X-LLaVA: Optimizing Bilingual Large Vision-Language Alignment

Dongjae Shin[‡*], **Hyeonseok Lim**[*], **Inho Won**[‡], **Changsu Choi, Minjun Kim, Seungwoo Song, Hangyeol Yoo, Sangmin Kim, Kyungtae Lim**[†]
Seoul National University of Science and Technology
[‡]Teddysum
{dylan1998,gustjrantk,wih1226,choics2623,mjkmain}@seoultech.ac.kr
{sswoo,21102372,sangmin6600,ktlim}@seoultech.ac.kr

Abstract

The impressive development of large language models (LLMs) is expanding into the realm of large multimodal models (LMMs), which incorporate multiple types of data beyond text. However, the nature of multimodal models leads to significant expenses in the creation of training data. Furthermore, constructing multilingual data for LMMs presents its own set of challenges due to language diversity and complexity. Therefore, in this study, we propose two cost-effective methods to solve this problem: (1) vocabulary expansion and pretraining of multilingual LLM for specific languages, and (2) automatic and elaborate construction of multimodal datasets using GPT4-V. Based on these methods, we constructed a 91K English-Korean-Chinese multilingual, multimodal training dataset. Additionally, we developed a bilingual multimodal model that exhibits excellent performance in both Korean and English, surpassing existing approaches.

1 Introduction

Recently, large multimodal models (LMMs) have evolved to respond in alignment with human intent through visual instruction-following (VIF) (Liu et al., 2023a; Dai et al., 2023; Bai et al., 2023; Chen et al., 2023a; OpenAI, 2023). In LLaVA1.0 (Liu et al., 2023b), a method was proposed to automatically construct a VIF dataset using GPT4, which demonstrated excellent performance in visual question answering (VQA). However, there are two main limitations to the data generated in LLaVA1.0: first, it was constructed using a text-only version of GPT4, which does not accept images as input; and second, it targeted only English.

Subsequently, LLaVA1.5 (Liu et al., 2023a) incorporated the multilingual instruction dataset ShareGPT (sha), demonstrating its potential in multilingual processing. However, ShareGPT uses an instruction following (IF) (Chen et al., 2023a) dataset for LLMs, still suffers from a lack of vision information. To address this issue, ShareGPT4V (Chen et al., 2023b), a VIF dataset created using GPT4-V, which accepts image information as input, was released. ShareGPT4V is also limited because it consists only of English question-answering, posing a constraint in aligning multiple languages to acquire multilingual information.

In this context, we propose constructing a multilingual VIF dataset based on object relational information and a multilingual LMM that efficiently utilizes this dataset. The proposed multilingual VIF dataset was composed of 23,496 question-and-answer pairs centered around objects, locations, atmospheres, and conversations to ensure the diversity of expressions. The target languages were selected considering linguistic diversity by choosing English, Chinese, and Korean, which belong to different language families (FitzGerald et al., 2023; Park et al., 2021).

We also propose the development of a multilingual LMM, X-LLaVA, utilizing the proposed data. X-LLaVA is a model that enhances LLaVA1.5, by applying the following three enhancement methods: (1) **vocabulary expansion** for target language, (2) pretraining for **connecting knowledge** across multiple languages, and (3) **multilingual VIF**. First, bilingual-based vocabulary expansion involves adding words to a pretrained language model to strengthen the relatively limited vocabulary of Korean compared to English (Lu et al., 2023; Cui et al., 2023). Second, additional pretraining was conducted to link the English and Korean knowledge. Third, we conducted multilingual training using the proposed VIF dataset.

Experimental results showed that the X-LLaVA model demonstrated an average improvement of approximately 5.2% in three Korean quantitative evaluations compared to the previously proposed

*These authors contributed equally.
†Corresponding author.

Findings of the Association for Computational Linguistics: NAACL 2024, pages 2463–2473
June 16-21, 2024

Table 1: Summary of multi-modal instruction tuning datasets. 'Visible' refers to the including of images in the data generation process. The availability of a 'Parallel' pertains to whether the dataset can be used translation task.

Dataset	Domain	Data Type	# of Words	Visible	Captioned by	# of Instances	Multilingual	Parallel	Open
MiniGPT4	Daily life	Description, Discourse	80 ∼	✗	Template-based	5K	✗	✗	✓
MultiInstruct	General	Description, Reasoning	∼ 100	✗	Template-based	∼ 235K	✗	✗	✗
InstructBLIP	Daily life	Description, Reasoning, Discourse	∼ 200	✗	Template-based	∼ 1.6M	✗	✗	✗
LLaVA	Daily life	Description, Reasoning, Discourse	∼ 200	✗	GPT-based	1.15M	✗	✗	✓
MultiModalGPT	General	Description, Discourse	∼ 200	✗	GPT-based	6K	✗	✗	✗
SharedGPT4V	General	Description, Reasoning, Discourse	∼ 200	✓	GPT-based	100K	✗	✗	✓
LVIS-INSTRUCT	Daily life	Description	∼ 100	✓	GPT-based	220K	✗	✗	✓
M^3IT	General	Description, Reasoning	∼ 200	✗	GPT-based	2.4M	✓	✗	✓
Ours	Daily life	Description, Discourse	∼ 200	✓	GPT-based	91K	✓	✓	✓

KoLLaVA model. In addition, it achieved the highest performance in two out of five English quantitative evaluations. In qualitative evaluations, preference assessments using GPT4-V demonstrated that our model generated responses in both English and Korean that were 19-93% superior to existing models. Through qualitative analysis, we highlighted that the proposed bilingual training enhanced specific language vocabulary, leading to better performance in writing evaluations. The contributions of this study can be summarized as follows:

- We propose a training framework of multilingual LMM for enriching a specific language availability
- We have constructed multilingual VIF dataset based on different task-oriented types
- Through an in-depth analysis, we demonstrate the real-world effectiveness of the multilingual approach employed in our dataset.

