Why do LLaVA Vision-Language Models Reply to Images in English?

Musashi Hinck[*1],
Carolin Holtermann[†2] **Matthew Lyle Olson**[†1], **Florian Schneider**[†2]
Sungduk Yu[1], **Anahita Bhiwandiwalla**[1],
Anne Lauscher[2] **Shaoyen Tseng**[1], **Vasudev Lal**[1],
[1]Intel Labs, [2]University of Hamburg,
Correspondence: musashi.hinck@intel.com

Abstract

We uncover a surprising multilingual bias occurring in a popular class of multimodal vision-language models (VLMs). Including an image in the query to a LLaVA-style VLM significantly increases the likelihood of the model returning an English response, regardless of the language of the query. This paper investigates the causes of this loss with a two-pronged approach that combines extensive ablation of the design space with a mechanistic analysis of the models' internal representations of image and text inputs. Both approaches indicate that the issue stems in the language modeling component of the LLaVA model. Statistically, we find that switching the language backbone for a bilingual language model has the strongest effect on reducing this error. Mechanistically, we provide compelling evidence that visual inputs are not mapped to a similar space as text ones, and that intervening on intermediary attention layers can reduce this bias. Our findings provide important insights to researchers and engineers seeking to understand the crossover between multimodal and multilingual spaces, and contribute to the goal of developing capable and inclusive VLMs for non-English contexts.

1 Introduction

Language fidelity in large language models (LLMs) refers to whether the model replies in the same language as it was queried in. While seemingly a simple task for humans, models with multilingual capabilities will often bias towards English replies, especially for queries in low-resource languages (Holtermann et al., 2024). In this work, we identify a surprising parallel pathology in LLaVA-style VLMs (Liu et al., 2023): when prompted with a multimodal query that includes an image, the model is more likely to reply in an incorrect language with respect to the query language. We term this Image-induced Fidelity Loss (IFL).

In the first part of the paper, we define this phenomenon and demonstrate the *extent* of this problem in existing LLaVA-style models. Using a collection of 7740 evaluation tasks drawn from three VQA benchmarks and spanning fourteen languages, we show empirically that adding an image to the query to a LLaVA model causes the probability of the response being in the correct language to decrease between 6% and 53%.

We analyze the cause of IFL through two complementary approaches. First we conduct a macro-level analysis by training numerous LLaVA-style models by systematically ablating the design space. This allows us to statistically estimate the impact of each design choice on the model's propensity for generating linguistically misaligned responses. Second, we provide a micro-level investigation of the model's intermediary representations. By studying the clustering patterns of vision tokens in relation to tokens from different languages, and directly intervening on hidden states within the language transformer layers, we gain insights into the internal dynamics that give rise to the observed phenomenon. The synthesis of these two levels of analysis – exploring the design space at a statistical level and probing the model's internal representations – provides a comprehensive characterization of the factors influencing cross-lingual response generation in multimodal models.

We find strong evidence that IFL occurs due to the language modelling component of the LLaVA model. Our statistical analysis indicates that changing the language backbone from an English-main LLM to a bilingual LLM has the strongest effect on reducing IFL. In our mechanistic analysis we find that visual inputs are not mapped to a similar space as text inputs, and that intervening on intermediary attention layers can reduce IFL.

In sum, our contributions are as follows:

[*]First author.
[†]Equal contribution, ordered alphabetically.

Findings of the Association for Computational Linguistics: EACL 2024, pages 13402–13421
November 12-16, 2024

- We systematically demonstrate IFL on multiple tasks over a wide range of languages for modern VLMs.
- We conduct a large scale analysis of the design choices in training VLMs, enabling robust statistical results for interpreting training architecture decisions.
- We perform detailed representation analysis to find the IFL problem is localized within the language model.
- We provide evidence that the key to reducing IFL for LLaVA-style models is in the language backbone of LLaVA.

2 Related Work

2.1 Global Inclusion via Multilingual Models

Aiming towards globally inclusive language technologies, much of natural language processing research has focused on multilingual and cross-lingual models (Conneau et al., 2018). Starting from static embedding models (cf., Ruder et al., 2019) and smaller multilingual pretrained transformer models like mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020), much of multilingual NLP has shifted to multilinguality in instruction-tuned LLMs. Here, researchers either focus on training multilingual LLMs (Lai et al., 2023), or, given that explicitly open multilingual chat models are still rare, aim towards understanding multilinguality of models intended for English use only (Blevins and Zettlemoyer, 2022). Moreover, the vast adoption of instruction-tuned models has transitioned their focus from task-specific responses to generating more natural language outputs. This allows models to answer more flexibly, unbound by rigid response frameworks, and autonomously choose the language of their responses. Despite this advancement, existing benchmarks primarily evaluate the accuracy of answers without adequately assessing the models' fidelity to the language used (Holtermann et al., 2024). However, given the multilingual use of these models (Zhao et al., 2024), it is crucial to ensure they can respond accurately in the appropriate language to foster inclusivity.

2.2 Efficient Integration for Multimodal Understanding

Humans interact with the world through multiple channels. Accordingly, many AI researchers explored how to integrate multiple modalities, particularly vision and language, into a single model (e.g., Kim et al., 2021; Wang et al., 2022). As an alternative to efforts that focused on models specific to particular tasks (Brooks et al., 2023), general-purpose vision-language models emerged. Given that pretraining larger models (as in Kim et al., 2021; Radford et al., 2021, *inter alia*) became prohibitively expensive, researchers moved to employing readily available encoders, keeping those (partially) frozen, and mapping them through learned projections (Li et al., 2023; Merullo et al., 2023). For instance, Mañas et al. (2023) employ a transformer mapping network, and Eichenberg et al. (2022) rely on adapters. In this work, we focus on efficient and thus, accessible methods for the integration of instruction-tuned LLMs, with LLaVA as a popular representative (Liu et al., 2023).

2.3 Multimodal Understanding and Multilinguality

Our research shares motivations with existing research in multimodal multilingual models; extending advances in multimodality to other languages (Elliott et al., 2016) and increasing global inclusivity through a broader geographical lens on language and vision (Liu et al., 2021). The cross-modal, cross-lingual setting brings its own set of challenges for researchers and engineers, such as the suitability of machine translation for training data (Qiu et al., 2022).

In the present moment, the popularity and resource efficiency of LLaVA provides an opportunity to conduct research on training decisions that is both comprehensive and widely relevant. Already several researchers have used LLaVA to extend vision-language capability to multilingual settings (Andersland, 2024; Shin et al., 2024; Song et al., 2024). In this context, our research aims to extend the field's understanding of the links between multilinguality and multimodality, and to draw attention to the specific challenges that emerge in this setting.

3 Image-induced Fidelity Loss

The phenomenon of interest in this study is the change in fidelity seen when adding an image to the input to a VLM. We refer to this as *image-induced fidelity loss*. We argue that this phenomenon is surprising, given that visual inputs should be language-agnostic (barring the cultural-linguistic context as-

sociated with the image) and therefore orthogonal to the language of the response.

3.1 Experimental Design

Given an input x (text and/or image), we define the function $L(x)$ as returning the (natural) language of the input.

The fidelity of a given langauge model $\theta(\cdot)$ and input x is defined as a binary indicator of whether the language of the input $L(x)$ equals the language of the output $L(\theta(x))$:

$$F(\mathbf{x}) = \begin{cases} 1 & \text{if } L(\mathbf{x}) = L(\theta(\mathbf{x})), \\ 0 & \text{otherwise} \end{cases}$$

In our experiments, we investigate the impact of including an image on fidelity. We compare inputs containing an image (x_{image}) against inputs where the image is replaced with a textual description of the image's content, $x_{description}$. The rationale for substituting the image with a textual description of the image's content is to maintain the semantic value of the input constant.

Thus, for each document (consisting of image and text pair) in our evaluation dataset, we define the Image Fidelity Loss (IFL) as:

$$\text{IFL} = F(x_{description}) - F(x_{image})$$

representing the fidelity loss incurred by substituting a text description of an image with the actual image.

3.2 Datasets

These image-text pairs are drawn from the three multilingual VQA benchmarks: MaXM (Changpinyo et al., 2023), PALO-LLaVAW (Maaz et al., 2024, hereafter LLaVAW) and ViSIT (Bitton et al., 2023), and summarized in Table 1.[1] These datasets all include a textual query referring to an image, plus a textual description of the image. In the case of ViSIT this description is generated conditional on the task instruction and verified by human annotators.

In total, for each model we collect 15480 responses spanning fourteen languages (7740 with an image plus 7740 with the description instead; see Table 2).

[1]Detailed dataset descriptions are included in the SI.

