How do VLMs balance reliance on image and text across languages?

Hoang Minh Anh (Minnie) Nguyen

Master of Science
Artificial Intelligence
School of Informatics
University of Edinburgh
2025

Abstract

Recent vision language models (VLMs) have demonstrated strong multimodal capabilities, achieving remarkable performance in vision-and-language tasks such as image captioning, multimodal chat, and visual question answering. The language model backbones employed in these multimodal architectures have showcased robust proficiency in diverse languages, thus enabling multilingual capabilities in VLMs. Although previous studies have been conducted to evaluate the degree of reliance on the vision and text modalities in VLMs, they mainly employ English-centric multimodal benchmarks, leaving the vast majority of languages understudied. In this work, we develop a framework to quantify and compare the degree of multimodality across typologically diverse languages in the most recent decoder-only VLMs. To this end, we employ and contrast two complementary methods – the perceptual score and MM-SHAP – using multiple high-resource and low-resource languages, question types, and models. Across the evaluated languages and question types, the perceptual score results indicate that model performance relies more heavily on the textual inputs than the visual inputs, while the MM-SHAP scores demonstrate that the text modality has a greater contribution towards the model prediction than the vision modality. In addition, the perceptual score reveals model biases and varying extents of modality reliance based on the different question types, whereas MM-SHAP indicates different textual and visual contributions among the models and languages, and offers a fine-grained analysis of the token-level contributions. We hope that the findings in this work provide useful insights for future developments of multilingual VLMs and multimodal datasets to improve cross-modal and cross-language alignment. The code for our experiments and analysis can be found at `https://github.com/nminnie/CC-SHAP-VLM`.

Research Ethics Approval

This project was planned in accordance with the Informatics Research Ethics policy. It did not involve any aspects that required approval from the Informatics Research Ethics committee.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Hoang Minh Anh (Minnie) Nguyen*)

Acknowledgements

The author would like to sincerely thank Dr. Ivan Titov and Dr. Rochelle Choenni for their supervision and support throughout the project.

This work used the Cirrus UK National Tier-2 HPC Service at EPCC (`http://www.cirrus.ac.uk`) funded by the University of Edinburgh and EPSRC (EP/P020267/1).

Table of Contents

Chapter 1

Introduction

There have been significant advancements to vision-language models (VLMs) in recent years, integrating language and visual capabilities into a unified model. Many contemporary VLMs leverage pretrained vision and text encoders instead of training these components from scratch, drastically reducing the computation overhead. The pretrained text encoder is often a powerful open-source large language model (LLM), whereas the vision encoder is typically a visual feature extractor such as CLIP [30]. When utilizing these pretrained backbones, a common approach is to train a projection layer to map the image embeddings to the text embedding space to connect the two modalities. This cross-modal alignment can be enhanced via visual instruction tuning [23], in which the VLM is trained to generate responses based on instructions pertaining to the given images. While this approach benefits from the strengths of existing models, such VLMs are susceptible to biases from the pretrained encoders [5]. Recent pretrained LLM backbones have become increasingly multilingual, raising the need to further investigate the alignment between text and vision modalities across languages.

This project proposes a systematic framework to measure and compare the degree of multimodality across different languages for multilingual decoder-only VLMs. More specifically, we investigate whether multilingual VLMs are able to effectively exploit both the visual and textual inputs in high-resource and low-resource languages. Ensuring robust multilingual capabilities in VLMs is important as it enables speakers of underrepresented languages to interact with technology in their native languages and mitigates linguistic and cultural biases that could arise from English-centric models. However, evaluating VLMs in the multilingual setting is challenging due to the data scarcity of low-resource languages, including both text-only as well as multimodal datasets, as discussed in [2]. In addition to the lack of high-quality evaluation benchmarks in under-resourced languages, imbalanced training data across languages could lead to varying levels proficiency in the LLM backbones.

In this work, we compare across two decoder-only models, 5 question types, and 8 languages, using two complementary methods. Assessing of the degree of multimodality is crucial for understanding the cross-modal and cross-language alignment in multilingual VLMs, resulting in improved model reliability and interpretability. Insights from this work could provide helpful guidance for future developments of multilingual VLMs to better align the modalities across multiple languages. In addition, this project addresses current concerns with VLMs, including the trend in which recent models rely increasingly more on the textual input and insufficiently utilize the visual input [11] and unimodal collapse [26], in which VLMs leverage a single modality and ignore the other.

Although some previous studies have evaluated the extent to which VLMs rely on the textual and visual inputs, most have focused exclusively on the English language, leaving the majority of under-represented languages largely underexplored. Moreover, these studies primarily employed encoder-only VLMs [11, 27] or earlier decoder-only VLM architectures [28]. In this work, we extend the analysis to a typologically diverse set of languages and to more recent decoder-based VLMs, including LLaVA-OneVision [18] and Pangea [38].

To this end, we employ and compare two complementary methods for quantifying the degree of multimodality in VLMs: the perceptual score [11] and MM-SHAP [27]. The perceptual score provides an intuitive measure of the importance of each modality towards the model performance, whereas MM-SHAP provides a more fine-grained analysis at the token level in addition to contribution scores of each modality towards the model prediction. Across all languages and question types employed in this study, the perceptual score results consistently indicate that in recent decoder-only VLMs, the textual inputs have a stronger influence on the model performance than the visual inputs, while MM-SHAP scores show that the text modality has a greater contribution towards the model prediction than the vision modality. In addition, the perceptual score highlights the model's tendency to make educated guesses based on its pretrained knowledge across different languages, with modality importance varying according to the type and difficulty of the questions, challenging the reliability of certain question types in multimodal evaluation benchmarks. In contrast, MM-SHAP results suggest that reliance on the textual input is related to the model's proficiency in the target language.

In the subsequent chapters, we outline the background and related work on quantifying the degree of multimodality in VLMs, introduce the dataset, models, and methods employed in this work, present and discuss the experiment results, and conclude with key findings and directions for future work.

Chapter 2

Background and Related Work

2.1 Multilingual Vision-Language Models

Vision–language models (VLMs) are artificial intelligence systems designed to integrate textual and visual information to enable cross-modal capabilities. Textual inputs are processed by a language model, such as a large language model (LLM), while visual inputs are handled by a vision encoder. Encoder-only VLMs use contrastive learning or masking techniques to directly construct the image and text features within a shared embedding space, making them very effective in multimodal representation learning. In contrast, in decoder-only architectures such as LLaVA, the visual features and text embeddings are processed separately by the vision encoder and the language model, then a trainable projection layer maps the visual features into the text embedding space as additional inputs to the language model, as illustrated in Figure 2.1. Thus, decoder-only VLMs are capable of generating outputs conditioned on the input texts and images. In particular, many recent decoder-only VLMs employ multilingual pretrained LLM backbones, enabling cross-modal capabilities in various languages.

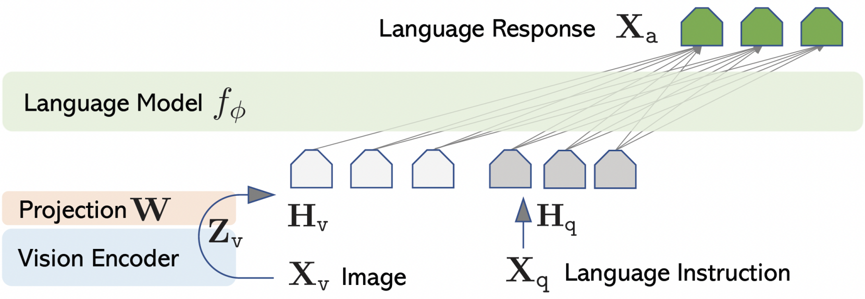


Figure 2.1: Diagram of the LLaVA architecture (figure from [23]).

This work primarily evaluates the most recent LLaVA-based models, which achieve state-of-the-art multimodal performance among open-source decoder-only VLMs. Drawing on the success of instruction fine-tuning in LLMs, LLaVA (Large Language and Vision Assistant) is among the first VLM frameworks to employ visual instruction tuning to enhance multimodal performance [23]. The original LLaVA model integrates CLIP [30] as a vision encoder and Vicuna [7] as the language backbone. To align the visual and textual representations, LLaVA is trained on large-scale GPT-4 generated multimodal instruction-following data for general-purpose language and visual understanding.

Since its initial release, several subsequent versions of LLaVA have emerged. These include LLaVA-1.5 [21] which improved upon the first version of LLaVA by using a fully connected multi-layer perceptron (MLP) layer and academic-focused VQA data, BakLLaVA [32] which is LLaVA-1.5 with a Mistral-7B [16] LLM backbone, LLaVA-RLHF [33] which employs Factually Augmented RLHF to align the model with human preferences to have reduced hallucinations and increased helpfulness, and LLaVA-NeXT (v1.6) [22] trained on high-quality user instruct data to achieve enhanced visual reasoning, logical reasoning, optical character recognition, and world knowledge.

While previous work such as [28] evaluates the degree of multimodality of earlier LLaVA-based models including BakLLaVA and LLaVA-NeXT, this work focuses on two of the most recent LLaVA variants: Pangea [38] and LLaVA-OneVision [18]. Both models integrate Qwen2 as the multilingual LLM backbone, which demonstrates strong proficiency in 30 different languages. Detailed descriptions of Pangea and LLaVA-OneVision are provided in section 3.2.

2.2 Multilingual Visual Question Answering

While there are various vision-and-language tasks such as image captioning, multimodal chat, and multimodal reasoning, this work focuses on visual question answering (VQA). In VQA, the model is given an image and a question about that image as input, and tasked with generating an answer to the question. While multimodal chat and image captioning can be subjective, VQA datasets contain ground truth answers which provide an efficient way to evaluate the model performance. In addition, there are multiple high-quality multilingual datasets available for this specific task.

Given that our intended analysis spans different languages, the chosen dataset needs to comprise a diverse set languages, ideally covering both high-resource and low-resource settings. Several multilingual VQA datasets were taken into consideration for this purpose, including xGQA, MaXM, MTVQA, and CVQA. xGQA [29] is a multilingual evaluation benchmark derived from the original English-only GQA dataset [15]. xGQA contains manually translated questions in 8 typologically diverse languages, providing high-quality parallel sentences for each image. MaXM [6] is a test-only VQA benchmark in 7 diverse languages, using a novel translated-based framework for large-scale multilingual data creation. MTVQA [35] is designed for text-centric VQA, in which the questions focus on the textual content in the images. This dataset consists of high-quality human annotations in 9 languages for 6,778 question-answer pairs and 2,116 images. CVQA [31] is the most culturally diverse, comprising culturally-driven images and questions from 30 countries and 31 languages, resulting in 10k question-answer pairs.

Ultimately, the xGQA benchmark is selected as the main dataset for this study as it presents several notable advantages over the alternatives, such as the availability of parallel sentences in both high-resource and low-resource languages. We provide more details on the xGQA dataset in section 3.1.

2.3 Measuring the Degree of Multimodality

The degree of multimodality refers to the extent to which a multimodal model relies on each of its modalities. Various methods have been proposed to measure and compare the unimodal and cross-modal influences in VLMs. However, the previous studies have applied these methods to English-only datasets, thus raising the need to extend the evaluation to the multilingual setting.

Cross-modal ablation is employed in [10] by masking the input from one modality either partially or entirely, and evaluating the model performance in predicting the masked data in the other modality. The findings indicate that encoder-only VLMs, such as LXMERT [34] and ViLBERT [24], are impacted more by masked images than masked texts. Although effective in measuring the degree to which the models exploit cross-modal information, this method assesses the model performance in modality-specific pretraining tasks, such as masked language modelling for text, and thus is not suitable for multimodal tasks such as VQA.

DIME [25] disentangles the model predictions into unimodal text contribution, unimodal image contribution, and multimodal interactions, enabling a fine-grained analysis of multimodality. DIME explanations demonstrate that certain encoder-only models, such as LXMERT, rely heavily on unimodal text contributions. However, the current implementation of DIME is only applicable to the classification task and to encoder-only VLMs, making it unsuitable for this study.

Text Preference Ratio [9] measures the model's preference for text-based over image-based predictions, but requires textual descriptions of the images in addition to the input questions, which are not readily available in most multilingual VQA datasets. Evaluation of recent proprietary and open-source decoder-only VLMs reveal the "blind faith in text" phenomenon, where the models predominantly depend on the textual descriptions over the visual inputs when the two are misaligned.

