M5 – A Diverse Benchmark to Assess the Performance of Large Multimodal Models Across Multilingual and Multicultural Vision-Language Tasks

Florian Schneider[1]
Language Technology Group
Universität Hamburg, Germany
florian.schneider-1@uni-hamburg.de

Sunayana Sitaram
Microsoft Research India
Bangalore, India
sitaram@microsoft.com

Abstract

Since the release of ChatGPT, the field of Natural Language Processing has experienced rapid advancements, particularly in Large Language Models (LLMs) and their multimodal counterparts, Large Multimodal Models (LMMs). Despite their impressive capabilities, LLMs often exhibit significant performance disparities across different languages and cultural contexts, as demonstrated by various text-only benchmarks. However, current research lacks such benchmarks for multimodal visio-linguistic settings. This work fills this gap by introducing M5, the first comprehensive benchmark designed to evaluate LMMs on diverse vision-language tasks within a multilingual and multicultural context. M5 includes eight datasets covering five tasks and 41 languages, with a focus on underrepresented languages and culturally diverse images. Furthermore, we introduce two novel datasets, M5-VGR and M5-VLOD, including a new Visio-Linguistic Outlier Detection task, in which all evaluated open-source models fail to significantly surpass the random baseline. Through extensive evaluation and analyses, we highlight substantial task-agnostic performance disparities between high- and low-resource languages. Moreover, we show that larger models do not necessarily outperform smaller ones in a multilingual setting.

1 Introduction

Since the release of ChatGPT, Natural Language Processing has experienced a significant surge in interest and research, with a particular focus on LLMs finetuned to follow human instructions. Besides proprietary models like GPT-4 (Achiam et al., 2023), Claude (Bai et al., 2022), or Gemini (Anil et al., 2023), there are also successful open-source variants such as Llama (Touvron et al., 2023), Phi (Gunasekar et al., 2023; Abdin et al., 2024), or Mistral (Jiang et al., 2023). While LLMs often demonstrate impressive performance on a wide range of tasks, quantifying and measuring this performance is challenging. Nevertheless, recent evaluation studies have shown that LLMs generally perform well in English but much worse in other languages (Ahuja et al., 2023a,b; Holtermann et al., 2024).

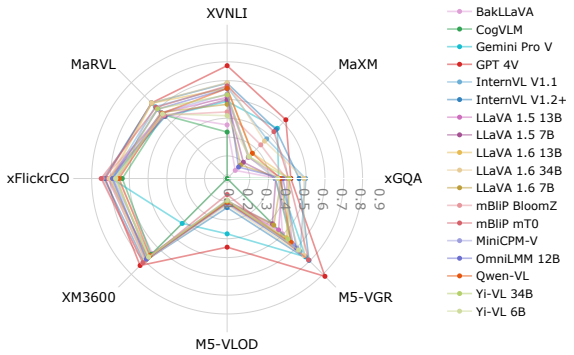


Figure 1: An overview of the average performance of the models on the datasets included in the M5 benchmark. For xFlickrCO and XM3600, we report BERTScore F1. For the other datasets, the accuracy metric is reported.

In this work, we focus on multimodal variants of LLMs, Large Multimodal Models (LMMs), such as GPT 4V (OpenAI, 2023), Gemini Pro V (Anil et al., 2023), or the popular open-source model, LLaVA (Liu et al., 2023a,b). LLMs are not text-only but are also capable of processing images in addition to text. Most open-source LMMs comprise three major components: an LLM, a vision-encoder model, and a mapping network that projects image embeddings into the text embedding space. With this architecture, where an LLM serves as the core, we argue that LMMs inherently suffer from the same issue as LLMs: they generally perform much worse in non-English languages. However, existing benchmarks are either text-only (Ahuja et al., 2023a) or multimodal but monolingual (Yue et al., 2023), thus unable to prove

[1]This works was done during a research internship with Microsoft Research India (Bangalore) between November 2023 and March 2024.