Dataset	MM	#Langs.	Size
PALO-LLaVAW	yes	10	600
MaXM	yes	7	2142
ViSIT	yes	10	5740
MultiQ	no	119	27400

Table 1: Overview of the employed datasets, indicating multimodality (MM), number of languages (#Langs.), and total number of observations (including parallel tasks repeated between languages).

Language	N	Language	N
Chinese (zh)	862	Japanese (ja)	585
Hindi (hi)	845	Spanish (es)	585
English (en)	842	German (de)	525
Hebrew (he)	805	French (fr)	324
Thai (th)	793	Romanian (ro)	284
Arabic (ar)	585	Russian (ru)	60
Bengali (bn)	585	Urdu (ur)	60

Table 2: Number of tasks per language in the datasets we study. ISO 639 language codes are provided in parentheses.

3.3 Measuring Language

We use the GlotLID model (Kargaran et al., 2023) to predict the language of the model output.[2] During the process, we observed non-random errors in the GlotLID predictions, such as having a lower accuracy on shorter texts. We correct for this bias in our downstream statistical analyses using the design-based supervised learning framework (Egami et al., 2023, DSL). DSL leverages a small number of randomly sampled expert annotations to correct for bias in downstream estimators caused by imperfect proxy measures. We manually label a stratified random sample of 1000 examples to use as our gold standard. The debiased results can be interpreted as being the results that would have been obtained if we had used expert annotation for all datasets. We provide details of the sampling weights and annotation method in the supplementary materials.

3.4 Prevalence of IFL

In order to motivate our research, we first assess the prevalence of IFL in existing LLaVA models. We apply the experimental design outline above to

[2]See supplementary materials for notes on manual post-processing of GlotLID outputs.

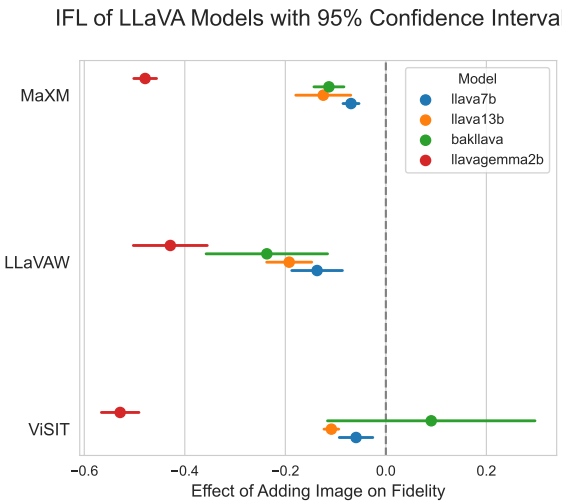

Figure 1: **IFL prevalence among existing LLaVA models.** Figure shows the effect of adding image to query on response fidelity (IFL) with 95% confidence intervals. All estimates are changes in probability, aggregated over languages within the benchmark and debiased using the DSL framework.

four popular LLaVA-style VLMs: LLaVA-v1.5-7b, LLaVA-v1.5-13b (Liu et al., 2023), BakLLaVA (SkunkworksAI, 2023) and LLaVA-Gemma-2B (Hinck et al., 2024).[3]

Figure 1 shows the debiased estimated effect and 95% confidence interval of adding an image on fidelity for each model and benchmark, aggregated across languages.

The magnitude varies by model and benchmark. The single largest drop is by LLaVA-Gemma-2b on ViSIT, where the response is 52.9 percentage points more likely to be in a different language than the query when an image is included in the input. With the exception of BakLLaVA on ViSIT, all effects are statistically significant at an alpha of 0.95. Because we use the DSL framework for estimation, these claims are statistically robust to errors from the language identification model.

Collectively, these results provide concrete evidence of a systematic issue: LLaVA models are more likely to reply in the *wrong* language when the user includes an image in the query. The remainder of this paper explores the source of this issue.

[3]We limit our scope to LLaVA-v1.5 models because the exact data mixture and training architecture for the newer v1.6 models has not been made public at the time of writing. Given our stated aim to explore the effects of the training decisions, we cannot provide insights into the newer models.

4 Effects of Design Choices

The LLaVA architecture combines a pretrained vision encoder and language model by using a small multi-layer perceptron (MLP) to project the penultimate hidden states of the vision encoder into the input embedding space of the language model (Liu et al., 2023). This architecture is then fine-tuned with two stages of training. In the first, the vision and language models are frozen and the projection MLP is trained on 558k image-caption pairs. In the second, the vision encoder is kept frozen and the projection MLP and language model are trained on 665k visual instruction-following and examples (Liu et al., 2023).[4]

4.1 Design Space

In constrast to related works that study the effect of architectural decisions in the VLM design space (Karamcheti et al., 2024), we focus our analysis on the effect of the choice of pretrained models and training data while holding the architecture constant.

LLaVA uses Vicuna-v1.5-7b (Zheng et al., 2023) as the language model, CLIP (Radford et al., 2021) as the vision encoder and English for more than 99% its training examples.

There are *a priori* reasons to think that any of these decisions could induce an "English bias" in the model. Vicuna is published as an English-language LLM trained primarily on English-language examples. The captions used to train the CLIP vision encoder are filtered for non-English texts (Radford et al., 2021, p.3), meaning that the representations produced by CLIP may be "biased" towards English language representations of visual data. Finetuning the model with primarily English data may "teach" the language model to reply to visual inputs from the vision encoder/MLP in English.

We ablate each of these design choices individually while holding architectural features constant to disentangle their effects. For our experiments, we focus on Chinese and German because these are languages for which there is an LLM at a similar size and architecture to Vicuna-7b that is not directly finetuned from Vicuna-7b. For Chinese, we use the Yi-6b-chat, a 6B-parameter LLM trained from scratch on a bilingual Chinese-English data mixture (AI et al., 2024). For German, we use LeoLM-7b-chat, a 7B-parameter LLM finetuned

[4]A technical primer on LLaVA is included in the SI.

Axis	Options
LLM	Vicuna-v1.5-7b (Zheng et al., 2023)
	Yi-6b-chat (AI et al., 2024)
	Leo-7b-chat (Plüster, 2023)
Vision	CLIP (Radford et al., 2021)
	DINOv2 (Oquab et al., 2024)
Data	English (Liu et al., 2023)
	Chinese (MT)
	German (MT)

Table 3: Summary of design space. Note that Yi is only combined with English and Chinese, and Leo with English and German.

Model	IFL	Accuracy
	Chinese	
LLM	0.17 [0.15, 0.19]	0.21 [-0.07, 0.50]
VE	-0.20 [-0.22, -0.18]	0.15 [-0.13, 0.43]
Data	-0.16 [-0.17, -0.14]	0.01 [-0.27, 0.30]
	German	
LLM	0.07 [0.04, 0.10]	0.28 [-0.35, 0.91]
VE	-0.11 [-0.15, -0.08]	-0.10 [-0.73, 0.53]
Data	-0.37 [-0.40, -0.33]	-0.24 [-0.87, 0.40]

Table 4: **Design Effects on IFL.** Each cell represents the point estimate and 95% confidence interval of the effect of changing the design feature (LLM, vision encoder or training data language) on IFL (left-hand column) and accuracy (right-hand column). The top half reports values for Chinese (LLM: Vicuna → Yi; Data: English → Chinese) and the bottom half reports values for German (LLM: Vicuna → Leo; Data: English → German). These are equivalent to β_2, β_3 and β_4 in Equation 1.

from Llama-2 (Touvron et al., 2023) on 65B German tokens (Plüster, 2023). For the vision encoder, we test the effect of substituting CLIP for DINOv2 (Oquab et al., 2024) because the latter is trained using a self-supervised training objective that does not incorporate language, while still using a Vision Transformer architecture (Dosovitskiy et al., 2021).

We use NLLB-1.7-distilled (Team et al., 2022) to machine translate all ∼1.2M training observations used in the LLaVA training data into Chinese and German. We provide estimates of the machine translation quality following techniques in (Qiu et al., 2022) in the supplement.

This design yields a total of fourteen combinations (Table 3), which we trained on using 8 × A6000 Nvidia GPU nodes on an internal cluster. All designs used the same training parameters as the original LLaVA-v1.5-7B model. We provide further training details in the supplementary materials.