Finally, we emphasize that the 91K datasets and models constructed in this study can be implemented with relatively small resources, costing approximately $3,200 and utilizing an A6000 GPU.

2 Related Work

2.1 Vision-Language Models

With the advancement of LLMs, proposals have been made to extend LLMs to include additional modalities (Zhang et al., 2023). The primary idea was to focus on aligning information between vision and language (Alayrac et al., 2022). A prime example of this is CLIP (Radford et al., 2021) and ALBEF (Li et al., 2021), which integrated representations of images and text using contrastive learning (Chen et al., 2020; Lee et al., 2022) to unify distinct types of information. Subsequent enhancements, as observed in BLIP (Li et al., 2022) and BLIP-2 (Li et al., 2023b), utilized assorted data and Q-Former's trainable query vectors to strengthen this alignment. Most recently, MiniGPT4 (Zhu et al., 2023) proposed a fine-tuning method to generate responses that are more aligned with the user intent, demonstrating the potential for conversational image-text models. Concurrently, InstructionBLIP (Dai et al., 2023), LLaVA1.0 (Liu et al., 2023b), and LLaVA1.5 (Liu et al., 2023a) have advanced our understanding of complex prompts through more sophisticated visual instruction fine-tuning (VIT) (Liu et al., 2023b).

2.2 Visual Instruction Following Datasets

In LLMs, IF is used to ensure that the language model generates responses that align with user objectives. Recently, there has been a proposal for research to create a VIF dataset that includes image data in the IF. The construction of a VIF dataset is costly and time-consuming because it requires the simultaneous consideration of images, queries, and answers. Therefore, automatic generation methods are commonly used, with two primary approaches: one using GPT for data generation and the other using a template-based method that transforms existing data using predefined templates.

Table 1 presents a comparison of the representative VIF datasets. The initial versions of the VIF dataset were constructed using template-based models. Multi-Instruct (Li et al., 2023a) and InstructBLIP, which fall under this category, are fast and cost-effective as they involve rule-based transformation of existing data. However, they have the limitation of being oriented towards specific tasks such as image captioning or classification.

In contrast to template-based construction, LLaVA introduces a more flexible generative data construction method that utilizes the GPT. Using object location and caption information from COCO (Lin et al., 2014), LLaVA constructed 158K diverse VIF datasets with three different styles: detailed description, complex reason-

Main objects

box, fruit, oranges, apples, pole, sticker, apple, orange

System message for object-centric data generation
You're a helpful vision AI assistant. You are given an image and a main object. Your task is to generate question-and-answer data that strictly focuses on the objects and elements that are clearly visible and identifiable in the image. Ensure that your descriptions are clear, factual, and definitive. Avoid any speculative, uncertain, or imaginative descriptions. Do not include or mention any elements that are not present in the image. Provide accurate and reliable question and answer data, based on what is definitively observable within the image. The question & answer data should be provided in the following order: English, Korean, Chinese.

System message for location-centric data generation
You are a good vision AI assistant. You are given an image and its main objects. Your task is to generate locational scene graph, question-and-answer data that focuses solely on the location of clearly visible and identifiable objects in the image. Make sure your descriptions are clear, factual, and definitive. Avoid speculative, uncertain, or imaginative descriptions. Do not include or mention elements that do not exist in the image. Provide accurate and reliable question and answer data based on what you can reliably observe in the image. The orientation of left, right, etc. is based on the person looking at the image. The question & answer data should be provided in English, Korean, and Chinese.

System message for atmosphere-centric data generation
You are a proficient vision AI assistant. You are presented with an image. Your task is to generate question and answer data that focuses on the overall ambiance and mood of the image. Ensure that your descriptions are clear, factual, and definitive, capturing the essence of the image's atmosphere. Avoid speculative, uncertain, or imaginative interpretations. Provide accurate and reliable question and answer data based on what you can definitively observe in the image. The question & answer data should be provided in English, Korean, and Chinese.

System message for conversation data generation
You are a useful AI assistant. I will provide you with two images and an 8-Turn Question-Answer Pair sample for each image. Based on the provided example images and 8-Turn QA samples, create an 8-Turn Question-Answer Pair for the last image you provide. Do not reference uncertain details when generating data. Provide detailed answers to complex questions. For example, present detailed examples or reasoning steps to make the content more persuasive and well-organized. Include multiple paragraphs if necessary. Create in the same format as the example templates, and generate Question-Answer Pairs in Korean, English, and Chinese.

