What Is Missing in Multilingual Visual Reasoning and How to Fix It

Anonymous ACL submission

Abstract

NLP models today strive for supporting multiple languages and modalities, improving accessibility for diverse users. In this paper, we evaluate their multilingual, multimodal capabilities by testing on a visual reasoning task. We observe that proprietary systems like GPT-4V obtain the best performance on this task now, but open models lag in comparison. Surprisingly, GPT-4V exhibits similar performance between English and other languages, indicating the potential for equitable system development across languages. Our analysis on model failures reveals three key aspects that make this task challenging: *multilinguality*, *complex reasoning*, and *multimodality*. To address these challenges, we propose three targeted interventions including a translate-test approach to tackle *multilinguality*, a visual programming approach to break down *complex reasoning*, and a method that leverages image captioning to address *multimodality*. Our interventions achieve the *best* open performance on this task in a *zero-shot* setting, boosting open models LLaVA-v1.5-13B by 13.4%, LLaVA-v1.6-34B by 20.3%, and Qwen-VL by 16.7%, while also minorly improving GPT-4V's performance.[1]

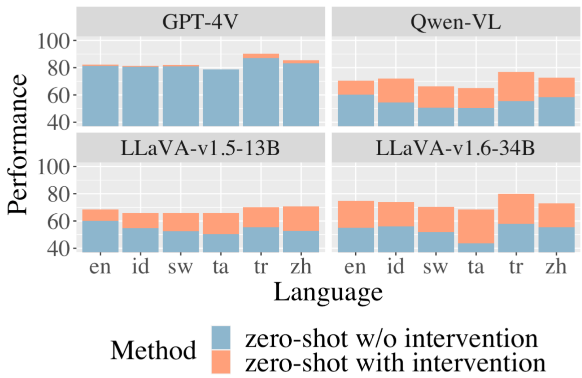

Figure 1: *Our Contributions*: First, we evaluate the multilingual visual reasoning abilities of various models; then, we analyze key challenges where models are falling short; lastly, we propose three interventions to address these challenges.

1 Introduction

Language technology today is continually evolving to be more inclusive, with models becoming increasing multilingual (Lai et al., 2023; Li et al., 2022), multimodal (Yang et al., 2023), or both (Chen et al., 2020; Zeng et al., 2023; Geigle et al., 2023; Achiam et al., 2023). Even though this promotes broader user accessibility, past research has consistently highlighted differences in model performance across languages (Blasi et al., 2022) and cultures (Liu et al., 2021). Notably, these models often favor North American or Western contexts, potentially leaving behind users from other regions. (Liu et al., 2021; Hershcovich et al., 2022).

The NLP community is currently witnessing a trend of moving away from openly releasing models to limiting their access through paid web APIs (Abdalla et al., 2023). Additionally, the cost to use these services is often higher for low-resourced languages, despite poorer performance (Ahia et al., 2023). While it is certainly desirable to have strong and inclusive models available regardless of the access method, open, well-documented, and reasonably sized models have advantages from the point of view of control, ownership, cost, and advancing scientific understanding.

In this work, we first compare and contrast the multilingual, multicultural capabilities of the proprietary system GPT-4V(ision) (Achiam et al., 2023) with a plethora of open models like LLaVA (Liu et al., 2023c,a, 2024), Qwen-VL (Bai et al., 2023b), mBLIP (Geigle et al., 2023), CCLM (Zeng et al., 2023), using two datasets on the same task of reasoning over texts and pairs of images, NLVR2 (Suhr et al., 2019) and MaRVL (Liu et al., 2021).

[1] An anonymized version of the code implementations and prompts can be found at `https://anonymous.4open.science/r/Multilingual_Visual_Reasoning-C1F7`.

We discuss this setup in more details in §2 and §3. We find that GPT-4V significantly outperforms all open models. One additional unprecedented and surprising result is, as shown in Figure 1, GPT-4V's consistency in performance across all languages, with performance on some even surpassing that on the NLVR2 dataset in English. In contrast, as we will discuss in §4, most open models still show a significant gap between English and other languages, perhaps due to deficiencies in training data, or due to the well-known "curse of multilinguality", where smaller models are less adept at processing low-resource languages (Conneau et al., 2020). This begs the question: "how can we take open models, and bring them closer to achieving the exciting language-equitable multimodal reasoning results demonstrated by the opaque (and presumably bigger) GPT-4V?"

Towards this end, we conduct a careful analysis of the results from testing models on the multilingual visual reasoning task and discover that failures can arise from any of the three challenging aspects of the task: *multilinguality*, *reasoning*, and *multimodality*. For *multilinguality*, we find a significant gap in performance for other languages as compared to English. For *reasoning*, we find a negative correlation of performance and the compositionality of the statement. For *multimodality*, we find that models were typically pretrained on single image-text pairs, but haven't seen pairs of images in pretraining, which may lead to a mismatch between pretraining and evaluation objectives. We will discuss this in more details in §5.