4.2 Design Effects

For each set of experiments (Yi/Chinese and Leo/German), we measure the effect of training choices on IFL using the following regression model with first-order interactions:

$$\begin{aligned} \textit{Fidelity} = \beta_0 &+ \beta_1 \textit{Image} \\ &+ \beta_2 \textit{Image} \times \textit{Lang. Model} \\ &+ \beta_3 \textit{Image} \times \textit{Vision Model} \\ &+ \beta_4 \textit{Image} \times \textit{Data Lang.} + \epsilon \end{aligned} \quad (1)$$

where:

- *Fidelity* is a binary indicator for whether a completion to a query in the target language (Chinese or German) is in the correct language
- β_0 is a constant intercept term that captures the average fidelity of the reference class (LLaVA-v1.5-7b)
- β_1*Image* measures IFL; the change in fidelity when an image is added to the query
- β_2*Image* × *Lang. Model* captures how IFL changes when the LLM backbone is changed from from Vicuna to Yi or Leo
- β_3*Image* × *Vision Model* captures how IFL changes when the vision backbone is changed from CLIP to DINOv2
- β_4*Image* × *Data Lang.* captures how IFL changes when the training language is changed from English to Chinese or German
- ϵ is an error term

Coefficients β_2, β_3 and β_4 with the corresponding 95% confidence interval are reported in the left-hand side of Table 4. We see similar patterns for both languages. Changing the language model from Vicuna to Yi/Leo improved the performance of the model, reducing IFL by 17 and 7 pp for Chinese and German respectively. Changing the vision encoder from CLIP to DinoV2 worsened IFL, increasing it by 20 and 11 pp respectively. Changing the training data language worsened IFL considerably, increasing it by 16 and 37 pp respectively.

4.3 Effect on Accuracy

Although the above section provides insights into reducing IFL, how do these design decisions af-

fect the factual accuracy of responses? To measure this, we used GPT-4o (OpenAI, 2024) to generate zero-shot predictions of the accuracy. Our GPT-4o prompt gave the question, dataset ground truth (where available) and model completion and asked if the completion is correct given the question and ground truth label. We then used the DSL procedure to debias these evaluations, whereby the authors manually annotated 1000 observations to provide a gold standard.

We use the same regression setup as Equation 1, substituting the outcome *Fidelity* for *Accuracy*, a binary variable indicating whether a given response is correct given the question and dataset ground truth. The right-hand column of table 4 displays the effect of each design decision on the accuracy of responses in the target language.

We find no evidence for a systematic effect of any of the design decisions on accuracy. All estimated coefficients are statistically indistinguishable from 0, meaning our data does not support the hypothesis that changing the LLM, vision encoder or training data in the way described has a systematic effect on the accuracy of the response.

5 Locating the Cause of IFL

5.1 UMAP Analysis of Embedding Spaces

The previous experiments provide macro-level evidence that the LLM influences IFL based on the inputs and outputs of the models, but does not shed any light on what happens inside the model. In this section, we explore whether IFL is primarily attributable to any one component, and find the problem to reside in the LLM. We additionally find that a remarkably simple training-free mechanistic intervention in the LLM can reduce IFL.

To gain a qualitative understanding for how image embeddings interact with text embeddings in the input space, we employ Uniform Manifold Approximation and Projection (UMAP) for dimensionality reduction (McInnes et al., 2018). UMAP is a non-linear dimensionality reduction technique that preserves global data structure, making it ideal for visualizing high-dimensional data.

In our experiment, embeddings from text inputs and image tokens are jointly visualized using UMAP. Figure 2 (Left) illustrates that image embeddings cluster distinctly from text, demonstrating a demarcated separation in the latent space. This segregation suggests image embeddings predominantly occupy a unique region of the embedding space, indicating they are not directly embedding in the same area as any particular language.

5.2 CKA Analysis of Vision Embeddings

To further understand what is happening with the image embeddings, we use Centered Kernel Alignment (CKA) to measure the similarity of internal representations across differently trained models (Kornblith et al., 2019). CKA measures the similarity between two sets of data by comparing kernel matrices, which transform data into a high-dimensional space. A CKA score close to 1 indicates high similarity between datasets, while a score near 0 suggests low similarity. We use CKA in order to measure how the vision embeddings compare between two seperately trained VLMs: LLaVA-Yi trained in Chinese and LLaVA-Leo trained in German.

Figure 2 (center and right) shows a surprising result that vision embeddings maintain a consistent structure in the latent space across various models, regardless of the language backbone or the training data specifics. This consistency supports the finding in the UMAP visualization that image embeddings are in their own region of the input space. Additionally, it is the language model's responsibility to interpret these out-of-distribution embeddings in the language it determines best; it is not the case that the MLP adaptor model is placing the image embeddings near a particular language. Next, we explore how to encourage the LLM to remedy the IFL problem.

5.3 Mechanistic Intervention: Experiment

An exciting nascent field in LLM research is mechanistic interpretability, in which the ability to explicitly steer LLMs has recently been shown. This field shows great promise in modifying language model outputs, as demonstrated in recent studies that ablate refusals from tuned LLMs (Arditi et al., 2024) and the isolation of interpretable attributes (Gao et al., 2024; Templeton et al., 2024).

Given this ability to manually modify an LLM without training nor modifying the weight, we propose an intervention approach to reduce IFL by directly modifying the model's intermediate representations at runtime. We use a strikingly simple steering mechanism, using just one text example per language, and achieve significant reduction in IFL.

Our steering mechanism works by computing a language attribute a_{lang} in an intermediate layer,

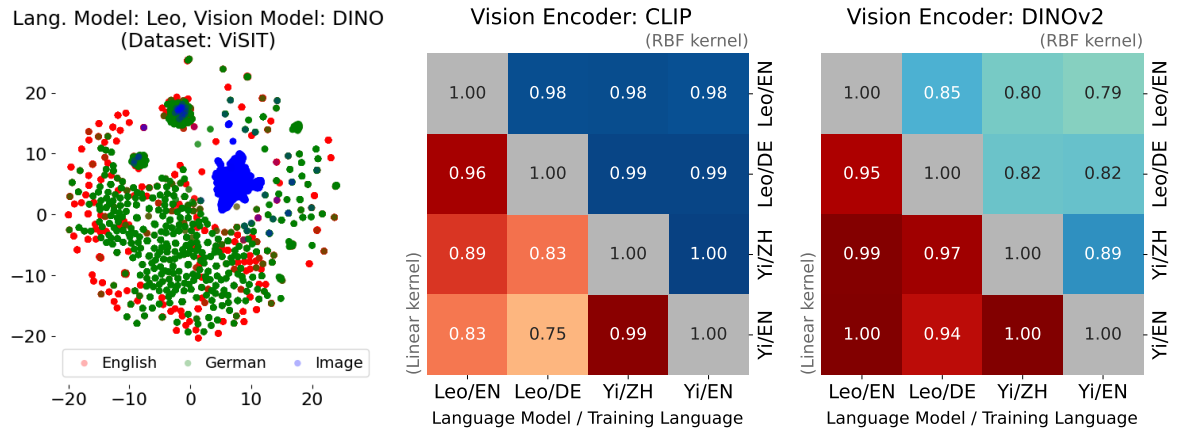

Figure 2: **(Left):** UMAP visualization of image and text embeddings from a multimodal language model. Image embeddings are shown clustering distinctly from text embeddings, indicating a unique separation in the latent space. This segregation highlights potential areas of focus for addressing fidelity loss in multimodal communication. **(Center and Right):** Centered Kernel Alignment (CKA) heatmap showing the similarity of vision embeddings across two differently trained language models. CKA based on a linear kernel is shown in the lower triangle; CKA based on an RBF kernel is shown in the upper triangle. The heatmap reveals a high degree of similarity in how vision data is embedded, regardless of the language model's architecture or training data specifics. This uniformity suggests that the method of integrating visual data into language models is a critical factor affecting fidelity.

then applying that attribute to every generated token. The attribute is computed as follows:

$$a_{lang} = LLM_l(x_{lang}) - LLM_l(x_{en})$$

where LLM_l represents the output at layer l, x_{en} is the sentence "Describe this image in detail.", and x_{lang} is the translated version of that sentence.

During inference, this direction is added to the output of layer l, effectively steering the model's behavior towards the desired language:

$$LLM*_l = LLM_l(o_{l-1}) + a_{lang}$$

where o_{l-1} is the output of the previous layer and $LLM*_l$ is the new, intervened layer. For our experiments layer l is selected to be partway through computation at one third depth (e.g., layer 10 out of 30).

5.4 Mechanistic Intervention: Results

The application of this mechanistic intervention has shown significant improvements in the fidelity of the model's responses across various languages. The quantitative improvements are displayed in Table 5, which presents a comparative analysis of fidelity metrics before and after the intervention. We find large relative improvements in performance across pretrained VLMs. Furthermore, we observe an improvement of IFL over the base LLM without image input.