Figure 1: System messages for four types of `mvif` dataset

ing, and conversational. However, because these datasets do not use images in their generation, SharedGPT4V (Chen et al., 2023b), and LVIS-INSTRUCT4V (Wang et al., 2023), which include images in their construction, were proposed. However, these datasets are predominantly written in a single language. To address the need for multilingual capabilities, the M^{3}IT dataset was released (Li et al., 2023c). M^{3}IT is an instruction-tuning dataset comprising 40 tasks translated into 80 languages that offers broad accessibility.

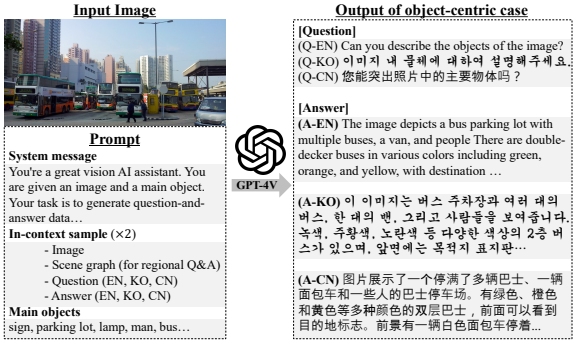

Figure 2: An example of prompt and result using data construction.

3 Data Generation

In this study, we were inspired by the VIF data generation method using the GPT of LLaVA and have built upon it. However, to minimize the loss of information from the images and include more detailed information, we directly input the image and object information into the GPT4-V model to construct our data. We constructed four types of multilingual VIF datasets (`mvif`) for three languages (English, Korean, and Chinese): (1) Object-centric, (2) Location-centric, (3) Atmosphere-centric, and (4) Conversation.

3.1 The Focus of Data Building

The `mvif` data proposed in this research concentrate on the relational factual information between objects. This focus diverges from the description and reasoning-centered question-answering proposed by LLaVA, leading to minimal information redundancy between the two datasets. Although LLaVA's data are commendable, we assessed whether data designed for reasoning purposes might incorporate subjective viewpoints, thereby potentially introducing bias toward certain objects. Therefore, our study aims to develop a functional-relationship-based multilingual VIF dataset that, deliberately avoids overlap with LLaVA.

The target languages selected were English, Chinese, and Korean, each belonging to a distinct language family. This choice was intended to evaluate how multilingual training affects the languages of different cultures and character systems.

3.2 Image Selection Criteria

To construct the `mvif` dataset, 23,496 images from the visual Genome (Krishna et al., 2017) were used. A challenge was encountered when generating data using GPT4: if an image contained fewer than three major objects, the constrained context could limit the diversity of question answers. However, answering questions generated using images with over ten objects often results in a focus on objects that are either exceedingly small or insignificant. Consequently, we speculate that images selected from the visual Genome, where the number of main objects corresponds to $3 \leq m \leq 10$.

3.3 Proposed VIF Dataset

Figure 2 shows an example of the method used to construct the proposed `mvif` dataset. As illustrated, an image and a prompt, which are metadata for question generation, were fed into GPT4-V. Subsequently, GPT4-V was designed to generate questions and answers in three languages. For conversation data, we designed a prompt to produce eight pairs of dialogues for each image in a multi-turn format. For the dataset construction, we provided two seed examples to GPT4-V to guide the construction of data suitable for the purpose through in-context learning. A total of $3,200 was used to generate 91K data points. Detailed prompts used in data construction can be found in Figure 1.

(1) Object-centric image description. Object-centric data focuses on providing detailed description of objects in an image, comprising questions and answers that include the shape, condition, and characteristics of the objects. The aim of constructing these data was to facilitate the learning of the intimate details of images by focusing on the specific attributes of the objects as they appear. Additionally, as shown in the "Main objects" section of Figure 2, a list of main objects was inputted into the GPT4-V prompt to prevent errors in object specification that might occur during question generation.

(2) Location-centric image description. Location-centric data is a type of question-answering data that focuses on describing the relative positions of objects within an image. However, when the same object appears multiple times in an image, this perspective can alter the location information. To address this effectively, we enabled GPT4-V to autonomously generate a relationship graph that served as the basis for answering the question. Consequently, when GPT4-V receives an image and a list of objects, it first generates a scene graph and then produces locational questions and answers regarding the image.

(3) Atmosphere-centric image description. Atmosphere-centric data include descriptions that focus more on the overall ambiance of an image than on individual objects. It encompasses a holistic depiction of the complex interplay among multiple objects.

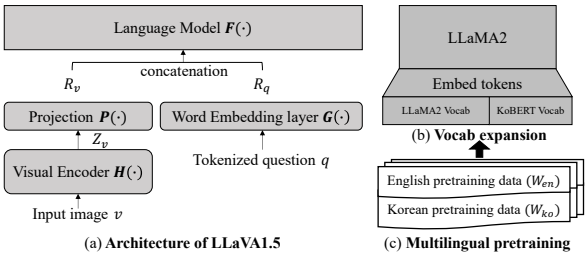

Figure 3: (a) Architecture of LLaVA1.5 & (b,c) The proposed language model pretraining

(4) Conversational question and answering Conversational data is structured as an 8-turn Q&A dataset to incorporate more in-depth and extensive information regarding the images. Unlike other datasets, this dataset is designed to infer human emotions or include subjective information about the mood of the image.