Given these constraints, we select and compare two alternative methods to assess the degree of multimodality in multilingual VLMs: the **perceptual score** [11] and **MM-SHAP** [27]. While both measure the reliance of VLMs on their modalities, their methodologies reveal different insights and model biases. These methods are discussed in further detail in section 3.3.

This study builds upon the work in [28], which adapts MM-SHAP to decoder-only VLMs and evaluates the degree of multimodality of three earlier LLaVA-based models – BakLLaVA, LLaVA-NeXT-Mistral, and LLaVA-NeXT-Vicuna – using English-centric visual question answering (VQA) datasets. The findings from the previous study indicate that the textual contributions are much higher than the visual contributions in all decoder-only VLMs across all examined datasets. In this work, we extend the analysis to diverse languages by using a multilingual dataset and assess two recent LLaVA variants – Pangea [38] and LLaVA-OneVision [18].

In addition, we extend the work in [11] which proposes the perceptual score as a novel method to quantify the impact of each modality towards the model performance. The results demonstrate the trend in which the most recent, high-performing VLMs increasingly rely on the textual input to a greater degree than the earlier models, sparking concerns that the models are not sufficiently leveraging the visual input. While this previous work assesses the modality reliance in encoder-only VLMs using English-centric datasets, this work evaluates the perceptual score of the latest decoder-only VLMs across a diverse set of languages.

Chapter 3

Method

This chapter introduces the multilingual benchmark (xGQA), the two models (LLaVA-OneVision and Pangea), and the two complementary methods (the perceptual score and MM-SHAP) employed in this work. In addition to providing their descriptions, we discuss why these models, dataset, and methods are suitable for this study.

3.1 Dataset

xGQA is an evaluation benchmark extended from the balanced test-dev set of the monolingual English-only GQA dataset [15], designed to assess multimodal capabilities on the VQA task. The original GQA dataset consists of 12,578 questions in English (EN) pertaining to 398 images, with the short-form answers provided as single words or short phrases. In xGQA, the questions have been manually translated into an additional 7 diverse languages covering 5 scripts: Bengali (BN), German (DE), Indonesian (ID), Korean (KO), Portuguese (PT), Russian (RU), and Chinese (ZH) – comprising languages in both high-resource and low-resource settings. Hence, the resulting dataset contains 12,578 questions in each of the 8 languages. However, the answers have been not translated and thus are only provided in English, requiring the model predictions to be translated from the target languages into English for evaluation.

Among the open-domain multilingual VQA datasets available, xGQA offers important benefits in its quality and suitability for this study. While MaXM [6] provides questions and answers in a diverse set of languages, its data creation process heavily relies on machine translation which could introduce *translationese* [13], which refers to undesirable features in the translated text. xGQA instead uses manual translation from English into the target languages by their native speakers, ensuring high-quality,

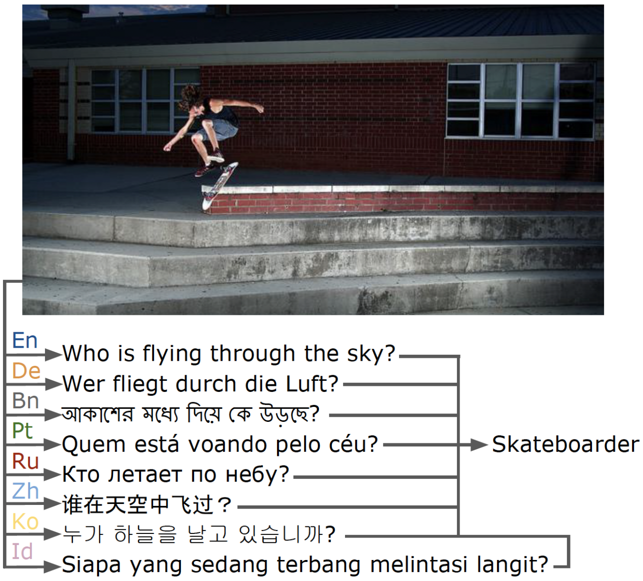
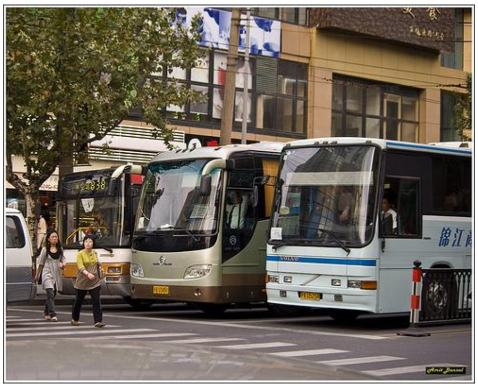

En
Who is flying through the sky?
De
Wer fliegt durch die Luft?
Bn
আকাশের মধ্যে দিয়ে কে উড়ছে?
Pt
Quem está voando pelo céu?
Ru
Кто летает по небу?
Zh
谁在天空中飞过？
Ko
누가 하늘을 날고 있습니까?
Id
Siapa yang sedang terbang melintasi langit?
Skateboarder

Figure 3.1: An example from the xGQA dataset with parallel sentences in 8 languages (figure from [29]).

natural-sounding translations. The MTVQA dataset, while multilingual, is specifically tailored to text-centric VQA which limits the scope of general multimodal analysis. CVQA contains culturally-driven images and questions which could bias the models to demonstrate stronger performances in languages that are culturally relevant to the content of the images and questions. In contrast, xGQA contains culturally neutral images, reducing the risk of such cultural biases and enabling a fair evaluation across the multiple languages.

A major advantage of xGQA is that it contains parallel sentences, where identical questions are provided in 8 different languages, as illustrated in Figure 3.1. This allows for direct cross-language comparison, while the other multilingual datasets can have varying questions for different languages. Another benefit of xGQA is that the questions are classified into 5 structural types: *verify* for simple yes/no questions, *logical* for complex yes/no questions involving and/or operations, *query* for open-ended questions, *choose* for selections between two options, and *compare* for comparisons between two or more objects. This categorization enables the evaluation of the degree of multimodality across different question types in addition to the different languages. These question types differ in complexity, as yes/no and selection type questions are straightforward while open-ended questions are more challenging. Examples of the 5 structural types are provided in Figure 3.2.



Verify: Does the person in front of the other person appear to be sitting? *no*

Logical: Is the bus to the right of the other bus red and large? *no*

Query: What is in front of the telephone pole? *bus*

Choose: Where in the photo is the person, on the right or on the left? *left*

Compare: Are the two people of the same gender? *yes*

Figure 3.2: Examples of the 5 question types in xGQA with the corresponding image.

3.2 Models

Both Pangea [38] and LLaVA-OneVision [18] follow the LLaVA-NeXT architecture with a LLM backbone, a vision encoder, and a trainable projector which maps the image features to the text embedding space. However, these two models differ significantly in their training data and fine-tuning strategies.

Pangea is a multilingual, multimodal VLM trained on a large-scale dataset of 6M instructions spanning 39 typologically diverse languages and 1 million culturally varies images. The training corpus consists of high-quality English instructions, carefully machine-translated instructions, and culturally diverse multimodal tasks, designed to address under-represented languages and cultural contexts. The model uses Qwen2-7B-Instruct [37] as the LLM backbone and CLIP-ViT-L/14@336px [30] as the vision encoder. Pangea outperforms its open-source counterparts of similar size in multilingual settings while maintaining strong capabilities in English when evaluated on various vision-and-language tasks such as multimodal chat, image captioning, and visual question answering.

Similar to Pangea, **LLaVA-OneVision** employs a Qwen2 LLM backbone [37]. However, the model uses SigLIP [39] as the vision encoder due to its strong multimodal performance. Visual instruction tuning for LLaVA-OneVision consists of two stages: single-image training and OneVision training. In the first stage, the model is trained on 3.2M single-image instructions from various visual tasks. Then during OneVision training, the model is further fine-tuned on a combination of single-image, multi-image, and video data to enhance its visual understanding capabilities. LLaVA-OneVision demonstrates performance comparable to, or exceeding, that of previous open-source VLMs on various single-image, multi-image, and video benchmarks.

3.3 Methods

3.3.1 Perceptual Score

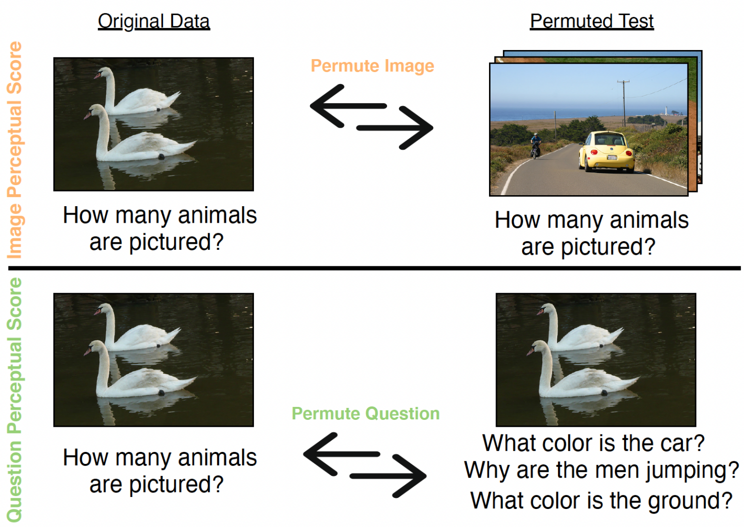

Original Data
Permuted Test
Image Perceptual Score
Permute Image
How many animals are pictured?
How many animals are pictured?
Question Perceptual Score
Permute Question
How many animals are pictured?
What color is the car?
Why are the men jumping?
What color is the ground?

Figure 3.3: Permutation test applied to the images and texts (figure from [11]).

The perceptual score is a conceptually simple yet effective method to assess a model's reliance on each modality with respect to the model performance. It addresses the concern that VLMs may be biased towards certain modalities, essentially taking "shortcuts" by using only one modality to perform a given task instead of integrating multimodal information. The method quantifies the reliance on a specific modality by measuring the degradation in model performance when features from that modality are randomly permuted during evaluation. Permutation of one modality minimizes its impact while maintaining the functionality of other modalities. Formally, given a dataset D, a multimodal model f, and a set of all modalities $\mathcal{M}$, the perceptual score of f towards the modality M_m is defined as

$$P_{f,D}(M_m) = \mathrm{Acc}_{f,D}(\mathcal{M}) - \mathrm{Acc}_{f,D}(\mathcal{M} \setminus \{M_m\}) \tag{3.1}$$

In this equation, $\mathcal{M} \setminus \{M_m\}$ effectively removes the influence of the modality M_m from the set of all modalities $\mathcal{M}$. Intuitively, the perceptual score $P_{f,D}(M_m)$ is higher if removing the influence of the modality M_m leads to a greater drop in the model accuracy, reflecting a stronger dependence on that modality. $P_{f,D}(M_m)$ inherently consists of both the contribution of the modality M_m as well as the model accuracy. To minimize the confounding effect of overall accuracy, the perceptual score is normalized by the model accuracy for the given task

$$Z_{f,D} = \mathrm{Acc}_{f,D}(\mathcal{M}) \tag{3.2}$$

in order to obtain a score that more directly reflects the reliance on a particular modality instead of absolute performance.

In previous work by [11], the perceptual score was computed for four encoder-only VLMs – BUTD [3], BAN [17], LMH [8], and LXMERT [34] – on the English-only VQAv2 [12] and VQA-CP [1] datasets. In most cases, the results indicate that these models consistently rely on the textual input to a greater degree than the image. This imbalance is more pronounced in newer models than their predecessors.

In this work, we extend the analysis of [11] by assessing the perceptual scores on state-of-the-art open-source decoder-only VLMs across multiple question types and languages. While the perceptual score is computationally efficient and easy to interpret, it has certain drawbacks: it requires the models to have reasonable accuracy on the target task and it does not provide the contribution scores at the token level for more fine-grained analysis. To address these limitations, we employ MM-SHAP as an additional method to complement the perceptual score.