Model	IFL	IFL + Remedy	Diff.	Relative Increase
llava7b	-0.085	-0.030	0.055	65%
llava13b	-0.175	-0.103	0.073	42%
bakllava	-0.073	0.098	0.170	233%
llava-gemma2b	-0.681	-0.513	0.168	25%

Table 5: Fidelity improvements by using mechanistic intervention (Remedy). Across all pretrained models, we find significant reduction in IFL by interventing on the LLM's intermediate layer. A full breakdown is available in the supplement.

While this experiment does require knowledge about which language attribute to select, it provides strong supporting evidence for the hypothesis that the LLM is responsible for IFL. Moreover, the successful application of a targeted mechanistic intervention highlights the potential of this approach to effectively address and resolve issues related to IFL, a future area of research we are interested in exploring further.

6 Implications

6.1 Multilingual Multimodal Understanding

Our results provide strong evidence in favor of the hypothesis that LLaVA models do not treat visual inputs orthogonally with respect to multilinguality,

and this bias occurs in the language backbone of the VLM.

In the first part of our analysis, we demonstrate the prevalence of IFL across different LLaVA models and languages. The second part of our analysis provides a controlled comparison of the effects of changing the linguality of the language backbone, the pretraining objective of the vision encoder, and the language of the training data.

We find that substituting the monolingual Vicuna language backbone for a bilingual Yi or Leo backbone reduces IFL in Chinese and German. In their exploration of the latent representations of multilingual inputs in Llama-2 models, Wendler et al. (2024) find that the abstract "concept space" of these LLMs lies closer to English than other languages. Correspondingly, if we think of LLaVA as instructing LLMs to map visual inputs to latent linguistic representations, an explanation for the positive effect of using a bilingual backbone may be that the higher proportion of the target language in the training data mix may reduce the extent to which visual concepts are mapped to spaces closer to English.

This interpretation is supported by our investigations of the intermediate representations of visual and text inputs to our suite of ablated LLaVA models. The UMAP and CKA results show that visual inputs are located in a separate space to textual inputs, and that this is remarkably consistent across training configurations. This indicates that LLaVA "instructs" the language backbone to interpret visual inputs, instead of learning to map visual inputs to a linguistic semantic space. Further evidence of the multimodal fusion occurring at an intermediary layer of the language model is shown in our mechanistic intervention analysis, where we show that the bias can be partially corrected through intervention on these layers.

Surprisingly, machine translating the training data language and changing the vision encoder to a language-agnostic one do not mitigate IFL. These null results should not necessarily be interpreted as negative ones; alternative explanations include limitations in performance stemming from the quality of the machine translated data, or there being a specific incompatibility between DINOv2 and the language backbones used.

On accuracy, we likewise find null results for a systematic effect of our ablations on accuracy. This is less surprising, given that our ablations were designed to affect IFL and not accuracy. Additional analyses on accuracy provide further evidence for the lack of a trade-off between accuracy and IFL for the models, languages and benchmarks explored.[5]

6.2 Building Robust Multilingual Multimodal Models

In terms of implications for the design of *multilingual* multimodal models, our research suggests that focus on improving the multilinguality of LLMs provides a productive way forward for building more robust multilingual VLMs. This finding is largely in line with the recent work on multimodal models, where progress has been fuelled by the increased availability of permissively licensed and capable pretrained LLMs. The mechanistic intervention results also suggest that effective and low-cost corrections to particular forms of bias (such as IFL) are a fruitful research direction.

7 Future Work

As the adoption of generative models proliferates, multilingual multimodal models that can service multiple regions will become an attractive solution. However, further efforts can be made to create more effective multilingual VLMs. One possible approach is the consideration of language fidelity (image-induced or not) during the training process through a mix of multilingual data and/or language-controlling instructions for in-context learning. Other in-depth approaches may include the design of better model architectures that integrate multitmodal representations at a suitable semantic level to prevent language bias. The phenomenon of IFL may also serve as a motivation to better understand the semantic role of visual representations within VLMs and how they are interpreted by the base language model.

We are particularly interested in understanding and improving the mechanistic intervention. We think this is a promising area of research, and that our simple approach could be improved either by automatic attribute selection conditioned on internal representation or better construction of the attribute. We also would like to study the broad class of fidelity loss in single modality models and to what extent mechanistic interventions are useful in those scenarios.

[5]These are located in the supplement because we wish to emphasize that a goal of this work is to motivate fidelity as an objective to maximize in its own respect.

8 Conclusion

In this work, we systematically examined the phenomenon of Image-induced Fidelity Loss in multimodal language models. Our analyses reveal that the addition of visual content to queries can unexpectedly bias the language output of these models, often resulting in responses in an incorrect language. Through a combination of empirical evaluations and introspective techniques, we quantified the extent of IFL across various VLM architectures and explored how different design choices impact language fidelity. Our findings are further substantiated by micro analyses within the models, illustrating the intricate interactions between image and text embeddings. Our introduction of a targeted mechanistic intervention demonstrates a potential method to mitigate IFL, indicating IFL is localized within the LLM itself, and suggesting a pathway for future enhancements to VLMs. Our work contributes to the broader understanding of multimodal systems and offers actionable insights for developing more robust and linguistically accurate AI technologies.

9 Limitations

9.1 Assumptions and Scope

When we debias our results with DSL, we assume that the expert annotated gold standard is in correct. Nevertheless, it is possible that the authors made mistakes during the data annotations, especially when identifying languages in unfamiliar scripts. The implication of this is that the bias-corrected results should be interpreted as "the answer we would get if we had manually annotated all of the examples" instead of "the truth".

The analyses in Section 5 in Section 6 contain a degree of speculation that we want to communicate clearly. The visible separation between image and text embeddings seen in the UMAP visualizations do not definitively prove that the inputs are mapped to distinct "semantic" spaces, and are not interpreted as such. Rather our takeaway is the lack of clustering between the visual tokens and a particular language that would be suggestive the projection layer learns to map images to a particular language.

The focus of this paper – IFL – could be framed as a form of model bias, and we sometimes describe it as such. We believe our definition of IFL is clearly articulated and operationalized: the decrease in linguistic fidelity in a VLM caused by adding an image to the input. We think that this constitutes a form of bias in the sense that it is a deviation from an intended/desirable output (i.e. replying in the same language as the query).

9.2 Risks and Ethics

A shortcoming of this work is the lack of focus on second- and third-order interactions, especially those looking at the differences in IFL between languages. The reason choosing to aggregate results in a setting where we expect heterogeneity is the focus of the paper: to define and demonstrate IFL as a problem that is not idiosyncratic to single languages. In future work we hope to better understand the heterogeneous patterns of IFL and how it relates both to structural differences between languages and their relative representation in LLM training corpora.

We do not foresee harm stemming from use of the models and data generated for this research. In particular, the models were trained for the purpose of assessing training decisions, and are unlikely to be competitive with other publicly available models engineered for optimal performance. Nevertheless, the model releases will be accompanied with clear caveats for safety.

While this paper has stressed the importance of research on multilingual VLMs for creating globally inclusive technology, we also want to stress the ethical risks of a mode of research that externalizes the contextualization and nuance necessary to achieve truly inclusive goals. Although the authors of this paper come from a broad constituency of cultures and languages, several of the language used in this study are not spoken by any of the authors. We take seriously the challenge of contributing to a global academic ideal while not appropriating languages or cultures, and will do our best to address any oversights in this respect.

Acknowledgements

Authors MH, MLO, SY, AB, ST and VL are employees of the Intel Corporation. The work of CH and AL is funded under the Excellence Strategy of the German Federal Government and the States.

References

01. AI, :, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang,

Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. 2024. Yi: Open foundation models by 01.ai. *Preprint*, arXiv:2403.04652.

Michael Andersland. 2024. Amharic llama and llava: Multimodal llms for low resource languages. *arXiv preprint arXiv:2403.06354*.

Andy Arditi, Oscar Balcells Obeso, Aaquib111, Wes Gurnee, and Neel Nanda. 2024. Refusal in llms is mediated by a single direction. *AI Alignment Forum*.

Yonatan Bitton, Hritik Bansal, Jack Hessel, Rulin Shao, Wanrong Zhu, Anas Awadalla, Josh Gardner, Rohan Taori, and Ludwig Schimdt. 2023. VisIT-Bench: A Benchmark for Vision-Language Instruction Following Inspired by Real-World Use. *arXiv preprint arXiv:2308.06595*.

Terra Blevins and Luke Zettlemoyer. 2022. Language contamination helps explains the cross-lingual capabilities of English pretrained models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3563–3574, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. 2023. InstructPix2Pix: Learning to Follow Image Editing Instructions. *arXiv preprint*. ArXiv:2211.09800 [cs].