4 Proposed Multilingual Model

In this section, we introduce the proposed X-LLaVA model, an effective approach for multilingual processing through multilingual VIT (Liu et al., 2023b). X-LLaVA applies the following three enhancement methods to the same model structure as LLaVA1.5: (1) vocabulary expansion for the target language, (2) pretraining for multilingual knowledge association, and (3) multilingual VIT. Figure 3 demonstrates the three proposed methods and the structure of LLaVA1.5.

4.1 Recap of LLaVA1.5

Figure 3 (a) shows the basic structure of the LLaVA1.5 model. LLaVA1.5 basically consists of a visual encoder and an LLM for natural language generation. The visual encoder utilizes a pretrained CLIP's Vision Transformer (Yuan et al., 2021) $H(\cdot)$, and the LLM $F(\cdot)$ utilized the pretrained LLaMA2-based models (Touvron et al., 2023; Peng et al., 2023). LLaVA uses image v and query q as inputs. In the case of image v, the output representation from the visual encoder, $H(v) = Z_v \in \mathbb{R}^{576\times1024}$, is converted into a vision-language representation $R_v \in \mathbb{R}^{576\times5120}$ through a projection layer $P(\cdot) : \mathbb{R}^{1024} \rightarrow \mathbb{R}^{5120}$. For text q, it passes through the embedding layer $G(\cdot)$ of LLaMA to generate the text representation $G(q) = R_q \in \mathbb{R}^{(|q|,5120)}$. R_q and R_v, generate through these two processes are concatenated and then passed through the entire layer of the LLaMA2 to produce a response. In this context, the projection layer serves the function of transforms image

representation Z_v into a word embedding format that can be understood using the LLaMA2.

To achieve image-language alignment, we train the process to connect the two representations, which LLaVA does in two steps. The first is image-text alignment through image captioning, and the second is VIT. X-LLaVA is trained in the same manner, and the details of the two phases are described in Section 4.3.

4.2 Enriching the LLM Vocabulary

In the LLaVA model, when querying in Korean for the LLaMA2-13B language model, issues arise, such as responses in English or English-Korean code-switching. This stems from a problem with the tokenizer, where 89.7% is in Latin script, while Korean only constitutes 0.37%, leading to insufficient Korean expressiveness and biases in the pretraining data owing to lexical bias. To address these issues, we expanded the Korean vocabulary in the LLaMA2 and conducted additional pretraining for knowledge infusion. (Figure 3 (b), (c))

Vocabulary expansion involves adding 7,478 words from the KoBERT[1] vocabulary to the LLaMA2 tokenizer. And we randomly initialize embeddings for these newly added words. Ultimately, the proposed tokenizer possessed a dictionary of 39,478 entries. As a subsequent step, the model was further enhanced with knowledge information using English Wikipedia data W_{en} and Korean Wikipedia data W_{ko}. Through this process, our model learns representations for the newly added vocabulary. If the pretraining dataset (7.8GB) is defined as $D_{pt} = \{\mathrm{W}_{en}, \mathrm{W}_{ko}\}$, then the loss function $\mathcal{L}_{PT}(\cdot)$ is expressed as follows.

$$\mathcal{L}_{PT}(\theta) = -\sum_{i}^{|D_{pt}|}\sum_{j}^{|x_i|}\log P(x_{i,j}|x_{i,<j};\theta) \quad (1)$$

Here, $|D_{pt}|$ is the size of D_{pt}, $|x_i|$ denotes the number of tokens in i-th data sample x_i. $x_{i,j}$ represents j-th token of sequence x_i, and $x_{i,<j}$ represents the sequence of tokens before the j-th token. In this context, $\mathcal{L}_{PT}(\theta)$ is the causal language modeling loss function, where θ denotes the model parameters.

4.3 X-LLaVA

In this section, we describe the method for training X-LLaVA using the LLaMA2 model, which has proceeded word expansion and bilingual dictionary pretraining, as previously introduced X-LLaVA, like LLaVA, is trained in two stages: image-language connection via captioning and multilingual VIT. However, unlike LLaVA1.5, to efficiently conduct multilingual training, we follow the cross-lingual language model pretraining method (Conneau and Lample, 2019), simultaneously utilizing a mix of English and Korean for training.

In the first stage, we train only the projection layer $P(\cdot)$ using the image-caption datasets LLaVA-CC3M (Liu et al., 2023b) (C_{en}) and its machine-translated Korean counterpart, LLaVA-KoCC3M(C_{ko}). This stage involves representation learning in which image representations are converted into word embeddings that are comprehensible to the LLaMA2. During this process, both Korean and English are learned concurrently while simultaneously aligning [image-English-Korean]. We define the dataset for Stage-1 as $D_{s1} = \{C_{en}, C_{ko}\}$.