Based on our analysis, we design three interventions that address these challenges in section 6. The first simply tackles *multilinguality* – we translate the MaRVL statements to English. Surprisingly, translation leads to a drop in performance for GPT-4V (which might indicate its advanced multilingual capabilities), but helps improve performance for all open models. Our next intervention tackles both *multilinguality and reasoning*, by generating programs to reason over the set of images using the translated statements, inspired by Gupta and Kembhavi (2023)'s VisProg method. Our third (and most effective) intervention tackles *all three* challenges by first captioning images conditioned on the statement, and then reasoning over the captions, rather than the images, using chain-of-thought capabilities of text-modality LLMs (Wei et al., 2022). Using this intervention, we obtain state-of-the-art zero-shot performance on the MaRVL dataset for open models, and also slightly improve the performance of GPT-4V itself, as shown in Figure 1.

Statement: இரண்டு படத்திலும் வெண்டைக்காய் செடியில் உள்ளது. (***Translation***: In both pictures there is a marrow plant.) ***Label***: True

Figure 2: Example from the MaRVL Dataset: Given two images and a statement, the task is to infer whether the statement is true or false of the given image pair.

2 Dataset Description

We evaluate on two visual reasoning datasets, both containing a statement in natural language and a pair of images. The task is to reason whether the statement is true based on the images, requiring reasoning over both images and the statement together. Figure 2 shows an example of this task.

NLVR2 NLVR2 contains 107,292 examples of English sentences with web photographs. Annotators paired visually-rich images and were encouraged to come up with compositional and linguistically diverse statements for each pair. The dataset contains a train-validation-test split. Images were collected using search queries generated from synsets derived from the ILSVRC2014 ImageNet challenge (Russakovsky et al., 2015), with each query resulting in 4 pairs of images from Google Images[2]. Queries for ImageNet (Deng et al., 2009) are based on the English WordNet (Poli et al., 2010), whose concepts are more reflective of the North-American or Western cultures.

MaRVL MaRVL explores the same task as NLVR2 in multilingual multicultural contexts. MaRVL is a test-only dataset collected for five diverse languages: Indonesian, Swahili, Tamil, Turkish, and Mandarin Chinese. Native speakers first select concepts that are reflective of their speaking population. Next, they curate images from the web that reflect those concepts within their specific cultural context. Finally, native speakers pair and write statements for each image pair, following the same protocol as that laid out for NLVR2.

[2] https://images.google.com/

Model	NLVR2-en	id	sw	ta	tr	zh	MaRVL-Avg.	MaRVL-Avg. - EN
Human	96.2	96.3	93.0	98.0	97.0	95.5	96.0	-0.2
Zero-Shot								
GPT-4V	81.4	**80.6**	**81.0**	**78.6**	**87.1**	**83.2**	**82.1**	**0.7**
mBLIP (mT0-XL)	67.3	64.9	64.8	69.6	68.0	65.9	66.6	-0.7
LLaVA-v1.5-13B	60.1	54.8	52.6	50.2	55.3	52.9	53.2	-6.9
LLaVA-v1.6-34B	54.9	56.0	51.8	43.4	57.9	55.3	52.9	-2.0
Qwen-VL	60.3	54.5	50.7	50.3	55.4	58.4	53.9	-6.4
Finetuned								
mBLIP (mT0-XL)	**85.2**	75.1	74.6	75.9	74.3	75.7	75.1	-10.1
CCLM-4M	80.2	67.6	64.4	60.5	69.0	69.2	66.1	-14.1
xUNITER	72.3	57.7	56.1	54.3	57.6	54.7	56.1	-16.2
mUNITER	73.2	55.0	51.5	52.2	54.7	56.8	54.0	-19.2

Table 1: NLVR2 and MaRVL performance across Human , Proprietary Models , and Open Models. On NLVR2 , mBLIP outperforms GPT-4V post finetuning, while GPT-4V shows the best performance across other languages.

3 Models and Evaluation Protocols

We evaluate various open models, including mBLIP (mt0-xl) (Geigle et al., 2023), LLaVA (Liu et al., 2023a, 2024), Qwen-VL (Bai et al., 2023b), CCLM (Zeng et al., 2023), and UNITERs (Chen et al., 2020); and a proprietary model GPT-4V(ision).[3] We describe these models in §A. We evaluate them in two settings:

Zero-shot. In this setting, models are not specifically fine-tuned for the task of visual reasoning. This setting is academically interesting, as it more generally tests the ability of models to perform tasks, and the results are more likely to be representative of performance on datasets for which training data is not available. In addition, it is practically useful since it can also be applied to LMs that cannot as easily be fine-tuned, such as GPT-4V (due to its closed nature), LLaVA, and Qwen-VL (due to their relatively large sizes). We test LLaVA, Qwen-VL, mBLIP, and GPT-4V in this setting.

Finetuned. We finetune models that can more easily be finetuned on the English NLVR2 dataset, and test on NLVR2 and MaRVL. This represents the realistic setting, adapting multilingual models to particular tasks using English data, which is relatively available. We test mBLIP, CCLM-4M, xUNITER, and mUNITER in this setting.

[3] *gpt-4-vision-preview* (`https://openai.com/research/gpt-4v-system-card`), abbreviated as "*GPT-4V*".