3.3.2 MM-SHAP

Inspired by Shapley values, MM-SHAP is designed to quantify the contribution of each text and image input token towards the model prediction, independent of its correctness. This is performed by iteratively masking subsets of the input tokens and measuring the change in the probability of the output tokens in the predicted sequence. In this case, the text tokens correspond to the tokenized subwords in the instruction, while the image tokens are constructed by dividing the raw image pixels into square patches, as shown in Figures 4.3 and 6.4. To ensure a fair comparison between modalities, the number of image patches is dynamically computed to be approximately equal to the number of text tokens. Thus, longer instructions resulting in more text tokens would correspond to more, smaller image patches while short instructions would lead to fewer and larger patches.

Each input token j is assigned a contribution score corresponding to each token in the output sequence T, denoted as $\{\phi_j^1, \phi_j^2, \ldots, \phi_j^T\}$. These token contribution scores are Shapley values which can be positive, negative, or zero, based on whether they increase, decrease, or have no effect on the probability of the predicted output sequence.

To ensure a fair comparison among the output tokens, the scores of each input token

j for each output token t are normalized as contribution ratios r^t_j:

$$r^t_j = \phi^t_j / \sum_i^N |\phi^t_i| \tag{3.3}$$

The contribution ϕ_j for each input token j towards the full output sequence T is computed as the average of the contribution ratios of token j over all output tokens t:

$$\phi_j = \sum_{t=0}^T r^t_j / T \tag{3.4}$$

Given N_T input text tokens and N_I input image tokens, the the modality-level contribution scores are computed as the sum of the contribution of the tokens from that modality:

$$\Phi_T = \sum_j^{N_T} |\phi_j|; \Phi_I = \sum_j^{N_I} |\phi_j| \tag{3.5}$$

The MM-SHAP score consists of the textual degree T-SHAP and visual degree V-SHAP, which are proportional to each other:

$$\text{T-SHAP} = \frac{\Phi_T}{\Phi_T + \Phi_I}; \text{V-SHAP} = \frac{\Phi_I}{\Phi_T + \Phi_I} \tag{3.6}$$

MM-SHAP [27] was initially implemented to measure the degree of multimodality in encoder-only VLMs such as LXMERT [34] and ALBEF [19], using vision-and-language datasets including MSCOCO [20], VQA [4], and GQA [15]. The method was later adapted for decoder-only VLMs [28] and evaluated on BakLLaVA, LLaVA-NeXT-Mistral, and LLaVA-NeXT-Vicuna using the same datasets. The results from these studies demonstrate that while encoder-only VLMs exhibit a relatively balanced reliance on the text and vision modalities, decoder-only VLMs have a strong dependence on the text modality compared to the vision modality. While previous MM-SHAP studies have focused on English-centric datasets, in this work we apply the method to the most recent decoder-only VLMs across multiple high-resource and low-resource languages in the xGQA dataset.

Chapter 4

Experimental Setup

The implementation in this work is an extension of the MM-SHAP code[1] for decoder-only VLMs [28]. In this chapter, we describe the experimental design for multilingual evaluation, modality permutations for the perceptual score, and MM-SHAP computations.

For both the perceptual score and MM-SHAP, we conduct experiments using two recent models – Pangea and LLaVA-OneVision – across the 8 languages and 5 question types in the xGQA benchmark, resulting in 80 configurations for each method. For the perceptual score, we evaluate using 1000 samples per language and question type. Since MM-SHAP is computationally expensive, we limit the evaluation to 20 samples per language–question type pair, corresponding to approximately 1-2 hours per run depending on the language.

4.1 Multilingual Evaluation

While the original MM-SHAP code is designed to process monolingual inputs using previous decoder-only VLMs such as BakLLaVA and LLaVA-NeXT, we extend the implementation to multiple languages, different question types, and more recent models. The multilingual evaluation pipeline is illustrated in Figure 4.1.

Since the xGQA dataset consists of short-form answers, we use a simple zero-shot prompt to instruct the model to generate a single word or short phrase as a response. This prompt is translated into the 8 languages in xGQA to guide the model to generate a prediction in the target language instead of English. As this instruction translation only needs to be performed once for each language, we employ Google Translate which has

[1] https://github.com/Heidelberg-NLP/CC-SHAP-VLM

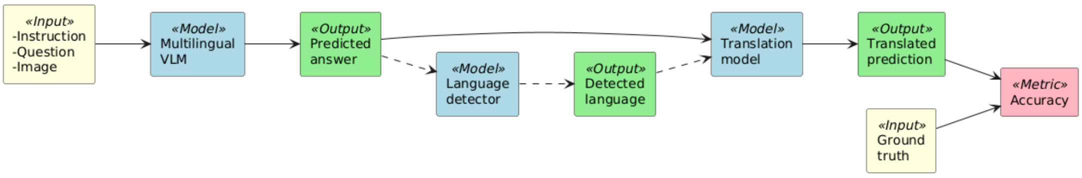


Figure 4.1: Diagram of the multilingual evaluation pipeline.

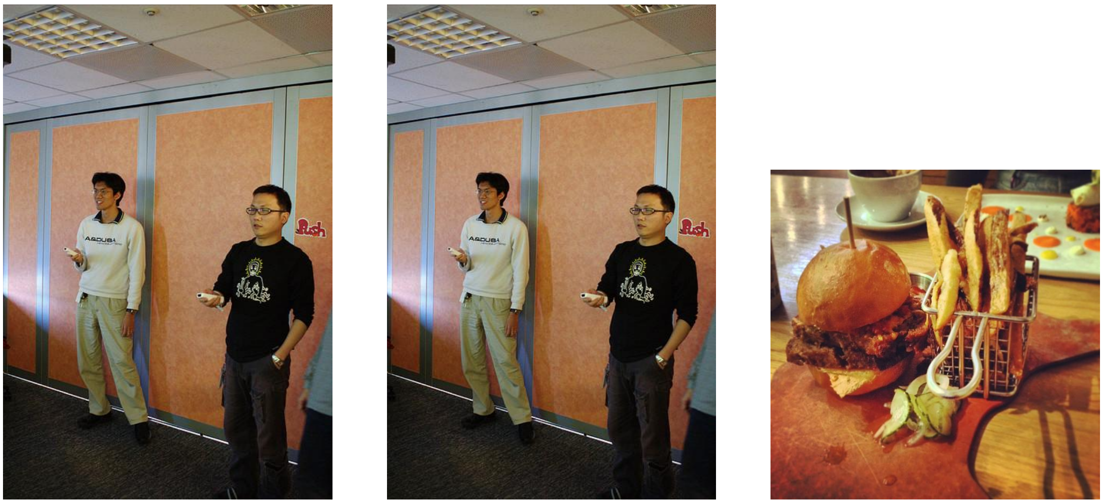
high translation quality in various languages.

The default decoding strategy uses top-k sampling with `k`=50 and `temperature`=1.0. As described in section 5.1, we explore using greedy encoding, which results in improved model accuracy. The xGQA dataset consists of short, factual answers, making greedy encoding advantageous as it leads to more deterministic answers. Moreover, greedy encoding allows for increased reproducibility which is beneficial for our analysis. In the subsequent experiments, we employ greedy encoding for generating model predictions, unless otherwise specified.

As the ground truth answers in xGQA are only provided in English and not in the target languages, we need to translate the model predictions from the target languages into English to evaluate the correctness. For this, we utilize NLLB-200-3.3B [36], which achieves state-of-the-art results among open-source translation models. Accuracy is computed using exact match such that the model prediction needs to contain the ground truth answer to be considered correct. We acknowledge that exact match is a stringent measure as it does not account for synonyms or paraphrasing, however we consider it to be sufficient since the answers in the dataset are short and simple.

To assist in the translation step, we experiment with using a language detection model to identify the language of the model prediction as additional input to the translation model, NLLB-200-3.3B. This is to address the phenomenon of *language infidelity* [14], in which LLaVA models exhibit the tendency to generate responses about an image in English even though the input question is in another language. We employ Lingua[2] as the language detector due to its strong performance on short texts and its support for limiting detection to a predefined set of languages. However, as shown in section 5.1, empirical results indicate that using the automatically detected language as additional input does not improve translation quality and degrades the model accuracy, likely due to the propagation of language detection errors. Therefore, we later remove the language detector from the pipeline.

[2] https://github.com/pemistahl/lingua-py



Q: What is the short person holding, a remote control or a phone?
Pred: Remote control

(a) Original image and text

Q: In which part of the image is the mug, the bottom or the top?
Pred: Bottom

(b) Permuted text

Q: What is the short person holding, a remote control or a phone?
Pred: Phone

(c) Permuted image

Figure 4.2: An example of permutation of the textual and vision modalities applied to the *choose* type question.

4.2 Perceptual Score Permutations

For experiments with the perceptual score, we first compute the baseline model accuracy $Acc_{\mathcal{M}}$ for each language and question type when the original image and text are provided as inputs to the VLM as in Figure 4.2a.

Next, we apply permutation to the text and vision modality individually and compute the new model accuracy, $Acc_{\mathcal{M}\setminus\{T\}}$ and $Acc_{\mathcal{M}\setminus\{V\}}$. Permutation of the text modality is performed by shuffling the questions and randomly sampling a question from the same distribution – same language and question type – as the original question, as illustrated in Figure 4.2b. Similarly, permutation is applied to the vision modality by shuffling the images and randomly selecting an image associated with the same question type as the original image, as shown in Figure 4.2c. To ensure that the permutation is reproducible across the languages, we apply a fixed seed for shuffling and random sampling.

Permutation causes misalignment between the question and the image. When permutation is applied to the text, the model still has access to the original image but is given an unrelated question. Conversely, when the images are permuted, the model is given the original question but not the relevant visual content to answer the question.

4.3 MM-SHAP Scores and Image Patches

4.3.1 T-SHAP and V-SHAP Scores

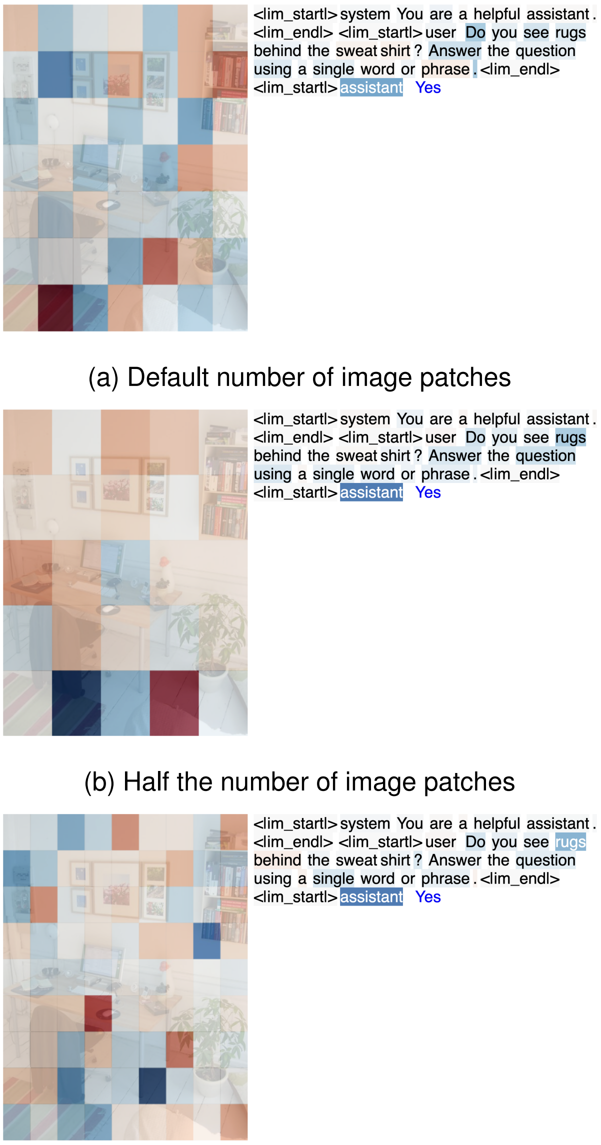
To compute the modality contribution scores with MM-SHAP, we first generate the predicted output sequence using the complete, original input question and image. Next, we obtain the text tokens by using the model's tokenizer to segment the input text into subwords. The image tokens correspond to square patches of raw pixels in the image. To ensure comparability between the proportional modality contribution scores, the number of image patches is dynamically calculated to closely match the number of text tokens for each sample.

