: "I am a king in a bygone era, but in this modern era, kingship seems to follow different rules. So, my sword is not only useful in combat, but also for cleaning floors and dealing with distractions in the supermarket. It's important that we adapt."]

Figure 4: **Qualitative results demonstrating the multilingual capabilities of PALO**. When presented with user queries, the model generates accurate textual responses related to the visual content and the relevant language. The figure highlights its ability to bridge vision and language understanding across diverse languages. In this illustration, we explore dialogues in *two high-resource languages—Spanish and Chinese*—and *two low-resource languages—Hindi and Arabic*. PALO accurately interprets the unusual aspects of an image featuring two individuals in medieval attire within a contemporary supermarket setting. The model exhibits its creative imagination in *Chinese*, proposing a backstory where these characters might be a king and queen from a storybook. In *Hindi*, PALO demonstrates scenario-building by describing a possible situation that brought the medieval couple into the current day as time travellers. At the bottom, PALO displays a touch of humour in *Arabic*, conjuring up a playful dialogue that a king might say, showcasing its subtle understanding of context and culture-specific humour. This image effectively visualizes the advanced ability to process and generate content in multiple languages, reflecting high linguistic precision and cultural intelligence.

guages such as Arabic, Hindi, Bengali and Urdu. To train our model on ten languages, we translate 150K instructions into each language using custom-tailored LLMs. To fine-tune an LLM on a language-translation task, we use 1K human-annotated conversations for each targeted language. Our final model simultaneously provides competency in ten languages and provides an overall performance improvement on vision-language evaluation. We train PALO across three scales (1.7B, 7B, and 13B) to demonstrate its generalization and scalability across ten languages. Our codes, models, and datasets will be publicly released.

## 6 Limitations

The semi-automated translation process we employ, while efficient, might not fully grasp the deep contextual and cultural nuances inherent to each language. This could impact the capability of the model to comprehend and generate content with the necessary cultural depth, accuracy and precision. Additionally, our selection of ten languages, though it spans two-thirds of the global population, still leaves out a considerable number of the world's languages, indicating room for further expansion to further expansion to enhance linguistic diversity and inclusivity within VLMs.

Si vous deviez préparer un plat pour le petit-déjeuner avec uniquement des aliments provenant de ce réfrigérateur, que pourriez-vous préparer ?
[French-English Trans: If you were to prepare one dish for breakfast with only items from this fridge, what could you make?]

Avec les aliments disponibles dans le réfrigérateur, vous pourriez préparer un plat de petit-déjeuner sain et nutritif comme un omelette aux légumes et aux œufs. Les œufs sont stockés dans le réfrigérateur, et il y a plusieurs types de légumes, y compris des tomates, des poivrons et des oranges, qui peuvent être utilisés pour la recette. Vous pouvez également utiliser les carottes disponibles pour ajouter une touche de douceur et de saveur. Pour préparer l'omelette, vous pouvez cuire les légumes et les œufs selon vos préférences, et vous pouvez ajouter des herbes fraîches ou des épices pour rehausser le goût. Ce plat de petit-déjeuner non seulement offre une variété de saveurs, mais il est également riche en vitamines et en nutriments essentiels pour commencer la journée sur une bonne note.
[French-English Trans: With the foods available in the refrigerator, you could prepare a healthy and nutritious breakfast dish like a vegetable and egg omelet. Eggs are stored in the refrigerator, and there are several types of vegetables, including tomatoes, peppers and oranges, that can be used for the recipe. You can also use available carrots to add a touch of sweetness and flavor. To prepare the omelette, you can cook the vegetables and eggs according to your preferences, and you can add fresh herbs or spices to enhance the taste. This breakfast dish not only offers a variety of flavors and textures, but it's also rich in essential vitamins and nutrients to start the day on a good note.]

Какие продукты в этом холодильнике богаты витамином C?
[Russian-English Trans: Which items in this fridge are rich in Vitamin C?]

В холодильнике есть несколько продуктов, богатых витамином C, включая апельсины и оранжевые морковь. Апельсины известны своим высоким содержанием витамина C, который является важным для поддержания здоровой иммунной системы, здоровья кожи и общего благополучия. Оранжевые морковь также содержат витамин C, который помогает поддерживать здоровую иммунную систему, способствует здоровью кожи и способствует усвоению железа из растительных источников. Включение этих продуктов в свой рацион может способствовать получению необходимых питательных веществ для поддержания хорошего здоровья.
[Russian-English Trans: There are several foods rich in vitamin C in the refrigerator, including oranges and orange carrots. Oranges are known for their high levels of vitamin C, which is important for maintaining a healthy immune system, healthy skin and overall well-being. Orange carrots also contain vitamin C, which helps support a healthy immune system, promotes healthy skin, and promotes the absorption of iron from plant sources. Including these foods in your diet can help you get the nutrients you need to maintain good health.]

この冷蔵庫のどのアイテムを幸福の比喩として使いますか?またその理由は何ですか?
[Japanese-English Trans: Which item in this fridge would you use as a metaphor for happiness and why?]

この冷蔵庫の中のオレンジは、幸福の比喩として使用することができます。なぜなら、オレンジはビタミンC、食物繊維、抗酸化物質などの必須栄養素が豊富であり、健康的な食事に貢献するからです。さらに、オレンジは鮮やかな色と爽やかな味わいで知られており、それが幸福と関連付けられることが多いです。
[Japanese-English Trans: This orange in the refrigerator can be used as a metaphor for happiness. Because oranges are rich in essential nutrients such as vitamin C, dietary fiber, and antioxidants, which contribute to a healthy diet. Additionally, oranges are known for their bright color and refreshing taste, which is often associated with happiness.]