4 How well do proprietary and open models perform on multilingual visual reasoning?

In this section, we perform an examination of how-well these various models perform on multilingual multimodal reasoning tasks. Table 1 shows performance of humans, open models, and proprietary models. For the models, we use the experiment protocols as in §3 in the zero-shot and finetuned settings. We ask the following questions:

Which model performs the best? *Answer:* GPT-4V on MaRVL, and mBLIP (mT0-XL) on English post-fintuning. However, in the zero-shot setting, the proprietary model GPT-4V performs the best across all languages,[4] and open models lag behind. Note that despite GPT-4V's impressive performance, it still lags behind human performance by 10% to 20% across all languages, emphasizing that this task still is not completely solved.

Which open model performs the best? *Answer:* mBLIP (mT0-XL), regardless of whether it is finetuned. The other open LMMs, LLaVA and Qwen-VL, are not explicitly trained on multilingual data, so the gap in MaRVL and NLVR2 performance is expected.

[4] We put GPT-4V in the zero-shot category because we evaluate the performance of GPT-4V on NLVR2 and MaRVL without finetuning on the NLVR2 training data. However, we do not know if GPT-4V has seen examples of NLVR2 or MaRVL during pretraining.

Do models perform equitably across languages? Under zero-shot setting, GPT-4V and mBLIP both show equitable performance across languages, which is encouraging, although the latter significantly lags in overall performance compared to GPT-4V. Interestingly, post finetuning on NLVR2, mBLIP shows better performance on NLVR2 than GPT-4V, but still has lower performance on MaRVL. The gap between English and MaRVL languages also significantly increases for mBLIP from the zero-shot to finetuned setting. Maintaining the equity in performance across languages during finetuning is an interesting future direction, which may help models surpass GPT-4V's performance on multilingual visual reasoning. Other models lag mBLIP, both in overall performance and equity with English.

5 What makes multilingual visual reasoning challenging?

As noted in Table 1, the best model still lags human performance by 10% to 20%. In this section, we aim to analyze what makes multilingual visual reasoning so challenging, and identify three key aspects as detailed below:

5.1 Multilinguality and Sub-Optimal Cross-Lingual Transfer

In the finetuned setting, we observe a significant drop in performance for MaRVL languages as compared to NLVR2 in English. This is expected since models are finetuned only in English but not in these languages due to lack of training data. We also note that performance on Swahili is consistently lower across models (excluding GPT-4V), which is the lowest-resource language amongst MaRVL languages, as laid out by the language resource taxonomy (Joshi et al., 2020). This observation motivates us to evaluate models with MaRVL data translated to English, as we discuss in §6.1.

In the zero-shot setting, GPT-4V and mBLIP both exhibit equitable performance on MaRVL as with NLVR2. While LLaVa is not expected to perform as well for non-English languages and Qwen-VL is not expected to perform as well for non-English and non-Chinese languages, they have poorer performance than mBLIP on NLVR2. While mBLIP is pretrained on multilingual multimodal data, LLaVA is not specifically trained on multilingual data. However Qwen-VL is pretrained on Chinese data (Bai et al., 2023b), and it is generally believed that LLaVA has multilingual abilities as it has seen multilingual data during pretraining (Liu et al., 2023c,a, 2024).

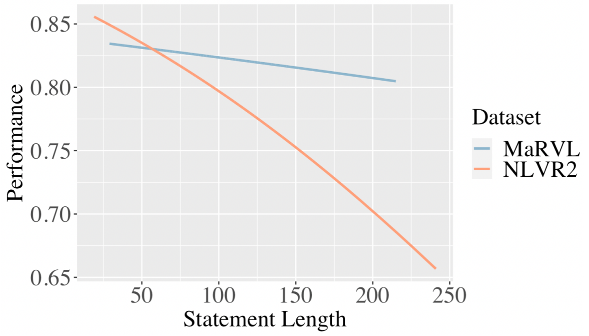

Figure 3: Performance of GPT-4V decreases as statement length increases.

Overall, models have better visual reasoning abilities when given English inputs from US/European-centric cultures, while still lagging behind when facing multilingual and multicultural inputs.

5.2 Complex Reasoning

Data points in both NLVR2 and MaRVL require complex reasoning. An example statement from NLVR2 is "one image includes a silver stylus and a device with a blue keyboard base and an open screen propped up like an easel", which is semantically diverse, long in length, and has a compositional structure, requiring models to perform compositional and complex reasoning to infer the label.

As a proxy to the complexity of reasoning, we measure the number of words of the NLVR2 and MaRVL statements (translated to English), and find that model performances drop as the number of words of the statement increases. Figure 3 shows a graph of the performance of GPT-4V plotted against the number of words in each statement. We can clearly see a downward trend in the graph. Based on this, we are motivated to examine methods that break down long, compositional statements, as will be discussed in §6.2.