Masking a text token involves replacing it by the whitespace token, while masking an image patch is applied by setting all of its pixel values to zero. When masking subsets of n input tokens, the number of possible combinations is 2^n, making computations very expensive. To improve computation efficiency in our experiments, we follow [27] and approximate the MM-SHAP scores by randomly sampling $2n+1$ token combinations. We then obtain the new model prediction logits using this masked input and measure the change in the probability of the original output sequence, obtaining the contribution scores for the input image and text tokens. To compute modality contributions, we aggregate the token contribution scores from each modality as in equation 3.5. The textual degree T-SHAP and visual degree V-SHAP are then the proportional contributions of the two modalities.

4.3.2 Varying Numbers of Image Patches

In addition to computing the textual and visual degrees, we perform experiments to assess the robustness of MM-SHAP to varying numbers of image patches. While the original setup ensures roughly equal numbers of text tokens and image patches, we explore increasing it to double the original number as well as reducing it by half, as illustrated in Figure 4.3. Given that the number of text tokens is unchanged, this essentially results in approximately half and twice the number of image tokens compared to the number of text tokens, respectively.

Since the modality contribution scores are an aggregated sum of their token contributions, the number of tokens could directly impact the modality scores. If the method is robust to the images patches, the visual degree V-SHAP and textual degree T-SHAP would remain stable across the different numbers of patches. In addition, we expect the method to assign positive and negative contribution scores to similar regions of the image and to the same text tokens.



<|im_start|>system You are a helpful assistant . <|im_end|> <|im_start|>user Do you see rugs behind the sweat shirt ? Answer the question using a single word or phrase . <|im_end|> <|im_start|>assistant Yes

(a) Default number of image patches

<|im_start|>system You are a helpful assistant . <|im_end|> <|im_start|>user Do you see rugs behind the sweat shirt ? Answer the question using a single word or phrase . <|im_end|> <|im_start|>assistant Yes

(b) Half the number of image patches

<|im_start|>system You are a helpful assistant . <|im_end|> <|im_start|>user Do you see rugs behind the sweat shirt ? Answer the question using a single word or phrase . <|im_end|> <|im_start|>assistant Yes

(c) Double the number of image patches

Figure 4.3: An example of MM-SHAP token contributions when varying the number of image patches.

Chapter 5

Results

This chapter presents the experiment results, starting with model selection based on multilingual performance and discusses pipeline improvements. After having selected the best two models based on their accuracy on the xGQA benchmark, we then evaluate their degree of multimodality using the perceptual score and MM-SHAP.

5.1 Model Selection

While the previous MM-SHAP study on decoder-only VLMs employed BakLLaVA and LLaVA-NeXT on English-only VQA datasets, our work extends this analysis by using 7 additional languages and two additional models – LLaVa-OneVision and Pangea. We evaluate the accuracy of 5 models to determine the models with strongest performance across the 8 languages in the xGQA dataset. To ensure comparability, this work uses the 7B parameter versions for all models. For each model, we assess the accuracy using 200 samples in each language and report the results in Table 5.1. By default, the model generation pipeline uses top-k sampling with `k`=50 and `temperature`=1.0, leading to more varied and creative responses.

While all 5 models achieve reasonable performance in English, their performance differs significantly in the other languages. We acknowledge that this could be partly because the model predictions in English do not require translation, while the predicted answers in the remaining 7 languages are susceptible to translation errors. Among the evaluated VLMs, the previous models – BakLLaVA, LLaVA-NeXT-Mistral, and LLaVA-NeXT-Vicuna – have poor performance in low-resource languages such as Bengali, Indonesian, and Korean. The more recent LLaVA-OneVision model attains reasonable performance across the languages except Bengali, likely because Bengali

Lang	BakLLaVA	LLaVA-NeXT-Mistral	LLaVA-NeXT-Vicuna	LLaVA-OneVision	Pangea
BN	1.5	2.5	1.5	8.0	22.0
DE	14.0	28.5	31.5	36.5	37.5
EN	56.5	42.5	45.5	59.0	55.0
ID	6.5	16.0	11.0	38.0	19.0
KO	11.0	10.5	5.5	27.5	36.5
PT	20.5	33.5	23.5	31.5	40.5
RU	23.5	16.0	13.0	27.5	35.0
ZH	23.5	31.0	21.0	40.5	36.5

Table 5.1: Initial accuracy scores (exact match) for BakLLaVA, LLaVA-NeXT-Mistral, LLaVA-NeXT-Vicuna, LLaVA-OneVision, and Pangea across 8 languages.

instructions are not present during its pretraining or visual instruction tuning steps. Pangea, fine-tuned on a diverse set of 39 languages which include the 8 languages in xGQA, achieves acceptable accuracy in all languages in the benchmark, demonstrating strong multilingual capabilities. Based on these initial results, we primarily employ LLaVA-OneVision and Pangea in our subsequent experiments with the perceptual score and MM-SHAP.

Having selected LLaVA-OneVision and Pangea as the main models for evaluation, we then apply certain modifications to the pipeline to improve accuracy. First, we remove top-k sampling and instead apply greedy decoding, making the prediction deterministic by always selecting the output token with the highest probability at each prediction step. Greedy encoding is more suitable for the xGQA dataset which consists of short, factual answers. Next, we improve the translation component by (1) removing the language detector as described in section 4.1, thus consistently translating the model predictions from the target language instead of dynamically identifying the language of each prediction, and (2) providing the question in the target language as additional context for the translation model alongside the predicted answer. Language detection is challenging for short texts and errors in language detection could propagate to the translation step. Similarly, translating a single word or a short phrase is difficult and thus providing a longer context helps to improve translation quality.

As reported in Table 5.2, these changes significantly enhanced accuracy for both LLaVA-OneVision and Pangea. Pangea demonstrates high performance across all languages, due to being fine-tuned on diverse multilingual and culturally diverse datasets. Despite not being trained specifically to be multilingual and multi-cultural, LLaVA-OneVision still performed well across the languages with the exception of Bengali as

Lang	LLaVA-OneVision	Pangea
BN	14.4	50.5
DE	49.6	48.4
EN	62.3	64.1
ID	46.8	42.9
KO	42.3	51.7
PT	47.5	55.1
RU	48.7	54.5
ZH	47.4	50.2

Table 5.2: Accuracy scores (exact match) for LLaVA-OneVision and Pangea with the improved pipeline across 8 languages.

this language is not present in the model's training data.

5.2 Perceptual Score

The image and text perceptual scores for LLaVA-OneVision and Pangea across the 5 question types and 8 languages are recorded in Table 5.3. Here, $Acc_{\mathcal{M}}$ is the baseline model accuracy in the default setup where the original text and image are provided to the model with any permutation. $Acc_{\mathcal{M}\setminus\{V\}}$ refers to the accuracy when permutation is applied to the images, and similarly, $Acc_{\mathcal{M}\setminus\{T\}}$ is the accuracy when the text modality is permuted. $P_V/Z_{f,D}$ and $P_T/Z_{f,D}$ are the normalized perceptual scores for the vision and text modalities, respectively, measuring the proportional degradation in model accuracy when the influence of a modality is removed. Intuitively, if the input from a given modality has a negligible impact on the model performance, its normalized perceptual score would be near zero and conversely, a modality with strong influence would have a normalized perceptual score close to 100.

For the perceptual scores to be meaningful, the baseline model accuracy $Acc_{\mathcal{M}}$ needs to be reasonably high. As LLaVA-OneVision has poor performance in Bengali in all question types and both models have relatively low accuracy in the *compare* questions, we will not discuss these results in detail but include them here for completeness.

According to the baseline accuracy scores $Acc_{\mathcal{M}}$ in Table 5.3, when the two modalities are unperturbed, both models achieve strong performance in *verify* (69.3% for LLaVA-OneVision, 74.0% for Pangea) and *logical* (61.3% for LLaVA-OneVision, 67.3% for Pangea) style questions which require validating a statement with either "yes"

	LLaVA-OneVision					**Pangea**				
		Image		Text			Image		Text	
Lang	$Acc_{\mathcal{M}}$	$Acc_{\mathcal{M}\setminus\{V\}}$	$P_V/Z_{f,D}$	$Acc_{\mathcal{M}\setminus\{T\}}$	$P_T/Z_{f,D}$	$Acc_{\mathcal{M}}$	$Acc_{\mathcal{M}\setminus\{V\}}$	$P_V/Z_{f,D}$	$Acc_{\mathcal{M}\setminus\{T\}}$	$P_T/Z_{f,D}$
					Verify					
BN	28.8	20.2	29.9	20.7	28.1	78.8	48.0	39.1	47.2	40.1
DE	76.5	45.1	41.0	44.6	41.7	65.7	57.3	12.8	54.3	17.4
EN	83.7	47.1	43.7	44.9	46.4	85.3	49.6	41.9	43.7	48.8
ID	73.9	45.2	38.8	43.4	41.3	57.9	56.1	3.1	55.1	4.8
KO	70.5	45.0	36.2	45.3	35.7	71.7	45.9	36.0	43.9	38.8
PT	75.8	47.6	37.2	44.9	40.8	80.8	48.6	39.9	45.7	43.4
RU	74.7	47.0	37.1	44.7	40.2	79.0	48.0	39.2	47.8	39.5
ZH	70.3	53.0	24.6	48.0	31.7	72.6	52.7	27.4	45.7	37.1
					Logical					
BN	26.4	25.3	4.2	22.8	13.6	70.0	54.1	22.7	49.0	30.0
DE	69.8	51.8	25.8	47.7	31.7	53.7	51.7	3.7	46.1	14.2
EN	80.3	56.1	30.1	49.5	38.4	79.4	57.2	28.0	49.3	37.9
ID	65.5	49.3	24.7	45.7	30.2	49.0	48.8	0.4	47.7	2.7
KO	61.5	49.6	19.3	49.0	20.3	71.8	52.0	27.6	50.7	29.4
PT	60.1	49.3	18.0	46.0	23.5	75.4	57.2	24.1	50.3	33.3
RU	65.7	52.1	20.7	47.2	28.2	75.3	54.6	27.5	51.3	31.9
ZH	61.2	48.5	20.8	47.5	22.4	63.7	51.1	19.8	49.2	22.8
					Query					
BN	2.3	2.0	13.0	1.1	52.2	36.5	9.6	73.7	3.8	89.6
DE	33.3	8.2	75.4	2.7	91.9	40.2	14.4	64.2	3.2	92.0
EN	48.7	12.6	74.1	3.4	93.0	51.1	16.7	67.3	3.9	92.4
ID	31.2	6.3	79.8	3.2	89.7	37.2	12.4	66.7	3.0	91.9
KO	28.2	5.6	80.1	4.1	85.5	39.3	12.6	67.9	3.6	90.8
PT	34.4	8.6	75.0	3.0	91.3	39.7	14.4	63.7	3.4	91.4
RU	34.8	8.5	75.6	2.7	92.2	39.3	13.5	65.6	2.5	93.6
ZH	37.7	10.5	72.1	3.2	91.5	37.8	10.2	73.0	4.4	88.4
					Choose					
BN	16.7	13.9	16.8	1.8	89.2	66.5	46.9	29.5	4.9	92.6
DE	64.5	44.7	30.7	4.4	93.2	70.0	47.6	32.0	4.8	93.1
EN	82.1	54.3	33.9	4.5	94.5	84.4	57.4	32.0	5.1	94.0
ID	60.2	41.6	30.9	4.4	92.7	65.6	47.4	27.7	4.2	93.6
KO	47.3	33.1	30.0	3.0	93.7	61.1	41.2	32.6	4.6	92.5
PT	66.4	45.8	31.0	4.4	93.4	65.5	49.7	24.1	4.5	93.1
RU	58.2	43.0	26.1	4.4	92.4	68.1	43.8	35.7	4.4	93.5
ZH	63.8	43.7	31.5	4.4	93.1	64.2	41.5	35.4	4.9	92.4
					Compare					
BN	12.6	11.7	7.1	11.9	5.6	28.6	27.5	3.8	18.3	36.0
DE	36.1	31.7	12.2	22.8	36.8	27.7	25.6	7.6	15.3	44.8
EN	39.6	34.4	13.1	23.2	41.4	40.2	36.0	10.4	23.0	42.8
ID	34.6	30.9	10.7	22.4	35.3	26.7	24.9	6.7	15.4	42.3
KO	27.9	27.9	0.0	22.1	20.8	38.3	34.5	9.9	23.7	38.1
PT	29.4	25.7	12.6	19.7	33.0	38.5	34.6	10.1	22.7	41.0
RU	33.7	31.3	7.1	22.9	32.0	38.1	33.7	11.5	22.5	40.9
ZH	27.8	25.5	8.3	15.8	43.2	28.4	26.3	7.4	20.1	29.2

Table 5.3: Normalized image and text perceptual scores, $P_V/Z_{f,D}$ and $P_T/Z_{f,D}$, for LLaVA-OneVision and Pangea across different languages and question types. $Acc_{\mathcal{M}}$ is the baseline accuracy without permutation, $Acc_{\mathcal{M}\setminus\{V\}}$ is the accuracy with permuted images, and $Acc_{\mathcal{M}\setminus\{T\}}$ is the accuracy with permuted texts.

or “no”, as well as *choose* type questions (57.4% for LLaVA-OneVision, 68.2% for Pangea) which involve selecting between two options. Accuracy is noticeably lower for *query* type questions (31.3% for LLaVA-OneVision, 40.1% for Pangea), as these free-form questions are more difficult to answer, especially in the zero-shot setting.