In the second stage, we conducted VIT on X-LLaVA to enhance its capabilities as a multilingual visual assistant. For VIT as described in (Liu et al., 2023b), we use the LLaVA instruct dataset (158K, L_{en}), its machine-translated counterpart (158K, L_{ko}), and the `mvif` dataset (91K, L_{our}) generated in Section 3. In this stage, unlike the first stage, we train the projection layer and language model simultaneously. Define the dataset for Stage-2 training as $D_{s2} = \{L_{en}, L_{ko}, L_{our}\}$. The formula for training the Stage-2 can be expressed as follows:

$$\mathcal{L}_{s}(\theta) = -\sum_{i}^{|D_s|}\sum_{t}^{T}\sum_{j}^{|a_i^{(t)}|}\log P(a_{i,j}^{(t)}|X_{i,<j}^{(t)};\theta) \quad (2)$$

Where $X_{i,<j}^{(t)} = \{v_i, q_i^{(1)}, a_i^{(1)}, \cdots, q_i^{(t)}, a_{i,<j}^{(t)}\}$, T represents the total number of conversation turns. In Stage 1, $T = 1$ because the dataset D_{s1} is composed of a single turn. In Stage 2, $T = 1$ is also true in all case, except for multi-turn conversations.

In the dataset D_s, which can be either D_{s1} or D_{s2} depending on the stage, v_i, $q_i^{(t)}$, and $a_i^{(t)}$ denote the i-th component of the image, the question (instruction) in turn t, and the answer in turn t, respectively.

5 Quantitative Evaluation

In this section, we describe the quantitative evaluation methods and criteria for the proposed X-LLaVA. Through these comparisons, we aim to

[1] https://github.com/SKTBrain/KoBERT

address the three research questions proposed in Section 1: (1) What impact does vocabulary expansion, intended to enhance multilinguality, have on vision-language models? and (2) How does bilingual training affect the relationship between these two languages? and (3) Which aspects of the model were strengthened by utilizing our proposed `mvif` data?

5.1 Experiment Environments

To ensure a fair comparison of LMMs, we must define task selection for evaluation and specify the LMM model used for evaluation. Below are the benchmark datasets used for evaluation, with the following characteristics for each benchmark:

- **(English) VQA2.0:** A dataset containing open-ended questions about images (Goyal et al., 2017), **GQA:** A VQA-format dataset considered Scene Graph (Hudson and Manning, 2019), **LV** (LLaVA^{w} from (Liu et al., 2023b)) and **POPE** (Yifan Li and Wen, 2023)
- **(Korean) KoViz:** A VQA-format dataset and **KoLiv:** A VQA-format dataset considered Korean culture and daily life (Kim et al.)
- **(English-Korean) BVQA** (Kim et al., 2024): A VQA dataset considering **B**ilingual Outside Knowledge

For our experiments, we converted the VQA2.0 and BVQA (Kim et al., 2024) datasets into the VIF format using the VQA-to-VIF data transformation method proposed in LLaVA1.5. Following this conversion, we proceeded with VIT over all the training sets from the proposed benchmark in only one epoch. The evaluation methodology and prompts were adopted directly as proposed in LLaVA1.5 . Experimental environments and answers generated for each model were made publicly accessible[2] to ensure reproducibility and facilitate comparison of the models.

5.2 Intrinsic Evaluation of X-LLaVA

An intrinsic evaluation was conducted to explore the three research questions we proposed. To achieve this, we train the three models under different conditions. Table 2 lists the training environments and performances of the three models. X-LLaVA refers to the model that underwent both vocabulary expansion and knowledge enhancement (4.2) as well as the VIT (4.3) proposed in Section 4. X-LLaVA(-P) is a model created to compare the effects of pretraining methods on Koreans and English data proposed in Section 4.2. This model is a version of X-LLaVA that does not utilize Wiki for **p**retraining during its training phase. X-LLaVA(-V,-P) represents a model that neither underwent **v**ocabulary expansion nor used Wiki for **p**retraining, essentially using pure LLaMA2. Finally, to assess the impact of the `mvif` data proposed in Section 3, we compared the results of each model with and without the addition of `mvif`.

Model	VIF	BVQA^{k}	BVQA^{e}	GQA
XLLaVA(-V,-P)		51.5	33.0	62.3
	+ O	51.9	**36.0**	61.9
XLLaVA(-P)		56.4	32.0	62.1
	+ O	56.6	32.3	62.5
XLLaVA		57.6	33.5	63.3
	+ O	**57.9**	34.3	**64.0**

Table 2: Intrinsic evaluation. Where (-V) represents without vocabulary expansion, and (-P) denotes without multilingual pretraining step. Metric is Accuracy(%).

The influence of Enriching Vocabulary. Comparing the X-LLaVA and X-LLaVA(-V,-P) models in Table 2, we observe an average of 6.1 points for Korean and 0.8 points for English. Therefore, the vocabulary expansion and pretraining proposed in Section 4.2 not only significantly improves the Korean performance of the model with expanded vocabulary but also enhances the performance of the existing English model.

The influence of Pretraining. A comparison between the X-LLaVA and X-LLaVA(-P) models showed that additional pretraining using Wikipedia uniformly enhanced the performance in both Korean and English, with a particularly notable improvement in Korean. Therefore, the effectiveness of pretraining in Korean and English using Wikipedia was evident.