5.3 Multimodality and Mismatch between Pretraining & Evaluation

NLVR2 and MaRVL contain two images per instance, along with a statement describing them, while vision-language models are typically trained on a single image-text pair (Cao et al., 2020), leading to a mismatch in the input between pretraining and evaluation. Further, multimodal reasoning is known to be harder than reasoning over text alone

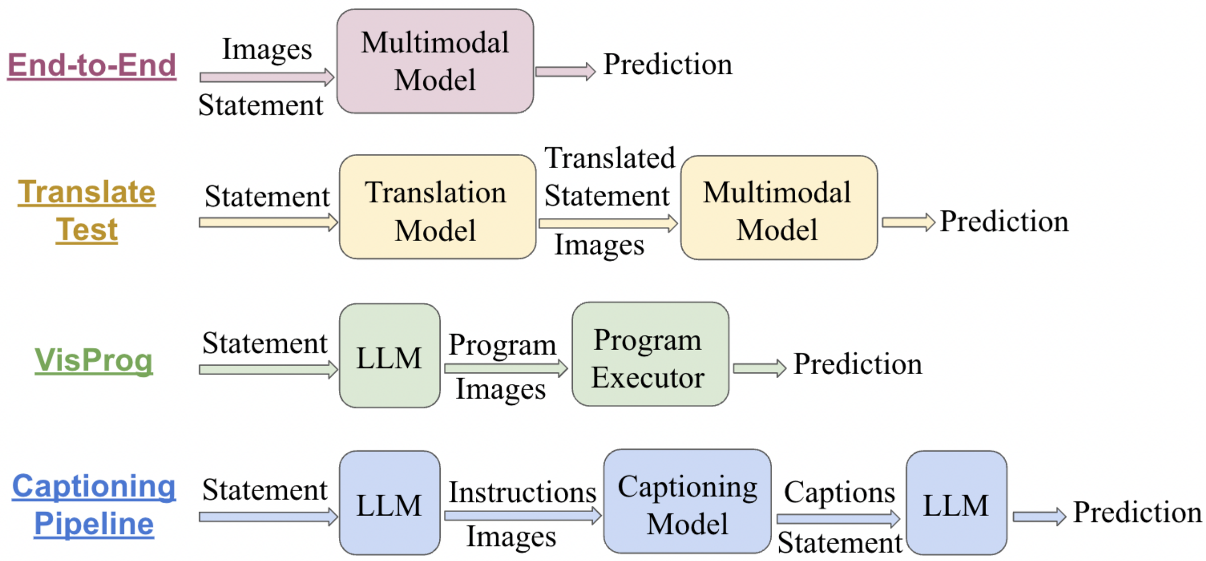

Figure 4: Flow chart visualizing the end-to-end testing in §4 and all interventions performed in §6.

Model	NLVR2-en	id	sw	ta	tr	zh	MaRVL-Avg.	MaRVL-Avg. - EN
				Zero-Shot				
GPT-4V	81.4	78.4	75.5	70.2	78.2	78.4	76.1	-5.3
mBLIP (mT0-XL)	67.3	65.2	66.9	68.7	68.2	67.2	67.2	-0.1
LLaVA-v1.5-13B	60.1	53.1	53.9	54.1	58.3	54.0	54.7	-5.4
LLaVA-v1.6-34B	54.9	55.7	53.1	52.8	55.3	55.4	54.5	-0.4
Qwen-VL	60.3	58.2	56.0	58.8	63.0	58.4	58.9	-1.42
				Finetuned				
mBLIP (mT0-XL)	85.2	77.2	78.6	78.3	73.1	75.1	76.5	-8.7
CCLM-4M	80.2	72.3	69.2	69.7	77.6	71.8	72.1	-8.1
xUNITER	72.3	63.2	63.8	62.1	67.5	62.1	63.7	-8.6
mUNITER	73.2	59.8	63.4	62.3	69.2	62.7	63.5	-9.7
ViLT	73.7	61.7	62.0	65.1	69.8	60.9	63.9	9.8

Table 2: MaRVL translate-test accuracies across Open and Proprietary models.

(Mogadala et al., 2021; Park and Kim, 2023). Although Qwen-VL has seen multi-image inputs during training (Bai et al., 2023b), it still encounters difficulties in handling the complexities presented by multimodal reasoning during evaluation.

These, and the inherent difficulty of aligning image data and text data during the reasoning process make this task particularly challenging. This motivates us to (1) move from processing a pair of images together to processing each image separately; and (2) break down the two modalities of image and text in the reasoning process, as in §6.3.

6 How can we address these challenges?

Based on our analysis from the previous section, we now move on to examining whether we can devise methods to further improve multilingual multimodal reasoning abilities, particularly those of open models. We examine three research questions, which we discuss in more details in the following subsections respectively. Figure 4 shows a flow chart visualizing the interventions we perform to address the research questions[5].

RQ1) (*multilinguality*) Does translating the text to English and reducing the cross-lingual gap aid performance? ***Short Answer***: it depends.

RQ2) (*multilinguality+reasoning*) Can we break down the complex reasoning into modular programs which can be executed on a vision-text input? ***Short Answer***: yes, we adopt the Visual Programming approach (Gupta and Kembhavi, 2023).