Soravit Changpinyo, Linting Xue, Michal Yarom, Ashish Thapliyal, Idan Szpektor, Julien Amelot, Xi Chen, and Radu Soricut. 2023. MaXM: Towards Multilingual Visual Question Answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2667–2682, Singapore.

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.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Naoki Egami, Musashi Hinck, Brandon Stewart, and Hanying Wei. 2023. Using imperfect surrogates for downstream inference: Design-based supervised learning for social science applications of large language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 68589–68601. Curran Associates, Inc.

Constantin Eichenberg, Sidney Black, Samuel Weinbach, Letitia Parcalabescu, and Anette Frank. 2022. MAGMA – multimodal augmentation of generative models through adapter-based finetuning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2416–2428, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30K: Multilingual English-German image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74, Berlin, Germany. Association for Computational Linguistics.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. 2024. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*.

Musashi Hinck, Matthew L. Olson, David Cobbley, Shao-Yen Tseng, and Vasudev Lal. 2024. Llava-gemma: Accelerating multimodal foundation models with a compact language model. *Preprint*, arXiv:2404.01331.

Carolin Holtermann, Paul Röttger, Timm Dill, and Anne Lauscher. 2024. Evaluating the elementary multilingual capabilities of large language models with multiq. *Preprint*, arXiv:2403.03814.

Siddharth Karamcheti, Suraj Nair, Ashwin Balakrishna, Percy Liang, Thomas Kollar, and Dorsa Sadigh. 2024. Prismatic vlms: Investigating the design space of visually-conditioned language models. In *International Conference on Machine Learning (ICML)*.

Amir Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schuetze. 2023. GlotLID: Language identification for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6155–6218, Singapore. Association for Computational Linguistics.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. *arXiv preprint*. ArXiv:2102.03334 [cs, stat].

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of neural network representations revisited. In *International conference on machine learning*, pages 3519–3529. PMLR.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. 2020. The Open Images Dataset v4: Unified Image Classification, Object Detection, and Visual Relationship Detection at Scale. *International journal of computer vision*, 128(7):1956–1981.

Viet Lai, Chien Nguyen, Nghia Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan Rossi, and Thien Nguyen. 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 318–327, Singapore. Association for Computational Linguistics.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. *arXiv preprint*. ArXiv:2301.12597 [cs].

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, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916, New Orleans, LT, USA. Curran Associates, Inc.

Muhammad Maaz, Hanoona Rasheed, Abdelrahman Shaker, Salman Khan, Hisham Cholakal, Rao M Anwer, Tim Baldwin, Michael Felsberg, and Fahad S Khan. 2024. PALO: A Polyglot Large Multimodal Model for 5B People. *arXiv preprint arXiv:2402.14818*.

Oscar Mañas, Pau Rodriguez Lopez, Saba Ahmadi, Aida Nematzadeh, Yash Goyal, and Aishwarya Agrawal. 2023. MAPL: Parameter-Efficient Adaptation of Unimodal Pre-Trained Models for Vision-Language Few-Shot Prompting. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2523–2548, Dubrovnik, Croatia. Association for Computational Linguistics.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Jack Merullo, Louis Castricato, Carsten Eickhoff, and Ellie Pavlick. 2023. Linearly Mapping from Image to Text Space. *arXiv preprint*. ArXiv:2209.15162 [cs].

OpenAI. 2024. Gpt-4o: Documentation. Accessed: 2024-06-15.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. 2024. Dinov2: Learning robust visual features without supervision. *Preprint*, arXiv:2304.07193.

Björn Plüster. 2023. LeoLM: Igniting German-Language LLM Research. Accessed: 2024-06-15.

Chen Qiu, Dan Oneață, Emanuele Bugliarello, Stella Frank, and Desmond Elliott. 2022. Multilingual multimodal learning with machine translated text. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4178–4193, Abu Dhabi, United Arab Emirates. 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, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR. ISSN: 2640-3498.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A Survey of Cross-lingual Word Embedding Models. *Journal of Artificial Intelligence Research*, 65:569–631.

Florian Schneider and Sunayana Sitaram. 2024. M5 – a diverse benchmark to assess the performance of large multimodal models across multilingual and multicultural vision-language tasks. *Preprint*, arXiv:2407.03791.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.

Dongjae Shin, Hyunseok Lim, Inho Won, Changsu Choi, Minjun Kim, Seungwoo Song, Hangyeol Yoo, Sangmin Kim, and Kyungtae Lim. 2024. X-llava: Optimizing bilingual large vision-language alignment. *arXiv preprint arXiv:2403.11399*.

SkunkworksAI. 2023. Bakllava-1. Hugging Face model repository.

Yueqi Song, Simran Khanuja, and Graham Neubig. 2024. What is missing in multilingual visual reasoning and how to fix it. *arXiv preprint arXiv:2403.01404*.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, et al. 2024. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*.

Ashish V. Thapliyal, Jordi Pont Tuset, Xi Chen, and Radu Soricut. 2022. Crossmodal-3600: A Massively Multilingual Multimodal Evaluation Dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 715–729, Abu Dhabi, United Arab Emirates.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2022. SimVLM: Simple Visual Language Model Pretraining with Weak Supervision. *arXiv preprint*. ArXiv:2108.10904 [cs].

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in english? on the latent language of multilingual transformers. *Preprint*, arXiv:2402.10588.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. Wildchat: 1m chatgpt interaction logs in the wild. *arXiv preprint arXiv:2405.01470*.

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

Appendix

A Computational Experiments

Computational Budget The training experiments for this paper were conducted on an internal cluster using nodes with 8 × A6000 Nvidia 48GB GPUs. In total, we trained 32 distinct configurations (not all of which were ultimately used). A single end-to-end training run with a 7-billion parameter LLM backbone takes 25 hours, meaning roughly 800 GPU hours were used for training. Inference experiments were run on a mixture of RTX3090 24GB cards, A6000 24GB cards and A6000 48GB cards. These required roughly an additional 900 GPU hours. Data analysis utilized CPU. The only sizable compute consisted of applying the DSL estimator to large datasets, which required on the order of $\sim$ 500 CPU hours. Finally, the GPT-4o annotation for the roughly 730k completions in our experiments required roughly USD 2k worth of completion calls.

B Expert Annotation

Sampling Weights We stratified on evaluation benchmark (i.e. we weighted the probability by the inverse proportion of the originating benchmark to the full dataset) and then upweighted German by a factor of 4, Chinese and Hindi by 2, and downweighted Romanian, Russian and Urdu by a factor of 2. We sampled a total of 1000 observations (without replacement) using these weights.

Annotation Procedure The 1000 observations were uploaded into spreadsheets for the authors to manually annotate. Where possible, annotations were matched to authors who could read the language used in the query. The annotation consisted of three questions: what language is the answer, does the model completion match the gold standard, and is the answer correct. The latter two questions were restricted to three categories: true, false and NA. NA was used where the model did not provide coherent output.

C Automated Evaluation

GlotLID We use the GlotLID v3 (Kargaran et al., 2023) model for automated language identification. We take the most-likely language as predicted by GlotLID, and then manually process the label to collapse what we thought were common misclassifications by the model, such as classifying Mandarin Chinese into various languages and dialects using the simplfied Chinese script when the outputs contained a mix of non-Chinese punctuation characters and Chinese glyphs.

The full parsing rule is as follows:

```
def parse_glotlid(lang: str) -> str:
    iso, script = tuple(lang.split("_"))
    match script:
        case "Hani":
            return "chinese"
        case "Jpan":
            return "japanese"
        case "Deva":
            return "hindi"
        case "Beng":
            return "bengali"
        case "Hebr":
            return "hebrew"
        case "Thai":
            return "thai"
        case "Cyrl":
            return "russian"
        case "Zzzz":
            return "none"
        case "Arab":
            match iso:
                case "urd":
                    return "urdu"
                case _:
                    return "arabic"
        case "Latn":
            match iso:
                case "deu":
                    return "german"
                case "eng":
                    return "english"
                case "spa":
                    return "spanish"
                case "ron":
                    return "romanian"
                case "fra":
                    return "french"
                case _:
                    return "other_latin"
        case _:
            return "other"
```

D Datasets Used

Here we provide an overview on the datasets we employ in our study.

MaXM was introduced by Changpinyo et al. (2023) and is a VQA dataset comprising seven lan-

guages in five scripts. In MaXM, the questions and their respective answers are in the same language. Moreover, in MaXM, the images are a subset of the XM3600 (Thapliyal et al., 2022) dataset and are chosen to match a region where the language of the question-answer pair is spoken. To increase the cultural diversity, the images selected to match the region where the language of the question-answer pair is spoken.