The influence of VIT using `mvif`. When models were tuned with the proposed dataset (+O), a performance improvement ranging from 0.2 to 3 was observed across almost models for the target language. Although the extent of improvement is modest, it is noteworthy that despite the grammatical differences between Korean and English, where knowledge loss might be anticipated, there was

[2] github.com/MLP-LAB/X-LLaVA

LMM	LLM	#PT	#VIT	BVQA^{k}	KoViz	KoLiv	BVQA^{e}	VQA	GQA	LV	POPE
BLIP-2	Vicuna13B	129M	-	-	-	-	-	41	41	-	85.3
InstructBLIP	Vicuna7B	129M	1.2M	-	-	-	-	-	49.2	-	-
InstructBLIP	Vicuna13B	129M	1.2M	-	-	-	-	-	49.5	-	78.9
LLaVA1.5	Vicuna7B	558K	665K	16.2	33.9	44.9	25.1	78.5	62.0	64.7	**85.9**
LLaVA1.5	Vicuna13B	558K	665K	27.9	24.4	33.4	26.1	**80.0**	63.3	65.7	**85.9**
LLaVA1.5(O)	Vicuna13B	558K	756K	32.6	24.6	23.2	29.1	78.1	45.3	**70.4**	85.8
LLaVA1.5(B)	Vicuna13B	558K	857K	54.5	50.3	52.1	33.5	76.4	63.0	22.8	85.8
KoLLaVA	Synatra7B	595K	612k	45.3	**55.9**	54.2	5.5	-	-	-	-
X-LLaVA	Ours	1.2M	407K	**57.9**	51.3	**61.7**	**34.3**	75.5	**64.0**	57.5	85.5

Table 3: Extrinsic evaluation results. Where (O), (B) represents training with `mvif` and BVQA dataset,`#PT` is the number of pretraining data, `#VIT` is the number of VIT data. POPE is a benchmark for evaluation of hallucination.

an observable enhancement in the English performance. This indicates that multilingual VIF can be expected to improve performance in both less- and high-resource languages.

5.3 Extrinsic Evaluation of X-LLaVA

We conducted a comparative evaluation of the performance of our X-LLaVA model in Korean and English against other LMMs. The models compared were BLIP-2, InstructBLIP, LLaVA1.5, and KoLLaVA, and the distinctive features of each model are presented in Table 3.

Overall. In the Korean evaluation (BVQA^{k},Koviz, and KoLiv) presented in Table 3, X-LLaVA demonstrated significantly higher performance, scoring on average 57.0 points. Interestingly, in the case of English (VQA, GQA, BVQA^{e}, LV, POPE), X-LLaVA also showed the highest performance in BVQA^{e} and GQA.

The effect of multilingual training. Typically, when training languages with different character systems, the performance of a relatively highly resourced language may deteriorate (Pires et al., 2019). However, when the multilingual training methods and data (`mvif`) we proposed, no decrease in performance was observed. When comparing the English BVQA^{e} and GQA scores of LLaVA1.5 and X-LLaVA, they showed 8.2 and 0.7 points higher performance, respectively. However, for VQA2.0, LLaVA1.5's performance was 4.5 points higher. During analysis, we observed that X-LLaVA generally performed better on GQA and BVQA, which asked about relationships and knowledge.

Comparison of X-LLaVA with KoLLaVA. KoLLaVA[3] is the Korean version of LLaVA1.5, a model trained after automatically translating CC3M, VQA2.0, GQA, and Visual Genome data used in LLaVA1.5. Additionally, it was trained using the Korean version of the BVQA. However, as only the 7B model is currently publicly available, it may be challenging were used to evaluate the same levels. However, the published LLaVA1.5 13B model shows an average of 0.96 points higher in english than that of the 7B model, X-LLaVA demonstrates a 5.2 point higher result in korean than KoLLaVA.

Comparison X-LLaVA with LLaVA1.5(O or B). LLaVA1.5 was trained on about 1.5 times more data (665K VIFs) then X-LLaVA. Nevertheless, BVQA data has never been utilized for training, which may be disadvantageous for the BVQA evaluation. We trained LLaVA1.5 on Korean and English data for three 3 epochs to tune the BVQA for a fair evaluation. LLaVA1.5(B) in Table 3 shows the results of the model tuned using the BVQA data. The results show a significant improvement in Korean performance on the BVQA. On the other hand, this model, being biased towards VQA data, showed lower performance in the writing evaluation (LV). Conversely, LLaVA1.5(O) in Table 3, a model trained on the LLaVA1.5 with `mvif` data, exhibited the highest performance on LV.

6 Qualitative Evaluation

In this section, we describe the qualitative evaluation methods and the results for X-LLaVA. In contrast to quantitative evaluations, which are similar to classification assessments, qualitative evaluations, such as writing evaluations, differ significantly. Although human evaluation may be the

[3] github.com/tabtoyou/KoLLaVA

fairest approach to qualitative assessments, it is practically challenging. Therefore, in LIMA (Zhou et al., 2023), a GPT preference evaluation method that closely resembles human evaluation results was proposed.

In our study, we directly employed the GPT preference evaluation method. The process is as follows: First, we input an image and a question into two models being compared to obtain answers A and B. Then, we provided GPT4 with the image, question, and both answers to receive feedback such as 'Answer A is better', 'Answer B is better', or 'Both answers are similar', and measured the proportions. To compare the standing and generation abilities of recent LMMs in vision language, we used the GPT evaluation dataset proposed by LLaVA[4]. However, because this dataset is in English, we translated it into Korean, followed by a review from five annotators to ensure data quality. Afterward, we proceeded with the evaluations.