Findings of the Association for Computational Linguistics: EACL 2024, pages 4309–4345
November 12-16, 2024

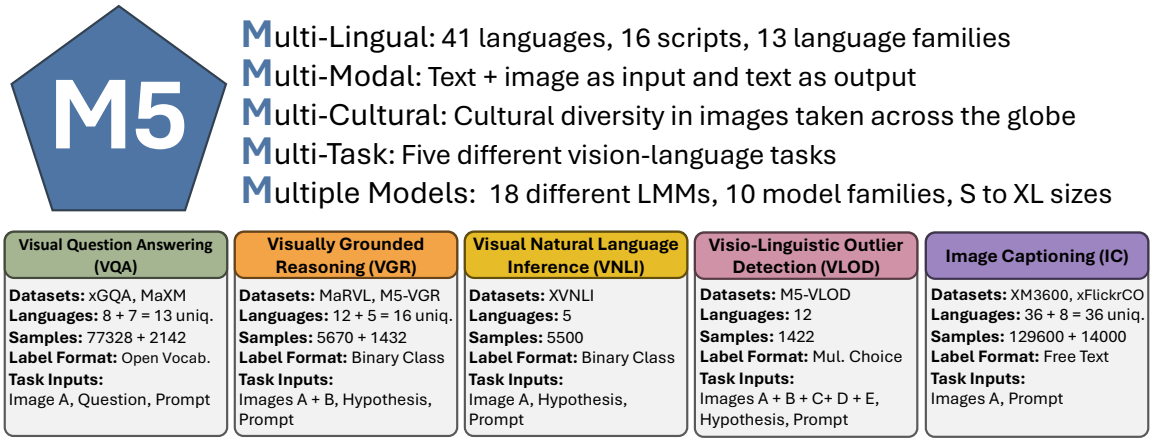


Figure 2: An informative overview of the M5 Benchmark introduced in this work.

this hypothesis. In other words, current research lacks multimodal multilingual benchmarks to examine LMMs' multilingual capabilities. In this work, we fill this gap by introducing the M5 Benchmark, taking a significant step towards identifying and measuring the performance disparities of current LMMs between various languages. Figure 2 and Figure 1 present a high-level summary of our benchmark. Moreover, we introduce two new evaluation datasets, including a novel vision-language task. Both datasets focus on African and Asian cultures, which are underrepresented or even non-existent in previous benchmarks. Our exhaustive analyses additionally investigate the influence of different factors on the performance, such as the models' size or language fidelity.

Major Contributions The major contributions of this work are (a) M5, the first multimodal benchmark to assess the performance of current LMMs across five tasks, eight datasets, and 41 languages; (b) Two novel datasets spanning 10 underrepresented African and Asian languages, English and German, with images depicting the respective cultures. (c) A novel vision-language task: Visio-Linguistic Outlier Detection (VLOD); (d) A large-scale evaluation of 18 recent LLMs and a thorough analysis of their multilingual performance. (e) A public release of our codebase and all datasets in a uniform schema to foster future research for more equitable and accessible LMMs or AI in general[1].

2 Related Work

Large Multi-Modal Models This work focuses on the multimodal counterpart of large language models (LLMs), often referred to as Large Multimodal Models (LMMs). LMMs are language models capable of processing and "understanding" data other than text. While this generally subsumes images, video, audio, or more, we concentrate on visio-linguistic LMMs, i.e., models that take text and/or images as input and generate textual output.

The vast majority of open-source LMMs comprise three major components: a pretrained generative LLM as the core, a pretrained vision-encoder model that computes semantically rich image embeddings, and a shallow mapping network that learned to project image embeddings into the text embedding space. One of this architecture's successful open-source implementations with a recent LLM, i.e., the Llama-based Vicuna (Chiang et al., 2023; Touvron et al., 2023), is LLaVA (Liu et al., 2023b), from which many others took inspiration also regarding the training data and process. Besides this, LMMs also exist, which use Cross-Attention (Wang et al., 2023; Bai et al., 2023), Q-Formers (Li et al., 2023; Geigle et al., 2023), Adapters (Eichenberg et al., 2022), or Preceiver Resamplers (Alayrac et al., 2022; Awadalla et al., 2023) to process image embeddings. For an overview including architectural details and the number of parameters of the 18 LMMs' components we employed in this work, please see Table 8.

Evaluation Benchmarks With the recent surge in the research of LLMs and LMMs, analyzing the models' performances is crucial yet challenging. Popular benchmarks like BIG-Bench (bench authors, 2023), HELM (Liang et al., 2022), or MMLU (Hendrycks et al., 2020) are the defacto-standard to evaluate LLMs on text-only tasks primarily in English. Efforts like MEGA (Ahuja et al., 2023a), MEGAVERSE (Ahuja et al., 2023b), or MultiQ (Holtermann et al., 2024) extended these monolingual benchmarks to a large set of diverse

[1] https://github.com/floschne/m5b

languages and showed that the LLMs' performance in English versus non-English languages differs significantly.