For both models, the normalized text perceptual scores $P_T/Z_{f,D}$ are consistently higher than the normalized image perceptual scores $P_V/Z_{f,D}$ across all languages and question types, indicating that the input question has a greater impact on the model performance than the input image. The differences between the text and image perceptual scores are relative small in *verify* and *logical* questions, but very pronounced in *choose* type question. For the open-ended *query* questions, both LLaVA-OneVision and Pangea have high perceptual scores for both modalities, implying that both the input text and image have high importance in these free-form questions.

There is no substantial difference in the perceptual scores for LLaVA-OneVision and Pangea. As LLaVA-OneVision was further fine-tuned on additional single-image, multi-image, and video data, we could expect it to generally have higher image perceptual scores than Pangea, but the results show that this varies based on the question type. For instance, image perceptual scores are generally higher for LLaVA-OneVision than Pangea in *query* type questions, but are lower for LLaVA-OneVision in *choose* type questions.

In addition, the experiment results show no notables differences in the perceptual scores among the 8 languages evaluated. While we could expect high-resource languages to have a higher text perceptual scores than low-resource languages, and likewise for the models to rely on the image more when given an input text in a low-resource language, we do not identify such patterns in these results.

5.3 MM-SHAP

5.3.1 T-SHAP and V-SHAP scores

For the second method, MM-SHAP, we begin by obtaining the textual degree T-SHAP and visual degree V-SHAP for LLaVA-NeXT-Mistral and LLaVA-NeXT-Vicuna, which are LLaVA-NeXT (v1.6) models with different LLM backbones. Given that previous work in [28] computed the MM-SHAP scores for these two models using English-only VQA datasets, we seek to compare against these results to ensure the correctness of our implementation and to extend the analysis to additional languages. Due to computation

	LLaVA-NeXT-Mistral				**LLaVA-NeXT-Vicuna**			
Lang	Acc	T-SHAP$_{all}$	T-SHAP$_c$	T-SHAP$_i$	Acc	T-SHAP$_{all}$	T-SHAP$_c$	T-SHAP$_i$
				Verify				
EN	50.0	93.5	93.0	94.0	60.0	89.5	89.1	90.0
ID	40.0	93.0	92.6	93.3	55.0	89.1	90.4	87.5
KO	25.0	93.6	93.9	93.5	5.0	89.8	93.6	89.6
RU	80.0	94.3	94.2	94.8	60.0	92.2	91.5	93.3
ZH	40.0	93.8	93.5	93.9	70.0	91.3	91.5	90.8
				Query				
EN	40.0	91.9	91.9	91.9	50.0	87.9	88.2	87.7
ID	0.0	92.9	N/A	92.9	0.0	89.6	N/A	89.6
KO	5.0	92.7	94.0	92.6	0.0	89.4	N/A	89.4
RU	10.0	93.9	92.8	94.1	10.0	93.1	91.6	93.3
ZH	20.0	91.6	91.1	91.7	20.0	88.3	89.1	88.1

Table 5.4: MM-SHAP results for LLaVa-NeXT-Mistral and LLaVA-NeXT-Vicuna using a subset of languages and question types.

constraints, we select a subset of simple and difficult question types – *verify* and *query* – as well as a subset of high-resource and low-resource languages – English, Indonesian, Korean, Russian, and Chinese. The T-SHAP scores for LLaVA-NeXT-Mistral and LLaVA-NeXT-Vicuna for these question types and languages are reported in Table 5.4. Here, T-SHAP$_{all}$ is the average textual degree across all samples evaluated, while T-SHAP$_c$ and T-SHAP$_i$ refer to the textual degrees for correct and incorrect cases, respectively. Since T-SHAP and V-SHAP are proportional modality contributions, the visual degree is simply V-SHAP $= 100\% -$ T-SHAP.

Both LLaVA-NeXT models demonstrate very high T-SHAP scores in both yes/no *verify* and open-ended *query* questions across the languages, indicating a strong reliance on the textual inputs. LLaVA-NeXT-Mistral has higher T-SHAP values than LLaVA-NeXT-Vicuna, aligning with previous findings from [28], which reported T-SHAP contribution scores of 96% for LLaVA-NeXT-Mistral and 89% for LLaVA-NeXT-Vicuna on the English-only balanced GQA dataset. While we can observe a difference in the textual degrees between the two models, the results do not reveal distinct patterns among the languages, question types, or correct and incorrect cases.

Having validated our MM-SHAP implementation by comparing with previous work, we then proceed with further experiments using the more recent decoder-only VLMs, LLaVA-OneVision and Pangea. The MM-SHAP results for these two models across the 5 question types and 8 languages are shown in Table 5.5 alongside the model accuracy.

As with Table 5.4, here T-SHAP$_{all}$ indicates the average textual degree over all samples, T-SHAP$_c$ corresponds to the correct cases, and T-SHAP$_i$ corresponds to the incorrect cases.

LLaVA-OneVision and Pangea have high T-SHAP scores, 87.7% and 91.7% on average respectively, indicating that both models rely predominantly on the textual input compared to the visual input. This result is in agreement with the previous findings from [28], in which LLaVA-based decoder-only VLMs are reported to have very high T-SHAP values and thus strong reliance on the textual modality. Among the two models, we observe that Pangea has higher T-SHAP scores than LLaVA-OneVision by 4.0% on average across the question types and languages, indicating that Pangea relies on the textual modality to a greater extent than LLaVA-OneVision. Conversely, LLaVA-OneVision relies on the vision modality to a greater degree than Pangea.

Among the question types, both models have slightly higher T-SHAP scores for yes/no questions (*verify* and *logical*) compared to open-ended questions (*query*), but the difference is small. Across the question types, model accuracy varies significantly while T-SHAP values remain stable, which aligns with [27] in that there is very low correlation between accuracy and MM-SHAP scores. While accuracy measures the model performance, MM-SHAP directly measures the degree of multimodality independent of the correctness. In addition, the T-SHAP$_c$ and T-SHAP$_i$ scores are similar for all question types and languages, demonstrating that the reliance on the text and vision modalities do not vary based on the correctness of the model predictions.

Across the languages, Pangea has similar T-SHAP scores for all the languages evaluated while LLaVA-OneVision has noticeably lower T-SHAP values in Bengali compared to the other languages. This is possibly because of the absence of training examples in Bengali during the pretraining and visual instruction tuning stages for LLaVA-OneVision, whereas for Pangea, all 8 languages are present in the multilingual, multi-cultural fine-tuning data.

5.3.2 Varying Numbers of Image Patches

We conduct additional experiments to evaluate MM-SHAP's sensitivity to the number of square image patches p used in the method and summarize the results in Table 5.6. Due to computation constraints, we limit this analysis to one model, Pangea, and two question types – *verify* and *query* – due to their contrast in difficulty.

Here, $\sqrt{p}$ is the average number of patches on one dimension of the image. T-

	LLaVA-OneVision				**Pangea**			
Lang	Acc	T-SHAP$_{all}$	T-SHAP$_c$	T-SHAP$_i$	Acc	T-SHAP$_{all}$	T-SHAP$_c$	T-SHAP$_i$
Verify								
BN	20.0	83.2	83.2	83.2	75.0	91.4	91.2	91.7
DE	70.0	89.2	89.7	88.2	60.0	92.8	92.6	93.1
EN	70.0	89.8	89.7	90.1	80.0	93.3	93.1	94.3
ID	70.0	89.2	89.3	89.0	60.0	92.8	93.1	92.2
KO	75.0	89.5	89.6	89.2	65.0	92.8	92.3	93.7
PT	65.0	88.1	88.6	87.1	70.0	91.9	91.7	92.4
RU	65.0	88.5	89.0	87.7	80.0	93.3	93.3	93.6
ZH	70.0	89.0	89.0	89.2	70.0	91.5	91.5	91.3
Logical								
BN	20.0	82.9	82.8	82.9	50.0	91.5	91.5	91.6
DE	70.0	88.1	88.3	87.6	35.0	92.3	92.8	92.0
EN	90.0	90.1	90.4	87.3	75.0	92.3	92.7	91.0
ID	75.0	88.4	88.1	89.3	30.0	92.2	92.8	91.9
KO	60.0	88.7	88.7	88.7	60.0	92.7	92.7	92.6
PT	45.0	89.5	89.8	89.2	90.0	92.6	92.5	92.7
RU	70.0	88.0	88.1	87.8	85.0	92.9	93.1	92.2
ZH	55.0	89.0	89.0	89.0	55.0	90.8	91.1	90.5
Query								
BN	5.0	82.7	81.5	82.8	30.0	90.2	90.6	90.1
DE	35.0	87.4	87.7	87.3	40.0	90.6	90.3	90.2
EN	50.0	89.0	89.3	88.7	55.0	90.4	90.7	90.5
ID	30.0	87.9	89.1	87.4	25.0	90.7	90.6	90.7
KO	25.0	87.7	87.0	87.9	35.0	91.6	92.2	91.3
PT	30.0	87.9	87.7	88.0	45.0	91.2	91.2	91.2
RU	25.0	87.8	88.3	87.6	35.0	92.0	92.1	91.9
ZH	30.0	87.6	85.7	88.4	15.0	89.8	90.5	89.6
Choose								
BN	20.0	80.9	81.7	80.7	60.0	90.4	90.5	90.3
DE	65.0	87.4	87.8	86.8	65.0	91.6	91.7	91.3
EN	85.0	89.4	89.4	88.8	80.0	92.1	91.8	93.4
ID	70.0	87.8	87.8	87.7	80.0	91.7	91.7	91.3
KO	50.0	87.9	89.1	86.6	55.0	92.8	92.8	92.7
PT	55.0	88.0	87.7	88.2	45.0	91.1	90.1	91.8
RU	55.0	86.9	86.9	86.9	60.0	92.5	92.4	92.6
ZH	60.0	87.3	87.2	87.5	50.0	89.2	88.9	89.6
Compare								
BN	15.0	83.5	84.1	83.3	50.0	90.9	90.6	91.2
DE	65.0	88.3	88.0	88.8	50.0	91.9	91.9	91.9
EN	70.0	89.9	89.6	90.4	55.0	92.1	91.9	92.3
ID	70.0	87.3	87.8	86.2	65.0	92.7	92.9	92.5
KO	55.0	89.2	89.2	89.1	60.0	92.1	91.8	92.6
PT	55.0	90.1	90.4	89.7	60.0	92.5	92.2	92.9
RU	50.0	87.5	88.1	86.9	60.0	92.6	92.1	93.3
ZH	40.0	88.8	89.0	88.6	50.0	90.1	88.5	91.7

Table 5.5: MM-SHAP results for LLaVa-OneVision and Pangea across 8 languages and 5 question types.

SHAP$_p$ denotes the textual degree when the default number of image patches is used, which is approximately equal to the number of input text tokens. This is compared with T-SHAP$_{\frac{1}{2}p}$, when the number of patches is half the number of text tokens, and T-SHAP$_{2p}$, when the number of patches is doubled. Effectively, this results in using half and twice the number of image patches compared the number of text tokens, respectively.