RQ3) (*multilinguality+reasoning+multimodality*) Can we alleviate the need for multimodal interaction during the reasoning process? ***Short Answer***: yes, we propose a new approach utilizing captions.

6.1 *Addressing Multilinguality:* Translate-Test

In §5.1, we find performance on NLVR2 is much better than that on MaRVL. While both are visual reasoning datasets, MaRVL is multi-cultural and contains statements in 5 diverse languages. Since NLP systems perform significantly better with English data (Song et al., 2023), we first simply translate the reasoning statements to English using the Google Translate API (Wu et al., 2016). A visualization of the process of translate test can be found in Figure 4.

In addition to the models we evaluate in §3, we also evaluate ViLT (Kim et al., 2021) for better comparisons, as our next proposed intervention in §6.2 uses ViLT. We didn't evaluate ViLT on MaRVL before translate test, since it doesn't support the MaRVL languages. Our evaluation protocols follows the ones introduced in §3 and results are shown in Table 2.

All prior works, as per our knowledge, have observed a gain in performance post translating to English (Liu et al., 2021). Our observation is consistent with prior findings for all models, except GPT-4V(ision). All models except for GPT-4V sees an increase in accuracy after performing translate test; while surprisingly, GPT-4V shows a sharp decrease in performance across all MaRVL languages after translate test. However, this is encouraging, because it speaks for the multilingual capabilities of this model, and indicates that the gains provided by translating to English are lower than the errors made in translating cultural-specific nuances in meaning.

For example, the MaRVL statement "右图有青绿色的苹果" is translated to "the picture on the right has turquoise apples", where "青绿色" is translated to "turquoise". However, the color "青绿色" means pure green with a little bit cyan in Mandarin Chinese, which is different from "turquoise". Given this, GPT-4V reasons correctly when provided the statement in Mandarin, but makes mistakes when given the translated statement.

6.2 Addressing *Multilinguality* + *Reasoning:* Visual Programming

To improve performance of LLMs on reasoning tasks, beyond naive prompting, several methods have been introduced (Nye et al., 2021; Zhou et al., 2022; Wei et al., 2022; Gao et al., 2023). Particularly, PAL (Gao et al., 2023) provides significant improvements by decomposing a natural language instruction into multiple programmatic submodules, executed in an inference step to obtain the final answer. Most recently, efforts like VisProg (Gupta and Kembhavi, 2023), ViperGPT (Surís et al., 2023), Visual ChatGPT (Wu et al., 2023) have followed suit to solve multimodal reasoning using LLMs to generate *visual* programs, that leverage off-the-shelf computer vision models for image processing during inference. Hence, we use VisProg to generate visual programs given translated statements as obtained in §6.1. VisProg uses ViLT (Kim et al., 2021) as its inherent vision module.

Figure 4 shows the flow of VisProg. For example, given the statement: *There is no one in the bedroom on the left, and there is someone in the bedroom on the right*, the generated visual program is:

```
ANSWER0=VQA(image=LEFT,question='Is
    there anyone in the bedroom?')
ANSWER1=VQA(image=RIGHT,question='Is
    there anyone in the bedroom?')
ANSWER2=EVAL(ANSWER0 == False and
    ANSWER1 == True)
FINAL_ANSWER=RESULT(var=ANSWER2)
```

Listing 1: Visual program example

If this program is executed on the images in Figure 5, then it will have ANSWER0 $= True$, ANSWER1 $= False$, so the final result is $False$.

Figure 5: VisProg example image pair.

For this intervention, we use text-davinci-003[6] as

[5] §C discusses additional computation cost incurred by the interventions.

[6] text-davinci-003 is the model that the VisProg authors utilized when running VisProg.

a representative of proprietary LLMs and LLaMA2-70B (Touvron et al., 2023) to represent open LLMs. Table 3 shows results to this method. Although this method does not achieve as high accuracy as models evaluated end-to-end in Table 1, this approach provides valuable insights of breaking down complex reasoning into modular modules and utilizing prompts to make use of LLMs' strong in-context abilities. In addition, this approach, without any additional training, achieves on par performance on MaRVL, as compared to ViLT post-fintuning.

Model	NLVR	MaRVL					
		id	sw	ta	tr	zh	Avg.
GPT-3	67.0	64.5	59.8	60.3	67.3	64.3	63.2
LLaMA2-70b	67.3	58.2	57.2	58.1	65.8	61.9	60.2

Table 3: VisProg performance across models.

6.3 Addressing *Multilinguality* + *Reasoning* + *Multimodality*: Reasoning with Captions

When analyzing errors for NLVR2, Gupta and Kembhavi (2023) note that 69% of them are caused by the vision module. This might be potentially worse for MaRVL, because open visual modules used in VisProg (Kim et al., 2021) are trained on Western-centric datasets like Imagenet (Russakovsky et al., 2015). Text-based LLMs, on the other hand, are trained on trillions of tokens, and are known to exhibit cultural awareness, albeit it may be limited (Yao et al., 2023). Hence, here we target the last remaining challenge, that of multi-modal interaction needed for the reasoning process, by working with image captions instead of images. Concretely, we first caption both images, and use LLMs to reason about the statement with the two captions, instead of with the two images. Figure 4 shows a flow chart of how this pipeline works.