VisIT-Bench stands for **Vis**ual **I**nstruction **T**uning **Bench**mark (Bitton et al., 2023). The dataset consists of 592 vision-language tasks written by human researchers, with GPT-4-generated responses and dense instruction-conditioned captions of the image that are rated by human coders. The 562 images are taken from the OpenImages (Kuznetsova et al., 2020) v7 dataset. In this work we use 525 examples where the GPT-4 generated responses are rated as correct by human annotators. We machine translate these examples into Arabic, Bengali, Chinese, German, Hebrew, Hindi, Japanese, Spanish and Thai using the Azure Translation API.

To check the translation quality, we inspected 25 randomly sampled translations in Chinese, Hindi, Hebrew, German, Japanese and Spanish (languages where the authors had access to native speakers). Among these, the majority (19 out of 25) of translations were deemed to not significantly change the meaning of the original. In the remainder, issues observed included omitting details (such as not mentioning an object or descriptor), or constructing words that were understandable but not "natural" in the target language. In general the question/instruction was correctly translated, but the translation of the gold standard varied in quality. This presents a limitation for this research, but one that cannot be overcome without greater resources for expert/higher quality translation.

PALO-LLaVA-Bench-In-The-Wild dataset is a multilingual VQA dataset created by the PALO authors (Maaz et al., 2024) by machine translating the original LLaVA-Bench-In-The-Wild (Liu et al., 2023) in 10 languages using a fine-tuned GPT-3.5 instance. The dataset comprises of 60 questions per language considering 24 diverse images with a caption describing the visual content.

MultiQ is an evaluation dataset for open-ended question answering covering 137 typologically diverse languages. It is specifically constructed to only contain questions that are simple, factual, and target common knowledge to only test the multilingual capabilities of language models, and no complex reasoning (Holtermann et al., 2024).

D.1 Machine Translation of Training Data

As noted in the main body, we machine translate (MT) the LLaVA training data into Chinese and German using the NLLB-1.7-distilled model (Team et al., 2022). The choice of this model was primarily motivated by resource availability for translating 1.2M texts into two languages.

We apply two automated translation quality checks for the training data based on the MT checks in Qiu et al. (2022). The first is the token-type-ratio (TTR) of each of the languages. A value close to 0 indicates a high degree of repetition, which is an observed pathology of neural MT models. The second is the BLEU score between the source and MT texts. A BLEU score close to 1 indicates the presence of copied English text.

The highest BLEU score for source to target across all translated examples is $1.6e-231$, indicating that copying did not occur. Figure 3 shows the values for the TTR check. We find that in both cases our MT data has a higher cumulative TTR curve than the English data; this indicates less token repetition. It is hard to directly interpret this value, given that baseline TTR should vary between languages, but the lack of an obvious negative pattern is reassuring.

E Models Used

Here we provide an overview of the models we employed in our study.

OpenAI/CLIP is a jointly optimized vision and text feature extractor trained using large-scale image-caption pairs (Radford et al., 2021). CLIP is focused on learning image representations from scratch that are trivially transferable to many downstream tasks without the need for domain specific training.

DINOv2 is a series of image encoders trained on curated data using unsupervised learning (Oquab et al., 2024). Through an improved training recipe and larger dataset, followed by a distillation process of larger to smaller models, DINOv2 is positioned as a ViT-based general-purpose image encoder that surpasses OpenAI/CLIP on most benchmarks.

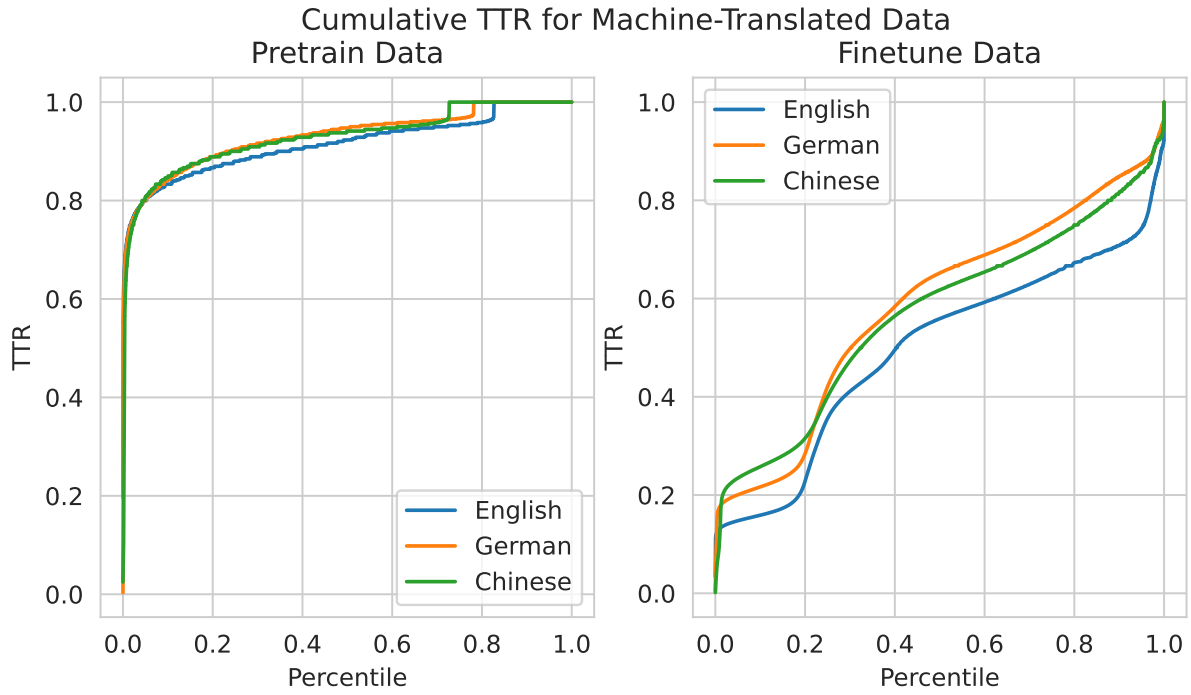

Figure 3: Token-type ratio (TTR) for pretraining (left) and finetuning training datasets.

LLaVA-v1.5 is a large multimodal model trained end-to-end with visual instruction following (Liu et al., 2023). The model combines a vision model — OpenAI/CLIP — with a large language model — Vicuna-v1.5 — achieving impressive visual and language understanding results that were state-of-the-art at its release. In this work we used the 7b and 13b variants of the model.

BakLLaVA is a large multimodal model based on the LLaVA-v1.5 architecture using Mistral-7b as the base LLM (SkunkworksAI, 2023). The model utilizes training data from LLaVA-v1.5 as well as additional sources including ShareGPT and private data with a permissive license.

Yi-6b-chat is a large language model trained from scratch on English and Chinese corpora (AI et al., 2024). In this work, we use the 6b variant that has been extended with chat-style training.

Leo-7b-chat is a large language model that extends Llama-2 into German through continued training on a large German corpus (Plüster, 2023).

GlotLID v3 is a language identification model that coveres 2102 languages. The data used to train this model was sourced from Wikipedia, news sites, translation corpora, religious text, and storybooks.

NLLB-1.7-distilled is translation model that support direct translation between 200 languages, including many low-resource languages (Team et al., 2022). The datasets used to train NLLB (No Language Left Behind) were sourced from professionally translated sentences in the Wikipedia domain in addition to publicly available translation datasets.

GPT-4o is a commercial large language model provided from OpenAI.

F Technical Explainers

F.1 Primer on LLaVA

What is LLaVA? Our study analyzes LLaVA, a multimodal model (VLM) that integrates a pretrained vision encoder, denoted as E_V, with a large language model (LLM), using a connecting multilayer perceptron (MLP). The process is defined in two main training stages: pretraining of the MLP and joint finetuning of the MLP with the LLM.

Model Architecture The VLM comprises the following components:

Vision Encoder: The vision encoder E_V processes the visual input X_v to produce a set of embeddings $E_V(X_v)$.

MLP Connector: A connecting MLP, defined as F, transforms the output of E_V into the dimen-

stionality of the LLM. This transformation is represented as $F(E_V(X_v))$.