In both question types and across all 8 languages, the textual degree decreases when the number of image patches is doubled (as shown by T-SHAP$_{2p}$ scores), but increases when the number of image patches is halved (as shown by T-SHAP$_{\frac{1}{2}p}$). Since the visual degree is V-SHAP $= 100\% -$ T-SHAP, this implies that the visual contribution is greater when more and smaller patches are used, and conversely, the visual degree is not as significant with fewer and larger image patches.

	Verify				**Query**			
	$\sqrt{p}$	T-SHAP$_{\frac{1}{2}p}$	T-SHAP$_p$	T-SHAP$_{2p}$	$\sqrt{p}$	T-SHAP$_{\frac{1}{2}p}$	T-SHAP$_p$	T-SHAP$_{2p}$
BN	11.2	93.9	91.4	87.4	10.9	93.1	90.2	86.0
DE	7.3	95.6	92.8	90.1	7.4	94.2	90.6	86.2
EN	6.8	96.2	93.3	89.7	7.0	93.9	90.4	86.6
ID	7.9	94.8	92.8	89.0	7.5	94.2	90.7	86.6
KO	7.3	95.6	92.8	88.8	7.3	94.5	91.6	88.0
PT	7.2	95.6	91.9	90.5	7.1	94.1	91.2	87.2
RU	7.5	95.6	93.3	90.2	7.5	94.6	92.0	88.5
ZH	6.7	94.6	91.5	87.1	6.5	92.5	89.8	83.9

Table 5.6: Varying T-SHAP scores of Pangea based on the number of image patches.

Chapter 6

Discussion

This chapter entails an in-depth analysis of the perceptual scores and MM-SHAP contributions from Chapter 5. While both methods indicate that the two VLMs have stronger reliance on the textual input than the visual input across all evaluated languages and question types, they provide different findings in the model behavior. The perceptual scores reveal the model's tendencies to make educated guesses by exploiting biases from its pretrained encoders in different question types, a phenomenon previously reported in [15]. In contrast, MM-SHAP results enable insightful comparisons between the models and languages, and offers an analysis of the contribution scores at the token level.

6.1 Perceptual Score

We do not observe a notable difference in the normalized perceptual scores $P_V/Z_{f,D}$ and $P_T/Z_{f,D}$ between the models and across the languages, with the exception of LLaVA-OneVision's perceptual scores in Bengali due to its low accuracy in this language. However, we can clearly identify a discrepancy among the question types for both models, indicating that the extent of the models' reliance of the inputs varies more strongly based on the type of questions than the languages. We conduct a deeper analysis across the different question types, which reveals strong biases in these models and the tendency to make educated guesses when permutation is applied to a modality.

6.1.1 *Verify* and *Query* Questions

We first analyze *verify* and *logical* type questions, which corresponds to simple and complex yes/no questions, respectively. The results in Table 5.3 show that for both

LLaVA-OneVision and Pangea, the text modality only has a slightly greater influence on the model performance than the vision modality in these question types. Moreover, Pangea demonstrates surprisingly low text and image perceptual scores in German and Indonesian compared to the other languages. Upon closer inspection, we discover tendencies in both models to take shortcuts and make guesses, a phenomenon previously discussed in [15].

When permutation is applied to one of the modalities causing misalignment of the text and image, both models exhibit strong biases in yes/no questions across all languages with the exception of Chinese. Pangea has a strong tendency to predict “yes” 85% of the time in German and 95% of the time in Indonesian, and predicts “no” 69% to 92% of the time in the other languages, as shown in Figure 6.1. Moreover, Pangea’s image and text perceptual scores for German and Indonesian are particularly low because the model tends to predict “yes” in these languages (89.5% of the time in German, and 97.3% in Indonesian) even when the visual and textual inputs are not permuted. Similar analysis indicates that LLaVA-OneVision is biased towards predicting “no” 72% to 86% of the time in all languages except Chinese. These biases likely arise from the pretraining data of the LLM backbone or during the visual instruction tuning stage of the VLM, causing the models to make guesses based on world tendencies [15]. These biases cannot be observed based on model accuracy alone, as the ground truth answers are quite balanced – 56% “yes” and 44% “no” – and thus the models can achieve acceptable accuracy by simply always predicting “yes” or always predicting “no”.

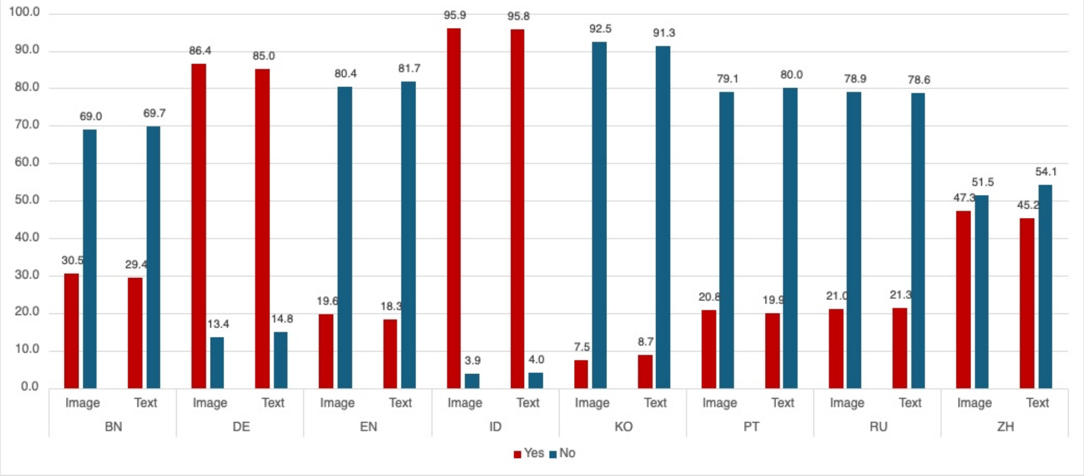


Figure 6.1: Pangea responses to *verify* type questions (yes/no) across languages when a modality is permuted.

We observe these tendencies in both *verify* and *logical* type questions, indicating that the models exhibit such biases and make guesses using the question priors irrespective of the simplicity or complexity of the yes/no questions. Thus, these types of questions might not be a suitable measure of the model's cross-modal understanding capabilities. The VLMs could be simply guessing based on the question priors and still attain reasonable accuracy scores.

6.1.2 *Choose* Questions

Across all 8 languages, both LLaVA-OneVision and Pangea demonstrate a remarkably stronger reliance on the input text than the input image in the *choose* type questions, in which the models are tasked to select between two alternatives. The average text perceptual score is 92.8% for LLaVa-OneVision and 93.1% for Pangea for this question type, while their respective average image perceptual scores are 28.9% and 31.1%. These results imply that the model performance degrades much more drastically when the VLMs receive a permuted question compared to a permuted image.

As illustrated in Figure 4.2b, when the question is permuted, the model still has access to the relevant image but is not presented with the appropriate options to select from in the input question. Hence, the model makes a guess based on the alternatives provided in the permuted question. However, when only the image is permuted as in Figure 4.2c, the model can still choose between the two options in the original question and thus has a decent chance of guessing correctly.

As with the yes/no questions, we can observe certain tendencies in the *choose* type questions. For instance, out of the 259 examples in which LLaVA-OneVision is tasked to choose between "left" or "right" and the image does not align with the question, the model guesses "right" in the majority of the cases in all languages except Chinese, as shown in Figure 6.2.

6.1.3 *Query* Questions

For open-ended *query* questions, LLaVA-OneVision and Pangea have average text perceptual scores of 85.9% and 91.3%, and average image perceptual scores of 68.2% and 67.8%, respectively. Hence, the models demonstrate strong reliance in both modalities, although the textual inputs still have a greater impact on the model performance than the visual inputs. *Query* questions are especially challenging, as they involve free-form, unrestricted answers instead of simply validating a statement or selecting between given options. As a result, the models have a much lower possibility of guessing correctly.

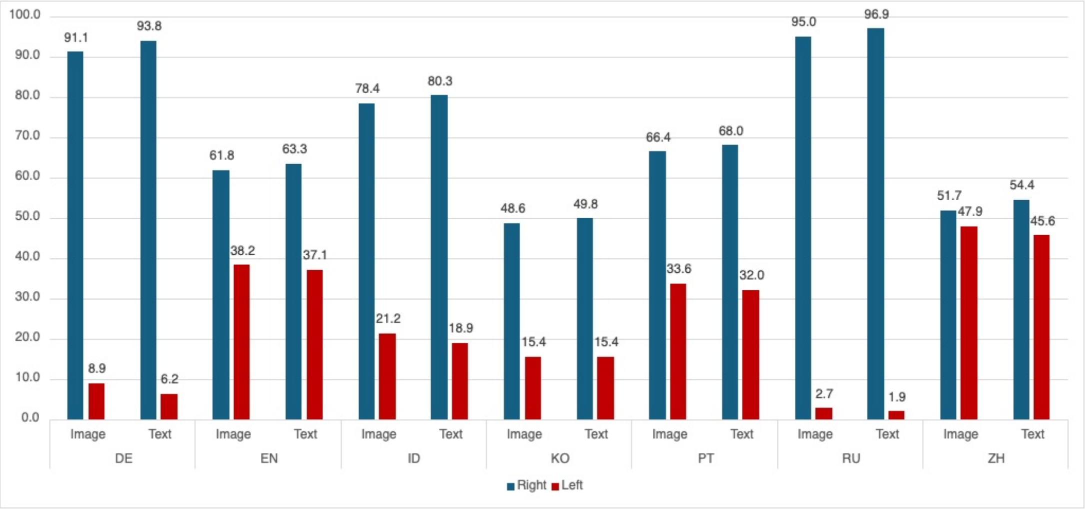


Figure 6.2: LLaVA-OneVision responses to *choose* type questions (right/left) across languages when a modality is permuted.

We investigate specific cases in the dataset and discover that certain model biases when the questions and images are misaligned. In particular, when asked "Who" type questions about the person or people in the image, both models are more likely to respond with "man"/"men" instead of "woman"/"women" in all languages, even though the ground truth answers contain more instances of "woman"/"women". This gender bias is especially pronounced in Pangea for German and Indonesian, as shown in Figure 6.3.

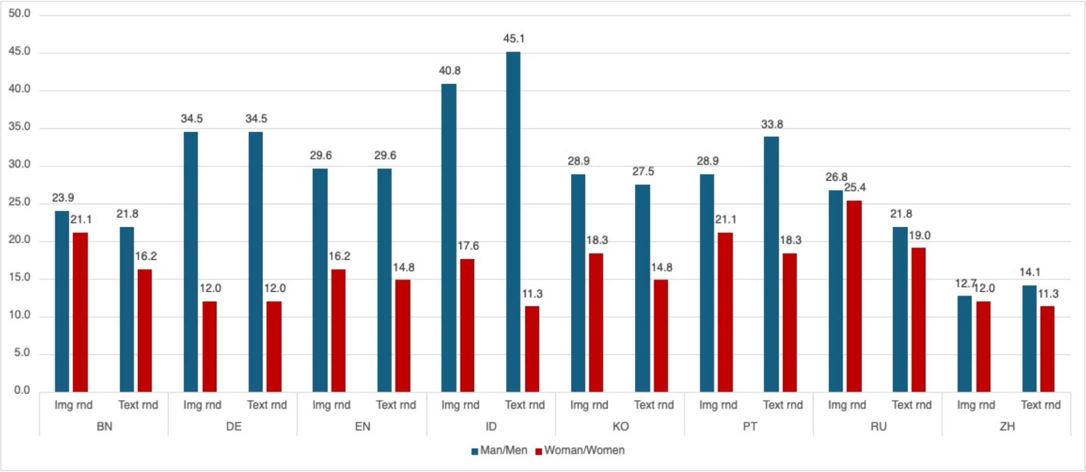


Figure 6.3: Pangea responses to *query* type questions ("Who") across languages when a modality is permuted.