To make sure the captions capture necessary information needed for reasoning about the statement, as a first step of this intervention we use LLMs to generate targeted instructions based on the statement. Consider the statement "*The picture on the left has several pencils of different colors, and the picture on the right has only one pencil*" from MaRVL-zh, the targeted instructions are:
Left image - "*Write a short caption describing the number and colors of pencils*;"
Right image - "*Write a short caption describing the number of pencils*".

As a second step, we generate captions following

Figure 6: Captioning example image pair.

the targeted instructions in step 1, using various captioning models, including InstructBLIP (Liu et al., 2023b), PromptCap (Hu et al., 2022), GPT-4V, LLaVA-v1.5-13B (Liu et al., 2023a), LLaVA-v1.6-34B (Liu et al., 2024), and Qwen-VL (Bai et al., 2023b). The instructions can point them to focus on targeted contents in the image. For instance, for the statement in step 1 and the images in Figure 6, the captions generated by GPT-4V are:
Left image - *A pencil case containing a single natural wood colored pencil.*;
Right image - *A young mind sketches ideas with a classic green pencil.*.

Lastly, we prompt LLMs to reason whether the statement match caption pairs. For instance, for the example above, GPT-4 reasons as follows:
Reasoning - *The statement is False. This is because the left image is described to include only a single pencil, which implies there are not several pencils of different colors in the left image. However, the statement claims that there should be several pencils of different colors in the left image. Meanwhile, the right image correctly shows a single green pencil, but the statement is negated by the inaccuracy of the part regarding the left image.*

Using this approach, visio-linguistic models only need to process one image at a time, instead of processing a pair of images simultaneously. In addition, reasoning is performed only over text, instead of over both image and text modalities.

Table 4 shows the performance of this intervention. Performance of InstructBLIP and PromptCap under zero-shot setting is on par with UNITER models post-finetuning. This intervention improves performance of LLaVA-v1.5-13B by 10% to 16%, LLaVA-v1.6-34B by 18% to 25%, and Qwen-VL[7] by 10% to 21% depending on language, while also minorly improving GPT-4V's performance. On average, our pipeline boosts LLaVA-v1.5-13B's performance on MaRVL by 13.4%, LLaVA-v1.6-

[7] §B discusses additional experiments on Qwen-VL.

Captioning	Reasoning	NLVR (en)	id	sw	ta	tr	zh	MaRVL-Avg.
InstructBLIP	LLaMA2-70B	65.1	61.3	60.8	60.2	62.6	62.8	61.5
PromptCap	LLaMA2-70B	63.2	59.3	58.9	58.3	59.2	59.9	59.1
GPT-4V	No Intervention	81.4	80.6	81.0	78.6	87.1	83.2	82.1
	GPT4	82.2	81.2	81.8	76.1	90.1	85.4	82.92
LLaVA-v1.5-13B	No Intervention	60.1	54.9	52.6	50.2	55.3	52.9	53.2
	LLaMA2-70B	68.6	65.8	65.9	65.8	69.9	70.8	67.6
LLaVA-v1.6-34B	No Intervention	54.9	56.0	51.8	43.4	57.9	55.3	52.9
	LLaMA2-70B	74.8	73.9	70.3	68.6	80.0	73.0	73.2
Qwen-VL	No Intervention	60.3	54.5	50.7	50.3	55.4	58.4	53.9
	LLaMA2-70B	70.3	72.1	66.3	65.1	76.7	72.8	70.6

Table 4: Captioning Pipeline Performance across Models. For rows with "No Intervention" stated in the "Reasoning" column, we pull over the end-to-end results of that model from Table 1, for the sake of comparison.

34B's performance by 20.3%, and Qwen-VL's performance by 16.7%. This intervention improves performance of LLaVA and Qwen-VL, achieving the best performance under zero-shot setting (without training on reasoning of pairs of images).

7 Related Work

From Pretraining to Instruction Tuning Previous research on instruction tuning sparks multiple works to finetune models on instructions, and create general-purpose models that are good at performing tasks under zero-shot settings (Ouyang et al., 2022; Liu et al., 2023b; Geigle et al., 2023). However, instruction tuning data is mostly in English (Touvron et al., 2023; Liu et al., 2023b). Due to the absence of multilingual instruction tuning data, models may struggle to effectively process multilingual inputs.

Moving Beyond English Past research efforts has predominantly centered around English language models, highlighting differences in model performance across languages (Blasi et al., 2022; Song et al., 2023). In the visio-linguistic domain, research in instruction tuning also center on English, due to a lack of multilingual instruction training data (Geigle et al., 2023). To this end, mBLIP (Geigle et al., 2023) translated instruction training data to various languages, and perform instruction tuning. This is the first multilingual instruction tuned vision LLM.