Similarly, efforts have been made to evaluate multimodal models. Benchmarks like MMMU (Yue et al., 2023), MME (Fu et al., 2023), or MMBench (Yuan et al., 2023) assess the performance of LMMs on a vast number of text-image tasks. However, these benchmarks primarily focus on English, with some tasks available in Chinese. Like MMMU, there is CMMMU (Ge et al., 2024), which focuses on text-image tasks in Chinese. Nonetheless, evaluating state-of-the-art LMMs in a massively multilingual large-scale setting remains largely unexplored. There are only a few multimodal multilingual evaluation datasets (see Section 3.2 and 8.6) and only two benchmarks: IGLUE (Bugliarello et al., 2022) and MEGAVERSE. However, IGLUE evaluates only non-autoregressive transformer-encoders, thus lacking state-of-the-art LLMs. In MEGAVERSE, only five recent LMMs are evaluated on two datasets.

3 The M5 Benchmark

This section describes the setup of the M5 Benchmark introduced by this work. Details about the experimental setup, including prompts and hyperparameters, are reported in Appendix A.

3.1 Models

We chose the LMMs included in this benchmark for the following reasons: Firstly, we focussed on publicly available models released on Hugging Face except for GPT-4 Vision and Gemini Pro. Secondly, we included LMMs well-performing on popular multimodal English-only benchmark s such as MMMU (Yue et al., 2023) and MME (Fu et al., 2023). Thirdly, we aimed to cover a mixture of different model families and a broad model size spectrum, including small models with 3B to 9B, medium models with 10B to 19B, and large models with 20B to 40B parameters. For an overview of all models, including their number of parameters and other architectural details, see Table 8.

3.2 Datasets

This section briefly introduces the existing datasets included in our benchmark. In addition to these, we crafted two novel datasets described in Section 4. For details about the languages covered by the datasets, please refer to Table 6.

xGQA The xGQA dataset (Pfeiffer et al., 2022) is a cross-lingual visual question-answering dataset. It extends the well-known English-only GQA dataset (Hudson and Manning, 2019) by manually translating the questions in the balanced *test-dev* set. Each of the 9666 questions is available in eight languages covering five scripts, while the answers are in English only. The dataset holds 300 unique images from Visual Genome (Krishna et al., 2017).

MaXM The MaXM dataset was introduced by Changpinyo et al. (2023) and is a VQA dataset comprising seven languages in five scripts. In MaXM, the questions and their respective answers are in the same language. 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. This ensures cultural diversity in the images in addition to the language diversity in the question-answer texts.

XVNLI The XVNLI dataset (Bugliarello et al., 2022) introduces the task of Cross-lingual Visual Natural Language Inference where a model needs to predict whether a textual hypothesis *entails*, *contradicts*, or is *neutral* concerning a visual premise. XVNLI comprises five languages covering three scripts and 357 unique images from Visual Genome.

MaRVL The MaRVL dataset (Liu et al., 2021) aims to benchmark models on Multicultural Reasoning over Vision and Language. A task sample comprises two images, a textual statement, and a binary true or false answer grounded in the images. MaRVL comprises five languages covering three scripts and 4914 culturally diverse images that match the respective languages. The images in a sample are chosen to match the culture of the annotator who has written the textual statement in his or her native language.

XM3600 The XM3600 dataset (Thapliyal et al., 2022) is a large multilingual image captioning dataset comprising 36 languages with 261375 captions covering 13 different scripts for 100 unique images per language. The images are selected to match the language's cultural background, ensuring cultural and linguistic diversity. The captions were not automatically translated but manually created by professional annotators who are native speakers of the respective language.

xFlickrCO The xFlickrCO dataset (Bugliarello et al., 2022) is an image captioning dataset and comprises 1000 images from Flickr30k (Young et al., 2014) and 1000 images from COCO (Lin et al., 2014). Each image is captioned in eight languages, covering four different scripts. For all languages except English and German, the captions were manually crafted by crowdsourcing workers instead of translated from English to prevent bias and increase linguistic diversity.