6.2 MM-SHAP

6.2.1 Model Comparison

Between the two models evaluated, LLaVa-OneVision has lower T-SHAP scores than Pangea across all languages and question types, with a 4.0% difference on average. This is possibly due to LLaVA-OneVision's two-stage visual instruction tuning strategy, in which the model was further fine-tuned on a large-scale dataset of single-image, multi-image, and video data. Hence, the model has strong visual capabilities which could result in greater reliance on the vision modality. In contrast, Pangea's training objective focuses on its multilingual and multi-cultural performance instead of visual capabilities.

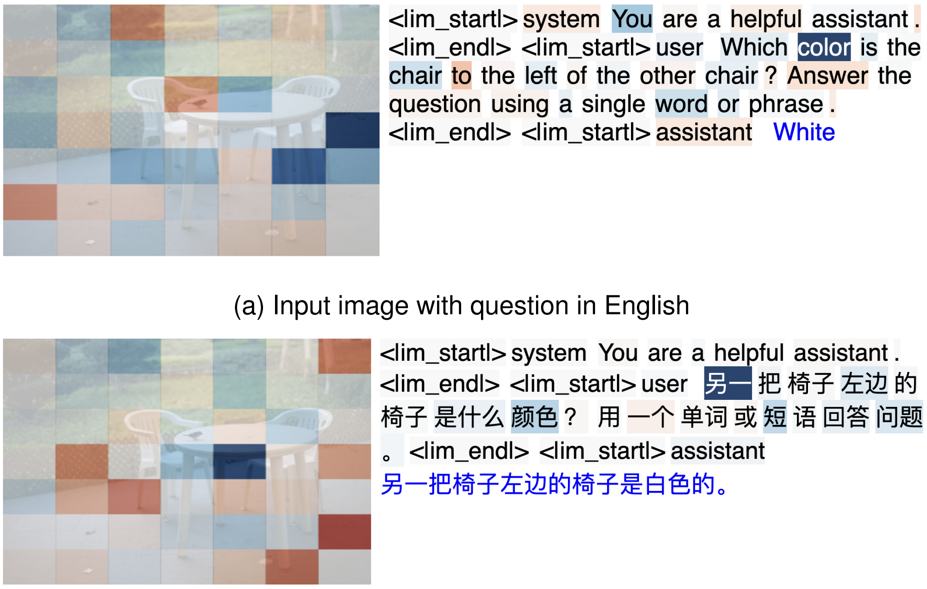
6.2.2 Language Comparison

Both LLaVA-OneVision and Pangea employ Qwen2 as the multilingual LLM backbone. Qwen2 achieves high proficiency in 30 diverse languages, including all languages in the xGQA except Bengali as this language was not present in the LLM's training data. Pangea is fine-tuned on multilingual multimodal instructions in 39 languages, including Bengali. Thus, Pangea is able to attain robust performance in Bengali during its multimodal instruction tuning stage, despite having a LLM backbone that is not trained on this language. In contrast, LLaVA-OneVision's visual tuning data does not consist of training examples in Bengali, leading to low performance in the language compared to the others.

These results suggest that the presence and the amount of training examples in a given language could affect the model's reliance on the text modality in that language. In this case, the lack of Bengali training data during LLaVa-OneVision's pretraining and fine-tuning stages causes the model to rely less on the text input in Bengali and more on the image compared to the other languages, as indicated by the lower T-SHAP scores in Bengali across the different question types. In contrast, LLaVa-OneVision has relatively high T-SHAP scores for English as this is the most common language in its training data, resulting in greater reliance on the textual inputs in English more than in other languages. On the other hand, Pangea is trained specifically to attain strong performance across different cultural settings and 39 different languages, addressing the concern that many state-of-the-art VLMs are English-centric. This could be the reason behind the comparable T-SHAP scores across the 8 languages for Pangea, indicating that for this model, the textual input has similar importance across the languages.

6.2.3 Token-level Contributions

After quantifying the modality contributions, we then conduct a more fine-grained analysis of the contribution scores of the input text and image tokens. More specifically, we want to assess if MM-SHAP is appropriately assigning highly positive scores to the text tokens and image patches that correspond to the main subject in the question, and negative scores to the background tokens. To this end, we visualize of the input tokens with their contribution scores. Figure 6.4 provides an example of such visualizations in English and in Chinese for the same image and question. The blue highlight indicates positive token contribution scores and red highlight refers to negative contributions.



<|im_start|> system You are a helpful assistant . <|im_end|> <|im_start|> user Which color is the chair to the left of the other chair ? Answer the question using a single word or phrase . <|im_end|> <|im_start|> assistant White

(a) Input image with question in English

<|im_start|> system You are a helpful assistant . <|im_end|> <|im_start|> user 另一 把 椅子 左边 的 椅子 是什么 颜色 ？ 用 一个 单词 或 短 语 回答 问题 。 <|im_end|> <|im_start|> assistant
另一把椅子左边的椅子是白色的。

(b) Input image with question in Chinese

Figure 6.4: Visualizations of token-level contribution scores by MM-SHAP in different languages.

In both the English and Chinese examples, positive contribution scores are appropriately assigned to the the text tokens with key information such as “color”, “chair”, and “left”. However, the image patches with high contribution scores do not correspond to the location of the object in question. We construct multiple other visualizations and observe similar patterns: the highly positive token-level scores are assigned to the right text tokens, but not to the relevant image patches (see Figures A.1 and A.2 in the Appendix). Additionally, even though identical images and questions are provided to the same model, with the only difference being the languages of the question, MM-SHAP assigns very different contribution scores to the patches in the same image and to text tokens in the same question.

We hypothesize that the method has difficulty in identifying the most important image patches because the image patches employed in its contribution computations are different from those processed by the VLMs. MM-SHAP segments the raw image pixels into patches, whereas LLaVA-based VLMs divide the image into patches based on the resolution of the vision encoder and construct an embedding for each image patch [21]. Hence, the image tokens employed in MM-SHAP do not correspond to the embeddings used for the model prediction. Masking square patches of the raw pixels only indirectly modify the underlying image patch embeddings instead of completely removing their presence, thus potentially resulting in a disadvantage for the image contribution scores. Future work should explore aligning the image tokens in the method with the image patch embeddings to ensure a more accurate measure of the model's visual degree.

In addition, based on the results in Table 5.6 with varying numbers of image patches, the modality scores do not remain consistent as the number of image tokens is varied. As indicated by T-SHAP$_{\frac{1}{2}p}$ and T-SHAP$_{2p}$ values, the vision modality has a higher contribution when more and smaller image patches are used. Conversely, using fewer and larger image patches reduces the visual degree.

Furthermore, the distribution of the token contribution scores changes significantly with varying number of image patches. That is, the method assigns very different image token contribution scores to different regions of the image when the number of patches is increased or decreased, as illustrated in Figure 4.3. While the distribution of the text token scores is also affected, this happens to a lesser extent, as the method is still able to identify the important text tokens in the question such as "see" and "rugs". To enhance robustness, future work could investigate normalizing the modality contribution scores with respect to the number of text and image tokens.

Chapter 7

Conclusions

This work proposes a framework to measure the degree of reliance on the different modalities in decoder-only VLMs across diverse languages, addressing the lack of multimodal evaluation in the multilingual setting, especially in under-represented languages. Our study employs two complementary methods – the perceptual score and MM-SHAP – and compares across two models, 5 question types, and 8 high-resource and low-resource languages. Across all languages and question types considered in this work, the perceptual score results demonstrate that model performance is more reliant on the textual inputs than the visual inputs, while the MM-SHAP scores indicate that the text modality has a greater contribution to the model prediction than the vision modality.

Further analysis of the perceptual scores reveals that when a modality is permuted, causing a mismatch in the image and text, the models rely on world tendencies and exploit biases to make educated guess. In particular, for simple question types involving yes/no answers or selecting between two alternatives, the models are able to attain high accuracy in different languages by simply guessing based on the input questions. This suggests that such simple questions may not be reliable to evaluate the multimodal capabilities in VLMs.

MM-SHAP results suggest that the degree of reliance on the textual modality could depend on the fine-tuning strategies as well as the presence and amount of training data of a given language. Token-level experiments demonstrate that the extent of reliance on the vision modality is not robust to varying numbers of image patches, as using more patches results in higher visual degrees. Furthermore, deeper analysis and visualizations reveal that the image patches of raw pixels employed in MM-SHAP do not correspond to the underlying image patch embeddings constructed by the models, which is a potential direction for further work.

This work has certain limitations which could be explored in future studies. In this study, we only employ one multilingual dataset, one vision-and-language task, and a fixed model size of 7B parameters. Moreover, both models considered in our study leverage Qwen2 [37] as the LLM backbone, which could constrain the scope of the evaluation. In addition, there are specific challenges with the two methods used for measuring the degree of multimodality in VLMs. The perceptual score is dependent on model performance and is less informative when accuracy is low. For MM-SHAP, obtaining the contribution scores for each input token is computationally expensive, thus restricting the experiments to a small subset of samples.

Future work could apply the analysis to additional models with different LLM backbones as well as other datasets with different languages and question types. While this work focuses on visual question answering, further studies could employ evaluation benchmarks for other vision-and-language tasks such as image captioning and multimodal reasoning. An interesting direction of research would be to measure and compare the degree of multimodality across languages with varying model sizes by evaluating smaller and larger decoder-only VLMs of the same architecture. We only consider the text and single-image inputs in this study, but this analysis could be extended to other types of inputs such as multi-image and video data, which are included in LLaVA-OneVision's training data [18]. Furthermore, future work could investigate multimodal correlation in addition to measuring the unimodal contributions, as proposed in [25].

Bibliography

[1] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4971–4980, 2018.

[2] Nahid Alam, Karthik Reddy Kanjula, Surya Guthikonda, Timothy Chung, Bala Krishna S Vegesna, Abhipsha Das, Anthony Susevski, Ryan Sze-Yin Chan, SM Uddin, Shayekh Bin Islam, et al. Behind maya: Building a multilingual vision language model. *arXiv preprint arXiv:2505.08910*, 2025.

[3] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018.

[4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[5] Florian Bordes, Richard Yuanzhe Pang, Anurag Ajay, Alexander C Li, Adrien Bardes, Suzanne Petryk, Oscar Mañas, Zhiqiu Lin, Anas Mahmoud, Bargav Jayaraman, et al. An introduction to vision-language modeling. *arXiv preprint arXiv:2405.17247*, 2024.

[6] Soravit Changpinyo, Linting Xue, Michal Yarom, Ashish V Thapliyal, Idan Szpektor, Julien Amelot, Xi Chen, and Radu Soricut. Maxm: Towards multilingual visual question answering. *arXiv preprint arXiv:2209.05401*, 2022.

[7] Wei-Lin Chiang, Zhuohan Li, Ziqing Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al.

Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6, 2023.

[8] Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. *arXiv preprint arXiv:1909.03683*, 2019.

[9] Ailin Deng, Tri Cao, Zhirui Chen, and Bryan Hooi. Words or vision: Do vision-language models have blind faith in text? In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3867–3876, 2025.

[10] Stella Frank, Emanuele Bugliarello, and Desmond Elliott. Vision-and-language or vision-for-language? on cross-modal influence in multimodal transformers. *arXiv preprint arXiv:2109.04448*, 2021.

[11] Itai Gat, Idan Schwartz, and Alex Schwing. Perceptual score: What data modalities does your model perceive? *Advances in Neural Information Processing Systems*, 34:21630–21643, 2021.

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

[13] Yvette Graham, Barry Haddow, and Philipp Koehn. Translationese in machine translation evaluation. *arXiv preprint arXiv:1906.09833*, 2019.

[14] Musashi Hinck, Carolin Holtermann, Matthew Lyle Olson, Florian Schneider, Sungduk Yu, Anahita Bhiwandiwalla, Anne Lauscher, Shaoyen Tseng, and Vasudev Lal. Why do llava vision-language models reply to images in english? *arXiv preprint arXiv:2407.02333*, 2024.

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

[16] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel,

Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023.

[17] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. *Advances in neural information processing systems*, 31, 2018.

[18] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[19] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[21] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306, 2024.

[22] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llavanext: Improved reasoning, ocr, and world knowledge, 2024.

[23] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[24] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

[25] Yiwei Lyu, Paul Pu Liang, Zihao Deng, Ruslan Salakhutdinov, and Louis-Philippe Morency. Dime: Fine-grained interpretations of multimodal models via disentangled local explanations. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 455–467, 2022.