Gap between Proprietary GPT-4V and Open Models Currently, there is a trend of shifting from openly releasing models to paid APIs (Abdalla et al., 2023). Previous research on examining GPT-4V demonstrates its unprecedented multimodal capabilities, and there is still a gap between this proprietary model and other open source models (Yang et al., 2023). However, it is still important for the community to have as strong open source multimodal models.

8 Conclusion

In conclusion, we explore the evolving domain of multilingual visual reasoning. We observe a trend towards inclusivity in models, yet recognize persistent disparities in performance across languages and cultures. While proprietary systems like GPT-4V exhibit notable and equitable accuracy across languages, open models still face challenges in bridging the gap, especially for low-resource languages. Our analysis highlights the superior performance of GPT-4V but also underscores the need for advancements in open models. Leveraging interventions addressing multilinguality, multimodality, and reasoning, we demonstrate significant enhancements in open model performance, achieving state-of-the-art results under zero-shot settings for open models. Our findings emphasizes the potential for further advancements in multilingual visual reasoning, with the aim of narrowing down the gap between human and machine performance, and the gap between proprietary and open models.

Limitations

With the goal of evaluating the multilingual visual reasoning capabilities of models, we employ NLVR2 and MaRVL, both of which engage in the task of determining whether a pair of images correspond to a given statement. This choice stems from MaRVL being the sole visual reasoning dataset with multilingual support, as far as our current knowledge extends.

Representing Visual Reasoning It's important to acknowledge that the task of NLVR2 and MaRVL solely represents a specific task of visual reasoning. Other aspects and dimensions of this domain may not be fully represented by this particular task.

Representing Multilinguality In addition, note that the combination of NLVR2 and MaRVL covers 6 distinct languages: English, Indonesian, Swahili, Tamil, Turkish, and Mandarin Chinese. This is only a small subset of all languages worldwide.

References

Mohamed Abdalla, Jan Philip Wahle, Terry Ruas, Aurélie Névéol, Fanny Ducel, Saif Mohammad, and Karen Fort. 2023. The elephant in the room: Analyzing the presence of big tech in natural language processing research. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13141–13160, Toronto, Canada. Association for Computational Linguistics.

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*.

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David Mortensen, Noah Smith, and Yulia Tsvetkov. 2023. Do all languages cost the same? tokenization in the era of commercial language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9904–9923, Singapore. Association for Computational Linguistics.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023a. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

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

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

Jize Cao, Zhe Gan, Yu Cheng, Licheng Yu, Yen-Chun Chen, and Jingjing Liu. 2020. Behind the scene: Revealing the secrets of pre-trained vision-and-language models. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 565–580. Springer.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Pal: Program-aided language models. In *International Conference on Machine Learning*, pages 10764–10799. PMLR.

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

Tanmay Gupta and Aniruddha Kembhavi. 2023. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14953–14962.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Yushi Hu, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A Smith, and Jiebo Luo. 2022. Promptcap: Prompt-guided task-aware image captioning. *arXiv preprint arXiv:2211.09699*.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR.

Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. *arXiv preprint arXiv:2304.05613*.

Junyi Li, Tianyi Tang, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2022. Pretrained language models for text generation: A survey. *arXiv preprint arXiv:2201.05273*.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

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

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

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

Aditya Mogadala, Marimuthu Kalimuthu, and Dietrich Klakow. 2021. Trends in integration of vision and language research: A survey of tasks, datasets, and methods. *Journal of Artificial Intelligence Research*, 71:1183–1317.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, et al. 2021. Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Sang-Min Park and Young-Gab Kim. 2023. Visual language integration: A survey and open challenges. *Computer Science Review*, 48:100548.

Roberto Poli, Michael Healy, and Achilles Kameas. 2010. *Theory and applications of ontology: Computer applications*. Springer.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252.

Yueqi Song, Simran Khanuja, Pengfei Liu, Fahim Faisal, Alissa Ostapenko, Genta Winata, Alham Aji, Samuel Cahyawijaya, Yulia Tsvetkov, Antonios Anastasopoulos, and Graham Neubig. 2023. GlobalBench: A benchmark for global progress in natural language processing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14157–14171, Singapore. Association for Computational Linguistics.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428, Florence, Italy. Association for Computational Linguistics.

Dídac Surís, Sachit Menon, and Carl Vondrick. 2023. Vipergpt: Visual inference via python execution for reasoning. *Proceedings of IEEE International Conference on Computer Vision (ICCV)*.

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*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. 2023. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. The dawn of lmms: Preliminary explorations with gpt-4v(ision). *arXiv*, abs/2309.17421.

Binwei Yao, Ming Jiang, Diyi Yang, and Junjie Hu. 2023. Empowering llm-based machine translation with cultural awareness. *arXiv preprint arXiv:2305.14328*.

Yan Zeng, Wangchunshu Zhou, Ao Luo, Ziming Cheng, and Xinsong Zhang. 2023. Cross-view language modeling: Towards unified cross-lingual cross-modal pre-training. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5731–5746, Toronto, Canada. Association for Computational Linguistics.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. 2022. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*.

A Models and Evaluation Protocols

In this section, we introduce all multimodal models that we evaluate in Section 4.