4 Novel M5 Datasets

In addition to the existing datasets introduced in the previous section, we crafted two novel multimodal and multilingual evaluation datasets. The principal motivation behind this is to fill the gap in existing vision-language datasets concerning the lack of underrepresented languages, tasks, and cultural diversity. Moreover, we aim to enable further examination of LMMs and their performance on non-English and non-Western data with a particular focus on African and Asian regions. More details, statistics, and examples are reported in Appendix B.

Common Characteristics

Languages Both datasets comprise samples in 12 languages covering seven scripts (see Table 6): *Amharic*, *Berber*, *Bengali*, *German*, *English*, *Filipino*, *Hausa*, *Hindi*, *Russian*, *Swahili*, *Thai*, *Zulu*. The languages were selected to enrich the set of languages covered by existing datasets, focusing on underrepresented languages from Asian and African countries or regions. To our knowledge, no other visio-linguistic evaluation dataset covers Amharic, Berber, Hausa, or Zulu.

Data Annotation The textual data in both datasets is manually created by professional annotators who are native speakers of the respective languages. All annotators work for a data annotation company, and fluent English-speaking correspondents handle communication and task delegations. In order to ensure that the annotators can fulfill the tasks as well as possible, detailed guidelines, including multiple good and bad examples, have been drawn up in English. These guidelines were explained in detail to the correspondents. The correspondents then delegated the tasks to the annotators by having internal company guidelines drawn up in the target languages. After the annotation tasks were finished, we conducted the following quality assessment procedure:

1. We translated all manually created annotations to English using the Bing Translate API.
2. We developed a small tool that displays a sample, including the images, target language, original and English-translated annotations, and other metadata.
3. We used the tool to manually inspect 20% of the samples and tagged them as "good", "bad", or "ambiguous/problematic".
4. We discussed in detail our findings with the annotators' correspondents, who then delegated the tasks to improve the quality of the annotations.
5. This loop was executed two times until no more issues were found by the authors and the annotators' correspondents.

Depicting Cultural Diversity The images in our datasets originate from the Dollar Street dataset (Gaviria Rojas et al., 2022), comprising around $38K$ photos taken in 63 different regions or countries around the globe. These photos depict the lives of families, including their homes, neighborhoods, or everyday objects, in a culturally diverse way. Further, each image in the original dataset is tagged with one or more "topics" that roughly describe its visual content.

Image Basis For our datasets, we sampled a subset of images from the Dollar Street dataset (Gaviria Rojas et al., 2022) taken in regions where the 12 target languages are spoken. In this subset, which forms the visual basis for both of our datasets and is referred to as $\mathbb{B}$, each image $i_l^t \in \mathbb{B}$ is tagged with exactly one topic $t \in \mathbb{T} = \{t_0, \ldots, t_{86}\}$ and was taken in a region r_l where language $l \in \mathbb{L} = \{l_0, \ldots, l_{11}\}$ is spoken. More information about the image topic distribution per language can be found in Appendix B.1.3.

4.1 M5-VGR

Inspired by MaRVL, the goal of the M5-VGR dataset is to provide a visually grounded reasoning (VGR) evaluation dataset that covers a wide range of topologically different languages and, at the same time, visually represents a diverse set of cultures in which the respective languages are spoken. However, since the MaRVL dataset contains

Figure 3: An Zulu example of the novel M5-VGR dataset. **Hypothesis:** *"Isithombe sokuqala nesithombe sesibili sibonisa iqanda elisehhokweni. (The first picture and the second picture show the egg on the head.)*", **Label:** *False*

only five languages, we chose 11 additional topologically diverse languages for our dataset. To guarantee visual and linguistic diversity and high data quality in our dataset, we hired professional native-speaker annotators of the respective languages to annotate the data. Moreover, we performed several rounds of data quality assessment in close collaboration with the annotators.

A task sample s in M5-VGR contains two images i_a and i_b, a textual visually grounded hypothesis h, and a binary label c which is either true or false concerning the two visual premises (see Figure 3). More specifically, for each language $l \in \mathbb{L}$, we created 120 tasks $s_l \in \mathbb{S}_l$ as follows: In the first step, we sampled 120 unique images $a_l^t \in \mathbb{B}$ from our image basis so that each topic $t \in \mathbb{T}$ occurs at least once across all 12 languages. Then, for each of the 120 images, we randomly selected another image $b_{l_2}^t \in \mathbb{B}$ associated with another language $l_2 \neq l \in \mathbb{L}$ that shares the topic t. In the third step, we asked the native-speaker annotators of the language l to manually create a hypothesis h and a label c which is either true or false concerning the image premises $\left(a_l^t, b_{l_2}^t\right)$. Further, the annotators were instructed to generate a hypothesis semantically related to the topic t if possible.