LLM: The LLM processes both textual query X_q and the transformed vision embeddings. The combined input to the LLM is given by concatenating the embeddings from the MLP with text embeddings, i.e., $LLM([F(E_V(X_v)); E_L(X_q)])$, where E_L denotes

The VLM is defined as a function that takes an image input X_v and a textual question X_q, and processes these through the vision encoder, MLP connector F, and LLM to produce an output X_a, which is the model's answer to the question based on the visual context. Formally, the VLM can be expressed as:

$$VLM(X_v, X_q) = LLM\left([\mathrm{F}(E_V(X_v)); E_L(X_q)]\right), \quad (2)$$

where $E_V(X_v)$ is the output of the vision encoder for the input image, $\mathrm{F}(E_V(X_v))$ is the transformed visual embedding suitable for the LLM, and $E_L(X_q)$ represents the embedded form of the textual question. The final output X_a is generated by the LLM, which synthesizes and integrates both the visual and textual information to produce a contextually appropriate answer.

Training Procedure The training of the VLM is structured into two distinct stages: pretraining and finetuning. During the pretraining stage, the MLP is trained while keeping E_V and the LLM frozen. The objective is to optimize the MLP to map the vision encoder outputs to a representation that is effectively integrable with the LLM. The training uses a custom dataset of 595k samples filtered from CC3M (Sharma et al., 2018):

$$\mathcal{L}_{\text{MLP}} = \sum_{(X_v, X_c) \in \mathcal{D}} L_{CE}(VLM(X_v, X_q)), \quad (3)$$

where X_c represents the captions associated with X_v, and $\mathcal{D}$ denotes the dataset.

Finetuning In the finetuning stage, the MLP and the LLM are jointly trained with a larger, diverse dataset of 665k multimodal instruction tuning examples, integrating both synthetic and established vision-language training sets. The entire conversation $C = (X_q, X_a)$ is fed into the LLM, with autoregressive masking applied to focus training on the answers using supervised cross-entropy loss L_{CE}:

$$\mathcal{L}_{\text{VLM}} = \sum_{C \in \mathcal{C}} L_{CE}(VLM(X_v, X_q)), \quad (4)$$

llava7b

dataset	Lang.	IFL	IFL + Remedy	Diff.
llavaw	ar	-0.250	-0.083	0.167
	bn	-0.117	-0.050	0.067
	zh	-0.233	-0.017	0.217
	fr	-0.183	0.000	0.183
	hi	-0.133	-0.033	0.100
	ja	-0.117	-0.050	0.067
	ru	-0.233	-0.017	0.217
	es	-0.200	-0.050	0.150
	ur	-0.050	0.083	0.133
maxm	zh	0.004	0.000	-0.004
	fr	0.004	-0.011	-0.015
	he	-0.132	-0.125	0.007
	hi	-0.042	-0.035	0.008
	ro	0.000	-0.014	-0.014
	th	-0.007	-0.011	-0.004
visitazure	ar	-0.038	-0.047	-0.009
	bn	-0.084	-0.038	0.045
	zh	-0.026	-0.047	-0.021
	de	-0.054	-0.037	0.017
	he	-0.038	-0.037	0.002
	hi	-0.026	-0.009	0.017
	ja	-0.021	-0.051	-0.030
	es	-0.024	-0.042	-0.017
	th	-0.045	-0.010	0.035
average	-	-0.085	-0.030	0.055

Table 6: mechint raw llava7b scores.

where $\mathcal{C}$ represents the conversation dataset, and training focuses exclusively on the answer parts X_a, leveraging the context provided by the entire conversation but training only through the answer segments.

G Training

G.1 Hyperparameters

All models were trained using the same hyperparameters as the original LLaVA-v1.5-7b model. This training takes place in two stages, as described above.

In the first ("pretraining") stage, we trained with a global batch size of 256 and a learning rate of $1e-3$. In the second ("finetuning") stage, we used a global batch size of 128 and a learning rate of $2e-5$. For both stages we trained for a single epoch, with a warmup ratio of 0.03 and a cosine annealed learning rate scheduler.

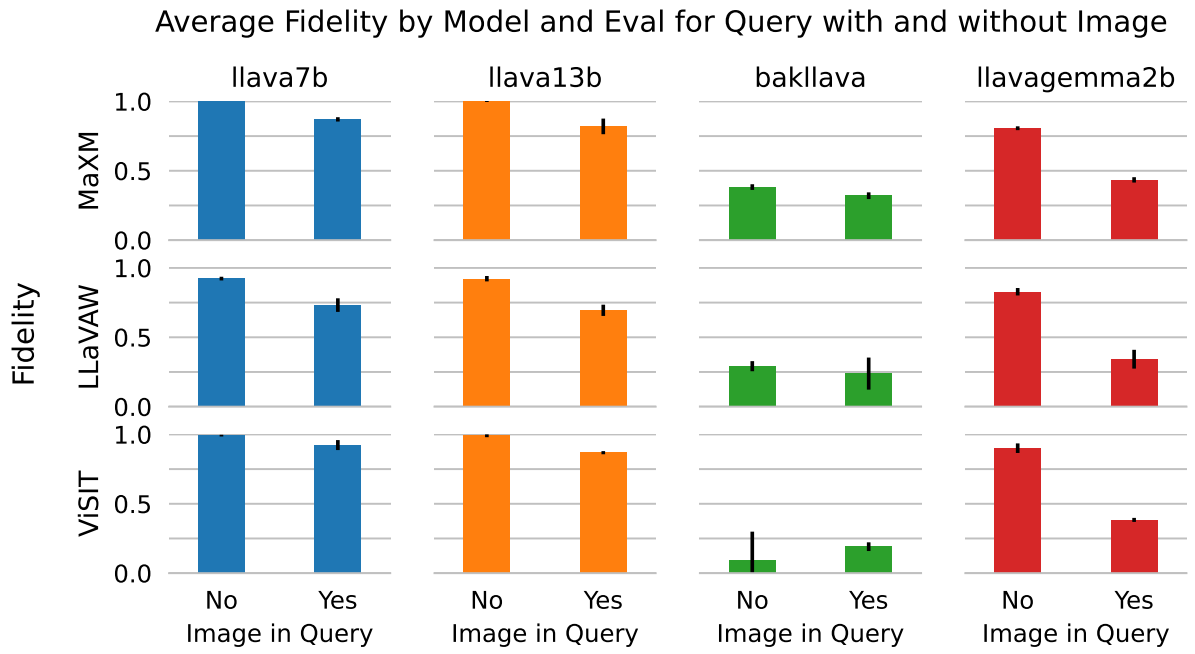

Figure 4: Average fidelity by model and eval for query with and without Images

G.2 Convergence

In order to ensure comparability across experiments, we trained every model the same amount (one epoch). However, as a hedge against random failures during training, we monitored the training loss curves. All checkpoints saw similar proportional decrease in training loss from their tenth to final training step, ranging from a 38.9% to 65.8% decrease in training loss. Figure 5 shows the loss curves for each model.

H Accuracy

H.1 Accuracy of Baseline LLaVA Models

The accuracy of the base LLaVA models is not very high for the languages and benchmarks considered. Table 10 provides a breakdown of accuracy by each of the languages in the benchmarks. We see that the 7B and 13B models fail to exceed even 40 and 50 percent accuracy respectively. These results are consistent with concurrent findings in Schneider and Sitaram (2024). We do not see these results as problematic for our research, as we want to emphasize the goals of fidelity and (factual) accuracy as being independently pursuable.

H.2 Accuracy-Fidelity Trade-off

In addition to the findings in the main body of this paper, further experiments indicate weak evidence in support of there being a trade-off between optimizing accuracy and fidelity. Table 11 provides the Pearson correlation coefficient between accuracy and fidelity for each of the models included in our analysis. We find that for only five out of 26 models is there a significant correlation, with the value ranging from -0.514 to 0.541. We do not find any pattern from these results to suggest a systematic finding for a trade-off.

I Use of AI Tools

The authors of this paper used Github Co-pilot for coding assistance for this research.