6.1 Preference Evaluation using GPT4-V

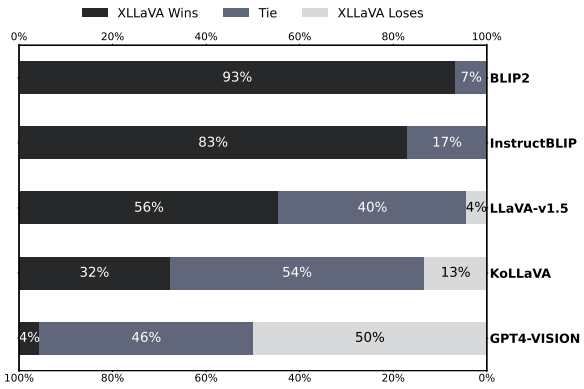

Figure 4: Korean Preference evaluation results by GPT4-V

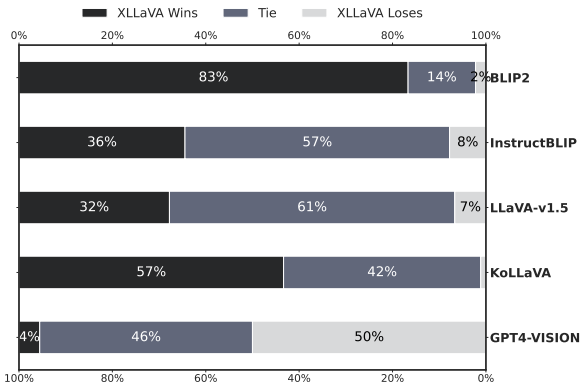

Figure 5: English Preference evaluation results by GPT4-V

Comparing X-LLaVA with others in Korean. Figure 4 presents the results of the GPT preference evaluation for each model. The X-LLaVA model outperformed all other models, except for the GPT4-V model. Notably, it obtained a 19% higher preference rate than the KoLLaVA, indicating the exceptional effectiveness of the proposed methods and datasets in enhancing Korean writing skills.

[4] 'qa90_gpt4_answer' at github.com/haotian-liu/LLaVA

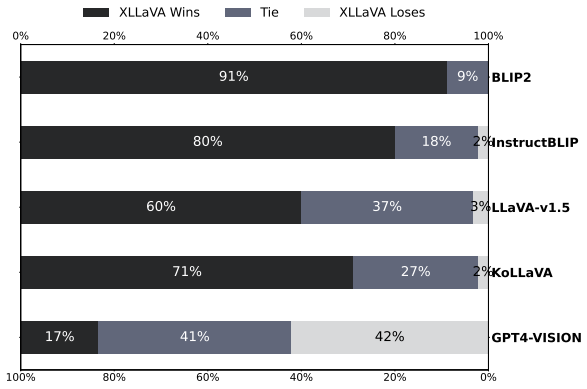

Figure 6: Korean Preference evaluation results by GPT4-V when limited to 30 Words.

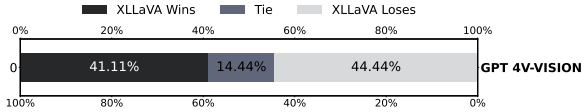

Figure 7: Preference evaluation results by human

Comparing X-LLaVA with Others in English. Figure 5 shows the results of English GPT preference evaluations. Interestingly, similar to Korean, the X-LLaVA received approximately 25% higher preference scores for English than LLaVA1.5. This indicates that pretraining of our proposed LLM and `mvif` datasets can also enhance English writing abilities.

X-LLaVA vs GPT4-V. Therefore, does evaluator GPT4-V generate better answers than X-LLaVA? We conducted the evaluations by comparing the GPT4-V and X-LLaVA models. Experimental results show that for both languages, GPT4-V's answers are preferred over those of X-LLaVA, with a significant performance difference. However, these results stem from GPT4-V generating answers that are more than 30% longer and more verbose compared to LLaVA-based models. This may also be because the GPT rates its own generated content more favorably as it becomes more familiar with it. To mitigate this, in experiments where the answers were limited to 30 words, the results changed significantly, with GPT scoring 42 compared to 17 for X-LLaVA, as shown in Figure 6.

Evaluator	XLLaVA Wins	Tie	XLLaVA Loses
GPT4-V(G)	15	37	38
Human(H)	37	14	39
G ∩ H	12	10	32

Table 4: It displays the number of samples chosen by GPT4-V and Human Evaluators for 'XLLaVA Wins', 'Tie', and 'XLLaVA Loses', respectively in Figure 6 and 7. 'G ∩ H' signifies instances where both evaluators (Human, GPT4-V) indicate the same outcome for each of the 90 samples.