4.2 M5-VLOD

Figure 4: A Swahili example of the novel M5-VLOD dataset. **Hypothesis:** "*Picha zote zinaonyesha sabuni inayotumika kwa mikono na mwili bila mtu yeyote. (All the images show soap applied to the hands and body without anyone.)*", **Outlier:** 1.

With the M5-VLOD dataset, we introduce a novel multimodal task: Visio-Linguistic Outlier Detection. The objective of the task is to detect an outlier image from a set of images considering a textual statement. An example of the task is shown in Figure 4, where five images related to the topic "soap for hands and body" are shown. The machine-translated English statement is: "All the images show soap applied to the hands and body without anyone.". Because only the first image shows a person, the statement is incorrect for the first image and, therefore, is considered the outlier image.

The dataset was collected similarly to M5-VGR, as described in the previous section. The major difference is that instead of sampling only one image in the second step, we sample four images so that a sample $s_{l_0} \in \mathbb{S}_{l_0}$ for language $l_0 \in \mathbb{L}$ comprises of five images: $\{a_{l_0}^t, b_{l_1}^t, c_{l_2}^t, d_{l_3}^t, e_{l_4}^t, \}$ associated with five different languages $\{l_0, \ldots, l_4 \in \mathbb{L}\}$ that share one topic $t \in \mathbb{T}$. In the third step, we asked the native-speaker annotators of the language l to manually create a textual statement h, valid for all but one of the images labeled as the outlier image.

5 General Results Discussion

This section discusses the models' performance on the datasets considered in our benchmark. Table 1 provides an overview of the performance in English compared to non-English languages for all models and datasets. Note that we use friendly names for the models for better readability (see Table 8). Detailed results for each dataset and all their respective languages are provided in Appendix D.

5.1 Summary of Findings

Table 1 shows a clear pattern: Generally, LMMs perform significantly worse in non-English languages across all tasks. More specifically, the average performance across all models and datasets in English is 0.63 versus 0.47 in non-English languages. Most models have an average performance difference from English to non-English larger or equal to 0.12. However, for `GPT 4V` and despite their much smaller size also for `mBlip BloomZ`, and `mBlip T0`, the difference is smaller than 0.1. For the two `mBLIP` models, the authors explicitly stated in their paper the language distribution in the training data, which covers 96 languages. Hence, it can be assumed that this is the reason for this slight absolute performance difference, and, further, this might indicate that `GPT 4V` was also trained in a