A.1 Open Models

A.1.1 Zero-Shot Evaluation (*no labeled data for task*)

Recently, there has been a rise in multimodal language models that are instruction-finetuned to solve tasks in a zero-shot manner (Chung et al., 2022). These systems may or may not be trained multilingually. We evaluate these models by providing the models with instructions on solving the task, utilizing the models' zero-shot learning abilities and chain-of-thought reasoning abilities (Wei et al., 2022). Below, we briefly describe the models that we experiment with under a zero-shot setting:

mBLIP mBLIP (Geigle et al., 2023) extends large multimodal models' capabilities to be multilingual. mBLIP re-align an image encoder previously tuned to an English LLM to a multilingual LLM. Re-alignment training of mBLIP utilizes multilingual data machine-translated from English data.

LLaVA Large Language and Vision Assistant (LLaVA) is a series of open large multimodal model that are instruction tuned on machine-generated instruction-following data (Liu et al., 2023c,a, 2024). LLaVA extends the capabilities of existing models by incorporating visual models and large language models. It connects a vision encoder CLIP and an LLM decoder. LLaVA is not explicitly trained to process multilingual data, but the LLM decoder (Vicuna is the default LLM) is known to have seen multilingual data in pretraining (Chiang et al., 2023).

Qwen-VL Qwen-VL is an open large multilingual multimodal model trained on English and Chinese data. It is based on Qwen-7B (Bai et al., 2023a), incorporating a language-aligned visual encoder and a positionaware adapter. It is trained to be able to process multi-image inputs.

A.1.2 Evaluation Post-Finetuning on NLVR2 *(labeled data for task in English)*

Several end-to-end encoder-based models have been proposed that are pretrained on multilingual multimodal data, and typically need to be fintuned prior to evaluation (Devlin et al., 2018). Pretraining objectives typically include masked language modeling (text), image-text matching, masked region modeling (image), and multimodal contrastive learning (Chen et al., 2020; Zeng et al., 2023).

To test on MaRVL, they need to be finetuned on task-specific data. Since MaRVL is a test-only dataset, we finetune on the training data of NLVR2 which is only in English. Note that these models are pretrained on a single image-text pair. To deal with a pair of images in finetuning, each image is separately paired with the statement in two forward passes, and a concatenation of obtained embeddings is passed to a linear classifier to make the prediction. Here, we experiment with CCLM and UNITER-based models as described below. We also finetune mBLIP, but not LLaVa, due to computational constraints introduced by its size.

UNITER The UNiversal Image-TExt Representation Learning (UNITERs) framework focuses on achieving end-to-end reasoning across different modalities (Chen et al., 2020). This model aims to unify the processing of textual and visual information, fostering more coherent and integrated reasoning capabilities. We experiment with mUNITER and xUNITER, which are initialized from UNITER with mBERT and XLM-R respectively.

CCLM The Crosslingual Cross-modal Language Model (CCLM) is an open pretrained multilingual multimodal that delves into conditional masked language modeling and contrastive learning techniques to enhance cross-modal understanding (Zeng et al., 2023). This model contribute valuable insights into improving the alignment between textual and visual representations in multilingual scenarios.

A.2 Proprietary Model GPT-4V

GPT-4V(ision) Incorporating multimodality into GPT-4, GPT-4V is able to process image inputs and text inputs together, paving the way for various downstream tasks including visual reasoning tasks (Achiam et al., 2023; Yang et al., 2023). Since GPT-4V is also know for its zero-shot learning abilities (Yang et al., 2023), plus finetuning is not supported by GPT-4V[8], we evaluate GPT-4V under a zero-shot setting as discussed in §A.1.1.

B Additional Experiments on Qwen-VL

To better understand multilingual and multicultural understanding abilities of our proposed pipeline, we performed additional experiments on Qwen-VL. This is because Qwen-VL is trained on Chinese data, while all other open models we evaluated are pretrained with a focus on English culture, without seeing much data from the local culture. Therefore, in addition to the experiments we discussed in Section 6.3, we also performed the third intervention with Qwen-VL on the MaRVL Mandarin Chinese dataset where we caption images using the native language. This experiment resulted in 73.4% accuracy, while using our interventions with English captions gives 72.8% accuracy, and using Qwen without interventions gives 58.4% accuracy. These results extended our points that visio-linguistic models need better understanding of culturally-specific elements. For example, Siheyuan is a culturally specific concept from Chinese culture, where if a model has never seen such concepts previously, it might not be able to generate the correct response for queries containing the concept Siheyuan.

C Additional Computation Cost

For the first intervention in §6.1, we use the translated statements provided in the MaRVL dataset, so no additional training cost is incurred.

For the second intervention in §6.2, training cost is not directly comparable, since we finetune ViLT if not using the intervention, and use the pretrained ViLT if using the intervention.

For the third intervention, with a 3% increase in total evaluation time, we see a 13% average improvement in performance for LLaVA-v1.5-13B. There is no additional training cost brought by the intervention. Noteworthily, total inference time using LLaVA is halved when using this intervention.

[8] https://platform.openai.com/docs/guides/fine-tuning/what-models-can-be-fine-tuned