اگر آپ اس فرج کے مواد کو فوڈ گروپس کے لحاظ سے درجہ بندی کریں تو ہر زمرے میں کیا ہوگا؟
[Urdu-English Trans: If you were to categorize the contents of this fridge by food groups, what would be in each category?]

اس فریج میں مختلف قسم کے کھانے کے گروپ پائے جاتے ہیں۔ اس میں کالی مرچ اور گاجر سمیت سبزیوں کا ایک گروپ شامل ہے۔ اس فریج میں سبزیوں کے علاوہ پھلوں کا ایک گچھا بھی ہے جس میں سنتری بھی شامل ہے۔ اس کے علاوہ، اس فریج میں انڈوں کا ایک گچھا ہے، جو پروٹین کا ذریعہ ہیں۔ آخر میں، اس فریج میں مختلف مٹھائیاں بھی ہوتی ہیں، جو میٹھے کھانے کا ایک گروپ بناتی ہیں۔ یہ فوڈ گروپ متوازن اور متنوع کھانے کے انتخاب کی نمائندگی کرتے ہیں جو صحت مند غذا کی حمایت کرتے ہیں۔
[Urdu-English Trans: A variety of different food groups are found in this fridge. It includes a group of vegetables including pepper and carrots. Apart from vegetables, this fridge also has a bunch of fruits including oranges. Also, this fridge has a bunch of eggs, which are a source of protein. Finally, this fridge also has various sweets, which make up a bunch of sweet treats. These food groups represent a balanced and varied food selection that supports a healthy diet.]

Figure 5: **Qualitative results demonstrating the visual reasoning of PALO and its adeptness in multiple languages**. PALO responds accurately to visual content in a contextually appropriate manner for each language. We illustrate a conversation in *three high-resource languages—French, Russian and Japanese* and *one low-resource language—Urdu*. In the *French* segment, the model shows practical reasoning by suggesting a recipe that utilizes the available ingredients in the fridge, connecting visual perception to culinary suggestions. In *Russian*, PALO identifies items rich in Vitamin C and in the *Urdu* example, the model organizes the fridge contents into food groups, demonstrating its ability to classify items and apply nutritional knowledge. This effectively highlights its ability to switch between languages while maintaining the context of the conversation, reflecting its capacity to generate relevant and culturally aware content in both high-resource and low-resource languages.

## 7 Potential Risks

The use of semi-automated translations could bring forward potential risks tied to biases inherent in LLMs, particularly for low-resource languages. The model must account for nuances in visual data, such as the interpretation of cultural symbols or gestures, to prevent any misrepresentations. The interpretations of the model, influenced by these biases, could lead to inaccuracies in contexts that are culturally sensitive. There is a need to evaluate and adopt necessary training to mitigate such risks.

## 8 Use of Data and AI Assistant

We use LLaVA-Instruct (Liu et al., 2023b) dataset, licensed under Creative Commons Attribution (CCA) 4.0 International, available for use in research. Further, the use of GPT models abides by (OpenAI). Respecting source license information, we will release all datasets created in this work under CCA 4.0 International license.

## 9 Human Annotations

The LLaVA-Bench (Liu et al., 2023b) evaluation for each language is verified and corrected by annotators selected to represent a diverse mix of genders and demographics. Annotators are provided with the English version alongside the translated version. They are given specific instructions to neutralize the tone and biases during the correction process.

## References


Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, et al. 2023. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Xiangxiang Chu, Limeng Qiao, Xinyang Lin, Shuang Xu, Yang Yang, Yiming Hu, Fei Wei, Xinyu Zhang, Bo Zhang, Xiaolin Wei, et al. 2023. Mobilevlm: A fast, reproducible and strong vision language assistant for mobile devices. *arXiv preprint arXiv:2312.16886*.

Together Computer. 2023. Redpajama: An open source recipe to reproduce llama training dataset.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv:2305.06500*.

David M Eberhard, Gary Francis Simons, and Charles D Fenning. 2015. Ethnologue: Languages of the world.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Huggingface. Huggingface dataset. `https://huggingface.co/datasets/Aeala/ShareGPT_Vicuna_unfiltered`.

Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, et al. 2022. The bigscience roots corpus: A 1.6 tb composite multilingual dataset. *Advances in Neural Information Processing Systems*, 35:31809–31826.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model. corr, abs/2211.05100, 2022. doi: 10.48550. *arXiv preprint arXiv.2211.05100*.

Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. 2023. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. *arXiv:2310.03744*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. In *NeurIPS*.

Junyu Lu, Dixiang Zhang, Xiaojun Wu, Xinyu Gao, Ruyi Gan, Jiaxing Zhang, Yan Song, and Pingjian Zhang. 2023. Ziya-visual: Bilingual large vision-language model via multi-task instruction tuning. *arXiv e-prints*, pages arXiv–2310.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2023. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv:2306.05424*.

OpenAI. Openai terms of use. https://openai.com/policies/terms-of-use.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision.

Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M. Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S. Khan. 2023. Glamm: Pixel grounding large multimodal model. *ArXiv 2311.03356*.

Teven Le Scao, Thomas Wang, Daniel Hesslow, Lucile Saulnier, Stas Bekman, M Saiful Bari, Stella Biderman, Hady Elsahar, Niklas Muennighoff, Jason Phang, et al. 2022. What language model to train if you have one million gpu hours? *arXiv preprint arXiv:2210.15424*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Xiangpeng Wei, Haoran Wei, Huan Lin, Tianhao Li, Pei Zhang, Xingzhang Ren, Mei Li, Yu Wan, Zhiwei Cao, Binbin Xie, et al. 2023. Polylm: An open source polyglot large language model. *arXiv preprint arXiv:2307.06018*.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv:2306.05685*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv:2304.10592*.