[26] Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. Valse: A task-independent benchmark for vision and language models centered on linguistic phenomena. *arXiv preprint arXiv:2112.07566*, 2021.

[27] Letitia Parcalabescu and Anette Frank. Mm-shap: A performance-agnostic metric for measuring multimodal contributions in vision and language models & tasks. *arXiv preprint arXiv:2212.08158*, 2022.

[28] Letitia Parcalabescu and Anette Frank. Do vision & language decoders use images and text equally? how self-consistent are their explanations? *arXiv preprint arXiv:2404.18624*, 2024.

[29] Jonas Pfeiffer, Gregor Geigle, Aishwarya Kamath, Jan-Martin O Steitz, Stefan Roth, Ivan Vulić, and Iryna Gurevych. xgqa: Cross-lingual visual question answering. *arXiv preprint arXiv:2109.06082*, 2021.

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[31] David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, et al. Cvqa: Culturally-diverse multilingual visual question answering benchmark. *arXiv preprint arXiv:2406.05967*, 2024.

[32] SkunkworksAI, LAION, and Ontocord. Bakllava. GitHub repository, 2023. `https://github.com/SkunkworksAI/BakLLaVA`.

[33] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*, 2023.

[34] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

[35] Jingqun Tang, Qi Liu, Yongjie Ye, Jinghui Lu, Shu Wei, Chunhui Lin, Wanqing Li, Mohamad Fitri Faiz Bin Mahmood, Hao Feng, Zhen Zhao, et al. Mtvqa:

Benchmarking multilingual text-centric visual question answering. *arXiv preprint arXiv:2405.11985*, 2024.

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

[37] Qwen Team. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

[38] Xiang Yue, Yueqi Song, Akari Asai, Seungone Kim, Jean de Dieu Nyandwi, Simran Khanuja, Anjali Kantharuban, Lintang Sutawika, Sathyanarayanan Ramamoorthy, and Graham Neubig. Pangea: A fully open multilingual multimodal llm for 39 languages. In *The Thirteenth International Conference on Learning Representations*, 2024.

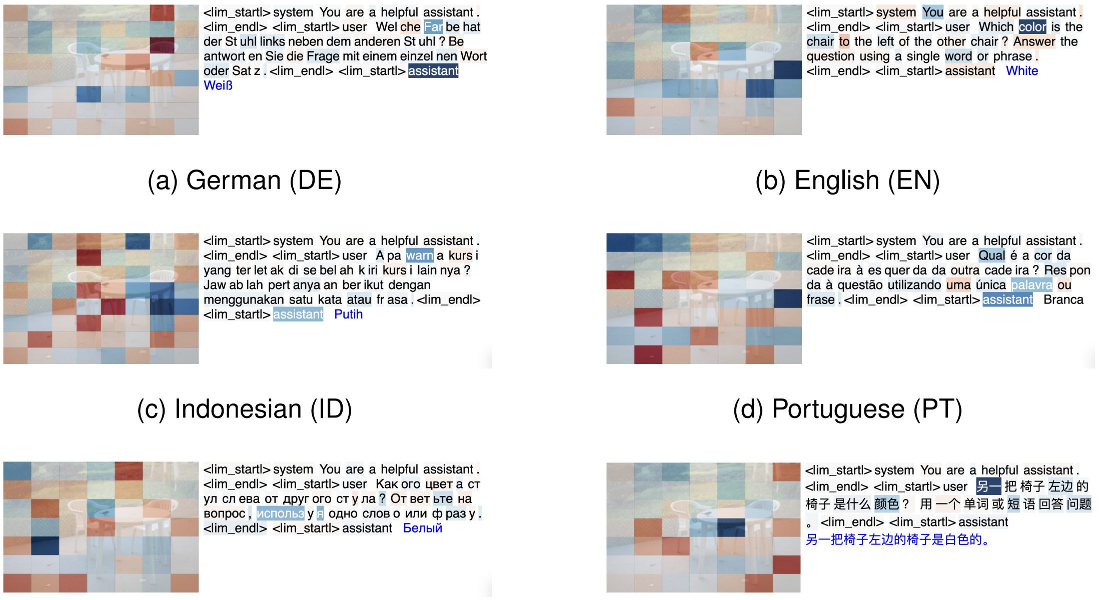
[39] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023.

Appendix A

Visualizations of MM-SHAP Scores

These visualizations for *query* and *verify* type questions show that MM-SHAP assigns contribution scores to image and text tokens differently with varying input languages. Qualitatively, the method seems to assign highly positive scores to the text tokens with important information in the input question but is unable to identify image patches that contain the object in question.

A.1 *Query* Questions



<lim_startl> system You are a helpful assistant . <lim_endl> <lim_startl> user Wel che Far be hat der St uhl links neben dem anderen St uhl ? Be antwort en Sie die Frage mit einem einzel nen Wort oder Sat z . <lim_endl> <lim_startl> assistant Weiß

(a) German (DE)

<lim_startl> system You are a helpful assistant . <lim_endl> <lim_startl> user Which color is the chair to the left of the other chair ? Answer the question using a single word or phrase . <lim_endl> <lim_startl> assistant White

(b) English (EN)

<lim_startl> system You are a helpful assistant . <lim_endl> <lim_startl> user A pa warn a kurs i yang ter let ak di se bel ah k iri kurs i lain nya ? Jaw ab lah pert anya an ber ikut dengan menggunakan satu kata atau fr asa . <lim_endl> <lim_startl> assistant Putih

(c) Indonesian (ID)

<lim_startl> system You are a helpful assistant . <lim_endl> <lim_startl> user Qual é a cor da cade ira à es quer da da outra cade ira ? Res pon da à questão utilizando uma única palavra ou frase . <lim_endl> <lim_startl> assistant Branca

(d) Portuguese (PT)

<lim_startl> system You are a helpful assistant . <lim_endl> <lim_startl> user Как ого цвет а ст ул сл ева от друг ого ст у ла ? От вет ьте на вопрос , использ у я одно слов о или ф раз у . <lim_endl> <lim_startl> assistant Белый

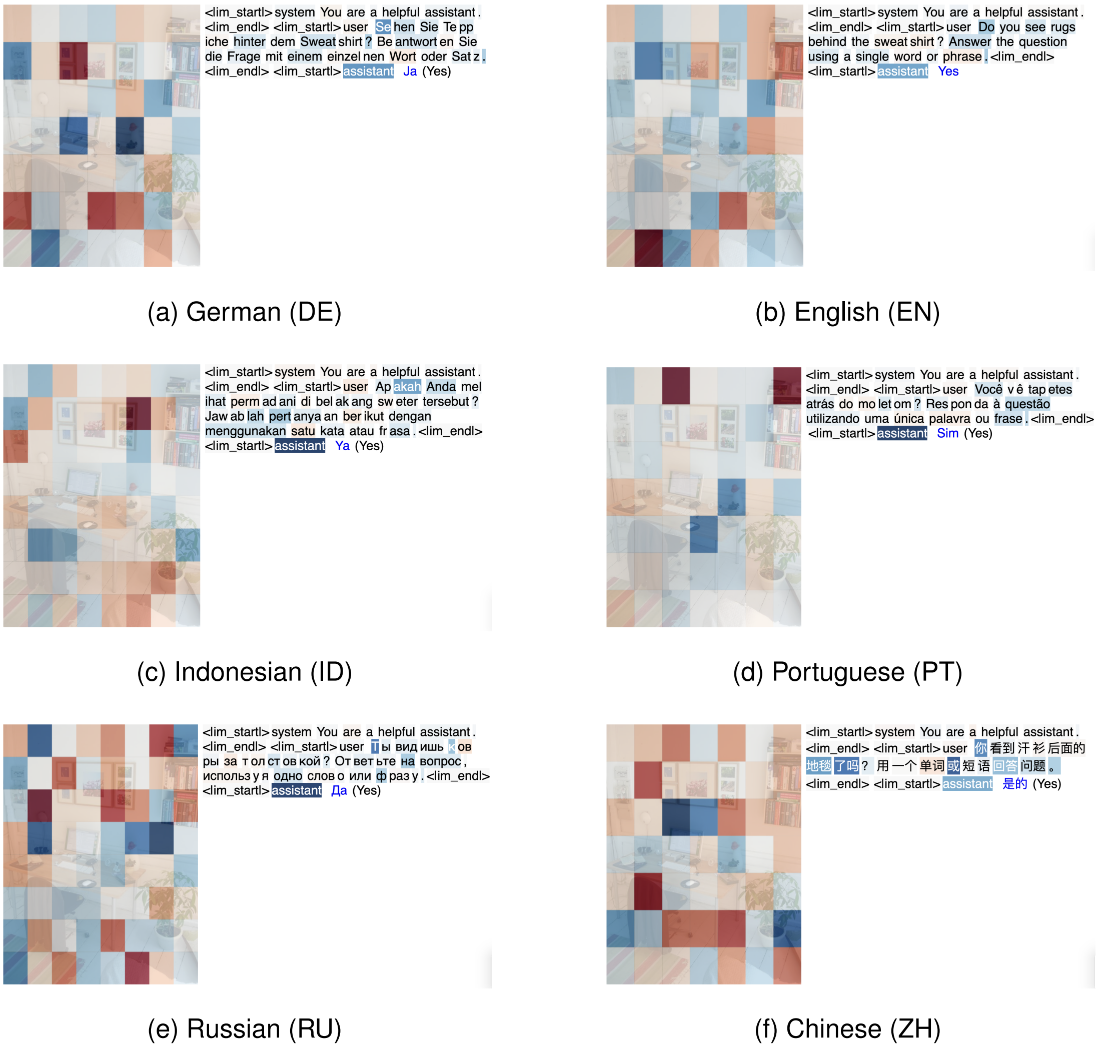
(e) Russian (RU)

<lim_startl> system You are a helpful assistant . <lim_endl> <lim_startl> user 另一 把 椅子 左边 的 椅子 是什么 颜色 ？ 用 一个 单词 或 短 语 回答 问题 。 <lim_endl> <lim_startl> assistant
另一把椅子左边的椅子是白色的。

(f) Chinese (ZH)

Figure A.1: Visualizations of MM-SHAP scores for image and text tokens across different languages for an example of the *query* type question.

A.2 *Verify* Questions



<|im_start|> system You are a helpful assistant .
<|im_end|> <|im_start|> user Se hen Sie Te pp
iche hinter dem Sweat shirt ? Be antwort en Sie
die Frage mit einem einzel nen Wort oder Sat z .
<|im_end|> <|im_start|> assistant Ja (Yes)

(a) German (DE)

<|im_start|> system You are a helpful assistant .
<|im_end|> <|im_start|> user Do you see rugs
behind the sweat shirt ? Answer the question
using a single word or phrase . <|im_end|>
<|im_start|> assistant Yes

(b) English (EN)

<|im_start|> system You are a helpful assistant .
<|im_end|> <|im_start|> user Ap akah Anda mel
ihat perm ad ani di bel ak ang sw eter tersebut ?
Jaw ab lah pert anya an ber ikut dengan
menggunakan satu kata atau fr asa . <|im_end|>
<|im_start|> assistant Ya (Yes)

(c) Indonesian (ID)

<|im_start|> system You are a helpful assistant .
<|im_end|> <|im_start|> user Você v ê tap etes
atrás do mo let om ? Res pon da à questão
utilizando uma única palavra ou frase . <|im_end|>
<|im_start|> assistant Sim (Yes)

(d) Portuguese (PT)

<|im_start|> system You are a helpful assistant .
<|im_end|> <|im_start|> user Т ы вид ишь к ов
ры за т ол ст ов кой ? От вет ьте на вопрос ,
использ у я одно слов о или ф раз у . <|im_end|>
<|im_start|> assistant Да (Yes)

(e) Russian (RU)

<|im_start|> system You are a helpful assistant .
<|im_end|> <|im_start|> user 你 看到 汗 衫 后面的
地毯 了吗 ? 用 一个 单词 或 短 语 回答 问题 。
<|im_end|> <|im_start|> assistant 是的 (Yes)

(f) Chinese (ZH)

Figure A.2: Visualizations of MM-SHAP scores for image and text tokens across different languages for an example of the *verify* type question.