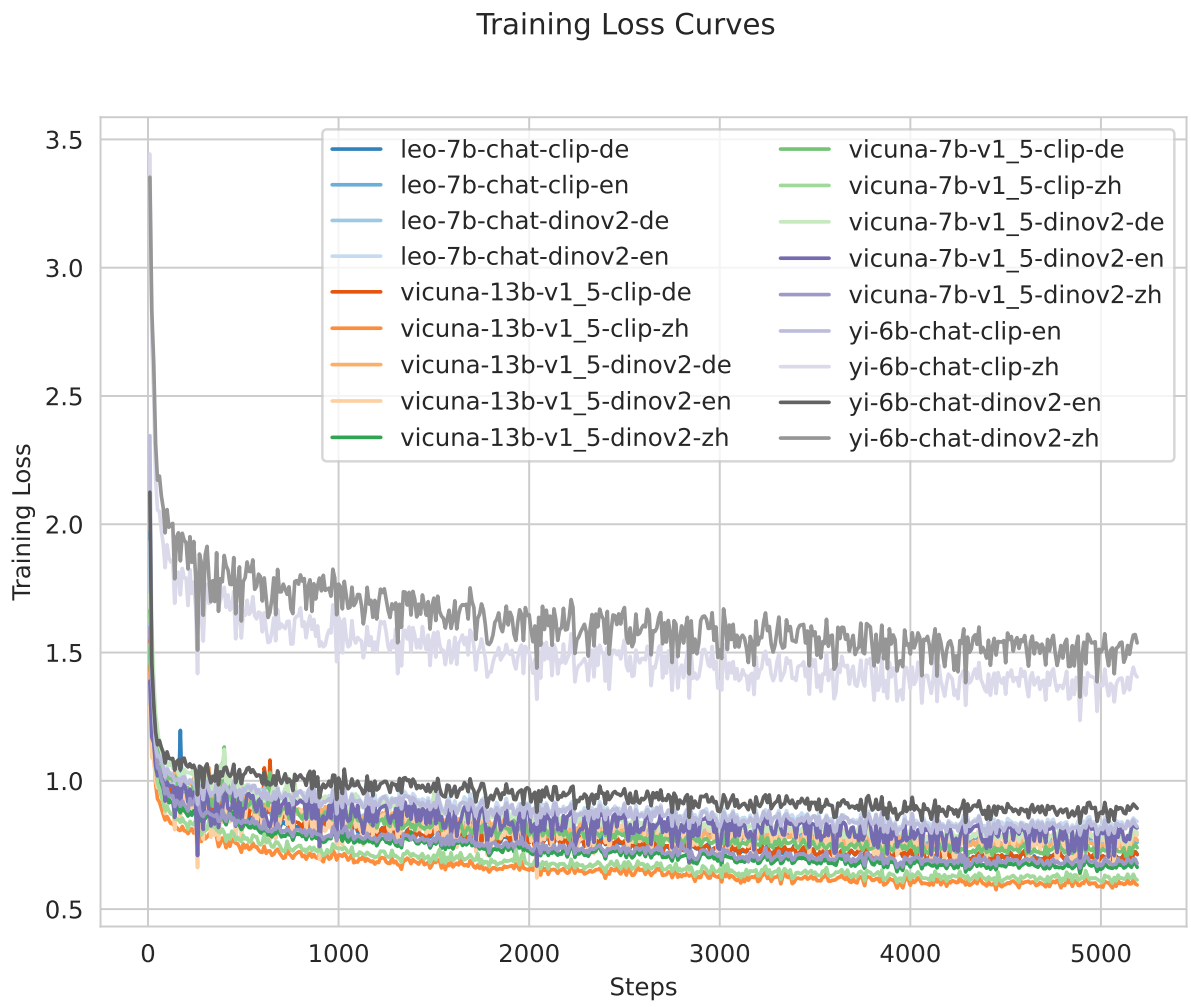

Figure 5: Training loss curves for finetuning. Legend indicates language backbone, vision encoder and training language as two-letter code.

llava13b

dataset	Lang.	IFL	IFL + Remedy	Diff.
llavaw	ar	-0.183	-0.033	0.150
	bn	-0.233	-0.083	0.150
	zh	-0.133	-0.033	0.100
	fr	-0.200	-0.100	0.100
	hi	-0.317	-0.200	0.117
	ja	-0.183	-0.117	0.067
	ru	-0.433	-0.317	0.117
	es	-0.233	-0.183	0.050
	ur	-0.550	-0.267	0.283
maxm	zh	-0.025	-0.007	0.018
	fr	-0.008	-0.045	-0.038
	he	-0.175	-0.121	0.054
	hi	-0.042	-0.035	0.008
	ro	-0.106	-0.085	0.021
	th	-0.157	-0.093	0.063
visitazure	ar	-0.174	-0.066	0.108
	bn	-0.244	-0.136	0.108
	zh	-0.071	-0.031	0.040
	de	-0.105	-0.094	0.010
	he	-0.125	-0.082	0.044
	hi	-0.136	-0.096	0.040
	ja	-0.056	-0.044	0.012
	es	-0.057	-0.042	0.016
	th	-0.258	-0.155	0.103
average	-	-0.175	-0.103	0.073

Table 7: Mechanistic intervention complete llava13b scores.

bakllava

dataset	Lang.	IFL	IFL + Remedy	Diff.
llavaw	ar	0.000	0.350	0.350
	bn	-0.050	0.217	0.267
	zh	-0.033	-0.067	-0.033
	fr	-0.117	0.000	0.117
	hi	0.000	0.050	0.050
	ja	-0.017	-0.067	-0.050
	ru	0.000	0.000	0.000
	es	-0.117	0.217	0.333
	ur	-0.017	0.183	0.200
maxm	zh	-0.018	0.014	0.032
	fr	-0.318	-0.223	0.095
	he	0.000	0.029	0.029
	hi	0.000	0.135	0.135
	ro	-0.567	-0.299	0.268
	th	-0.119	-0.078	0.041
visitazure	ar	-0.010	0.608	0.618
	bn	-0.012	0.557	0.570
	zh	-0.007	0.019	0.026
	de	-0.136	-0.108	0.028
	he	0.000	0.078	0.078
	hi	-0.007	0.113	0.120
	ja	-0.014	0.026	0.040
	es	-0.183	0.291	0.474
	th	-0.007	0.294	0.301
average	-	-0.073	0.098	0.170

Table 8: Mechanistic intervention complete bakllava scores.

llavagemma2b

dataset	Lang.	IFL	IFL + Remedy	Diff.
llavaw	ar	-0.583	-0.533	0.050
	bn	-0.483	-0.483	0.000
	zh	-0.733	-0.567	0.167
	fr	-0.800	-0.433	0.367
	hi	-0.500	-0.317	0.183
	ja	-0.600	-0.650	-0.050
	ru	-0.883	-0.650	0.233
	es	-0.900	-0.700	0.200
	ur	-0.517	-0.483	0.033
maxm	zh	-0.852	-0.762	0.090
	fr	-0.905	-0.652	0.254
	he	-0.768	-0.482	0.286
	hi	-0.731	-0.546	0.185
	ro	-0.810	-0.637	0.173
	th	-0.646	-0.455	0.190
visitazure	ar	-0.718	-0.578	0.139
	bn	-0.483	-0.420	0.063
	zh	-0.672	-0.552	0.120
	de	-0.688	-0.258	0.430
	he	-0.617	-0.280	0.336
	hi	-0.589	-0.375	0.214
	ja	-0.526	-0.509	0.017
	es	-0.793	-0.608	0.185
	th	-0.538	-0.373	0.166
average	-	-0.681	-0.513	0.168

Table 9: Mechanistic intervention complete LLaVA-Gemma-2b scores.

Query Language	**LLaVA 7B**	**LLaVA 13B**
English	0.372	0.392
French	0.340	0.417
Urdu	0.317	0.317
Russian	0.183	0.233
Bengali	0.161	0.222
Spanish	0.099	0.162
Japanese	0.115	0.130
Chinese	0.129	0.160
German	0.107	0.154
Romanian	0.025	0.271
Hindi	0.099	0.128
Thai	0.108	0.091
Arabic	0.080	0.087
Hebrew	0.063	0.080

Table 10: Performance of LLaVA Models Across Different Languages

LM	**VE**	**Data**	**Corr**	**p-val**
vicuna13b	dino	en	-0.514	0.006
yi6b	dino	en	-0.432	0.025
yi6b	clip	en	-0.413	0.032
yi6b	dino	en	-0.413	0.032
yi6b	clip	en	-0.386	0.047
leo	dino	en	-0.284	0.151
leo	dino	de	-0.239	0.231
leo	dino	en	-0.218	0.274
vicuna13b	clip	zh	-0.175	0.382
leo	clip	en	-0.152	0.448
leo	clip	de	-0.131	0.515
vicuna13b	clip	de	-0.112	0.579
vicuna13b	dino	zh	-0.060	0.768
yi6b	dino	zh	-0.057	0.776
yi6b	dino	zh	-0.054	0.788
yi6b	clip	zh	-0.050	0.804
vicuna7b	dino	zh	-0.035	0.862
vicuna7b	dino	en	-0.034	0.865
leo7b	clip	en	-0.029	0.885
yi6b	clip	zh	-0.023	0.908
vicuna13b	dino	de	0.023	0.909
vicuna7b	clip	zh	0.069	0.731
vicuna7b	clip	de	0.088	0.664
vicuna7b	dino	de	0.099	0.624
leo	dino	de	0.192	0.338
leo	clip	de	0.225	0.258

Table 11: Correlation between accuracy and fidelity by model.