6.2 Human-assisted Preference Evaluation

As previously described, the performance of GPT preference evaluation may vary according to the number of words. Consequently, a question arises: Can LIMA's assertion that GPT evaluations are akin to human assessments be extended to the vision-language model proposed in this study? We conducted a human preference evaluation using three human annotators. The Human Preference Evaluation was carried out with three evaluators using the following criteria: For a result to be classified as 'XLLaVA Wins,' either all three evaluators needed to select it or at least two did. A 'Tie' was determined either when all evaluators agreed on it or when their selections were evenly split across 'XLLaVA Wins,' 'Tie,' and 'XLLaVA Loses.' Similarly, 'XLLaVA Loses' was classified when all three agreed on it or at least two of the three chose it. Figure 7 presents the results of the human evaluation for GPT4-V and X-LLaVA in the comparative assessment, with the response length restricted to 30 words. Although GPT maintained a slight advantage, the preference scores were almost identical, as shown in Table 4. However, we observed that GPT evaluations resulted in ties 2.9 times more frequently than human evaluations. This observation can be interpreted to suggest that GPT tends to avoid ambiguous decisions compared to humans, who possess relatively clear criteria. Thus, the vision-language model can be considered as augmenting rather than substituting human evaluations.

7 Conclusion

In this study, we propose a framework for constructing data and training models for the efficient multilingual expansion of LMM. For data construction, we suggested a method to easily build multilingual VIF dataset based on the relational metadata between images and objects using GPT4-V. We also demonstrated a framework for efficient multilingual learning, which includes vocabulary enhancement, knowledge reinforcement based on pretraining, and a multilingual VIT framework. The experimental results confirmed that the proposed X-LLaVA model exhibited similar or superior performance compared to existing models that primarily focused on Korean and English as single languages. Finally, our proposed multilingual expansion framework can be trained in 7.5 days with a single A6000 GPU, and the 91K training data can be managed with relatively minimal resources, costing around $3,200.

Limitations

The ultimate goal of this research is to create a multilingual Large Multimodal Model (LMM). However, in this study, we first conducted pretraining in Korean-English and then proceeded with multilingual visual instruction following in Korean-English-Chinese. Consequently, as the Chinese component of the model did not undergo word expansion, it more closely resembles a Korean-English bilingual enhanced model. Therefore, there is a need for further investigation and research into models that have undergone vocabulary enhancement and knowledge connection for more than three languages. An additional factor was the difficulty in finding publicly available Chinese VQA evaluation data, which hindered diverse assessments.

Acknowledgements

This research was supported by the National Research Foundation of Korea (2021R1F1A1063474) for KyungTae Lim and Institute of Information & communications Technology Planning & Evaluation (IITP) by the Korea government(MSIT) (2022-0-00078, Explainable Logical Reasoning for Medical Knowledge Generation). This research used datasets from The Open AI Dataset Project (AI-Hub) (No. 2022-데이터-위41, 2023-지능데이터-위93).

References

Sharegpt. https://sharegpt.com/%7D%7D,year={2023}.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel

Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikoł aj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. 2022. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 23716–23736. Curran Associates, Inc.

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

Delong Chen, Jianfeng Liu, Wenliang Dai, and Baoyuan Wang. 2023a. Visual instruction tuning with polite flamingo.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023b. Sharegpt4v: Improving large multi-modal models with better captions.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and effective text encoding for chinese llama and alpaca.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. 2023. MASSIVE: A 1M-example multilingual natural language understanding dataset with 51 typologically-diverse languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4277–4302, Toronto, Canada. Association for Computational Linguistics.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

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.

Jin-Hwa Kim, Soohyun Lim, Jaesun Park, and Hansu Cho. Korean localization of visual question answering for blind people.

Minjun Kim, Seungwoo Song, Youhan Lee, Haneol Jang, and Kyungtae Lim. 2024. Bok-vqa: Bilingual outside knowledge-based visual question answering via graph representation pretraining. *arXiv preprint arXiv:2401.06443*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73.

Youhan Lee, KyungTae Lim, Woonhyuk Baek, Byungseok Roh, and Saehoon Kim. 2022. Efficient multilingual multi-modal pre-training through triple contrastive loss. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5730–5744, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. 2023a. Otter: A multi-modal model with in-context instruction tuning.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*.

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*.

Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng Yang, Jingjing Xu, Xu Sun, Lingpeng Kong, and Qi Liu. 2023c. M^3it: A large-scale dataset towards multi-modal multilingual instruction tuning. *arXiv preprint arXiv:2306.04387*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham. Springer International Publishing.

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

OpenAI. 2023. Gpt-4 technical report.

Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Jiyoon Han, Jangwon Park, Chisung Song, Junseong Kim, Yongsook Song, Taehwan Oh, et al. 2021. Klue: Korean language understanding evaluation. *arXiv preprint arXiv:2105.09680*.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

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. In *International conference on machine learning*, pages 8748–8763. PMLR.

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*.

Junke Wang, Lingchen Meng, Zejia Weng, Bo He, Zuxuan Wu, and Yu-Gang Jiang. 2023. To see is to believe: Prompting gpt-4v for better visual instruction tuning. *arXiv preprint arXiv:2311.07574*.

Kun Zhou Jinpeng Wang Wayne Xin Zhao Yifan Li, Yifan Du and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zihang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. 2021. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 538–547. IEEE.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2023. Vision-language models for vision tasks: A survey. *arXiv preprint arXiv:2304.00685*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2023. Lima: Less is more for alignment. *arXiv preprint arXiv:2305.11206*.

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