Model	Dataset																				
	xGQA		MaXM		XVNLI		MaRVL		M5-VLOD		M5-VGR		xFlickrCO		XM3600		ALL				
	E	NE	E	NE	E	NE	E	NE	E	NE	E	NE	E	NE	E	NE	E	NE	Δ		
CogVLM	0.59	0.30	0.43	0.02	0.47	0.29	0.60	0.51	0.10	0.08	0.68	0.55	0.87	0.60	0.88	0.65	0.58	0.38	−0.20		
BakLLaVA	0.62	0.32	0.53	0.08	0.48	0.34	0.59	0.53	0.14	0.20	0.71	0.48	0.91	0.63	0.88	0.64	0.61	0.40	−0.21		
LLaVA 1.6 7B	0.60	0.34	0.34	0.16	0.59	0.45	0.62	0.53	0.14	0.21	0.55	0.42	0.88	0.64	0.88	0.67	0.57	0.43	−0.15		
LLaVA 1.5 7B	0.62	0.30	0.52	0.15	0.60	0.47	0.57	0.52	0.15	0.20	0.48	0.42	0.92	0.68	0.89	0.67	0.59	0.43	−0.17		
Yi-VL 6B	0.57	0.32	0.53	0.20	0.56	0.38	0.59	0.53	0.20	0.19	0.73	0.61	0.91	0.64	0.91	0.66	0.62	0.44	−0.18		
MiniCPM-V	0.55	0.31	0.56	0.19	0.66	0.49	0.61	0.53	0.20	0.20	0.80	0.56	0.91	0.65	0.90	0.65	0.65	0.45	−0.20		
LLaVA 1.5 13B	0.62	0.34	0.56	0.19	0.59	0.49	0.60	0.54	0.16	0.21	0.57	0.46	0.91	0.69	0.90	0.69	0.61	0.45	−0.16		
Qwen-VL	0.59	0.33	0.50	0.23	0.62	0.54	0.60	0.53	0.16	0.21	0.82	0.54	0.89	0.62	0.90	0.65	0.64	0.46	−0.18		
Yi-VL 34B	0.58	0.38	0.53	0.20	0.59	0.51	0.62	0.58	0.26	0.19	0.77	0.52	0.91	0.64	0.90	0.66	0.65	0.46	−0.19		
Gemini Pro V	0.46	0.34	0.48	0.23	0.49	0.49	0.55	0.55	0.52	0.36	0.79	0.66	0.86	0.67	0.63	0.41	0.60	0.46	−0.13		
OmniLMM 12B	0.49	0.36	0.48	0.11	0.64	0.54	0.64	0.56	0.19	0.21	0.78	0.59	0.91	0.66	0.89	0.68	0.63	0.46	−0.16		
LLaVA 1.6 13B	0.65	0.38	0.46	0.24	0.61	0.55	0.65	0.65	0.14	0.21	0.78	0.50	0.90	0.67	0.88	0.68	0.63	0.48	−0.15		
mBliP BloomZ	0.44	0.39	0.55	0.29	0.40	0.44	0.55	0.56	0.14	0.21	0.69	0.56	0.92	0.72	0.91	0.71	0.58	0.49	−0.09		
InternVL V1.1	0.63	0.48	0.58	0.34	0.61	0.56	0.63	0.60	0.13	0.21	0.73	0.62	0.92	0.66	0.91	0.68	0.64	0.52	−0.12		
LLaVA 1.6 34B	0.65	0.46	0.58	0.32	0.62	0.58	0.64	0.66	0.26	0.22	0.87	0.64	0.89	0.68	0.88	0.70	0.67	0.53	−0.14		
mBliP mT0	0.44	0.40	0.50	0.42	0.59	0.57	0.60	0.63	0.12	0.17	0.74	0.69	0.92	0.73	0.91	0.71	0.60	0.54	−0.07		
InternVL V1.2+	0.67	0.43	0.60	0.42	0.63	0.58	0.68	0.61	0.28	0.23	0.86	0.68	0.92	0.71	0.90	0.70	0.69	0.55	−0.15		
GPT 4V	0.45	0.41	0.49	0.53	0.69	0.68	0.64	0.66	0.70	0.42	0.88	0.81	0.90	0.70	0.89	0.72	0.70	0.62	−0.09		
Random Baseline	–		–		0.33		0.50		0.20		0.50		–		–		–		–		
Average	0.57	0.37	0.51	0.24	0.58	0.50	0.61	0.57	0.22	0.22	0.73	0.57	0.90	0.67	0.88	0.66	0.63	0.47	−0.15		

Table 1: Average performance in English (E) and non-English languages (NE) on all datasets for all models. For each dataset and the Δ column, the heatmaps are created individually, indicated by the column gutter. The column "ALL" represents the average across all datasets. For xFlickrCO and XM3600, we report BertScore F1 and for the rest of the datasets, we report the relaxed accuracy.

multilingual fashion. Due to the difference in size and the architecture[2] of the `mBlip` models and `GPT 4V`, applying this multilingual training strategy for LMMs would generally lead to more robust multilingual performance.

The average performance difference of the models is most significant on the MaXM, XM3600, and xFlickrCo datasets, for which the models are required to generate non-English text.

Interestingly, for the M5-VLOD dataset, the models that performed worse than the random baseline of 0.2 in English performed better in non-English languages. An explanation for this could be false assumptions drawn from the English text. This finding also explains why the average English versus non-English performance disparity across all models is equal for the dataset and lies around the random baseline, indicating the challenge introduced by our dataset.

5.1.1 Dataset-Specific Discussion

Note that due to brevity constraints, we report exact numbers and diagrams of the language-specific results for each dataset in Appendix D.

xGQA All models perform best in English mostly, with a significant gap in accuracy to the second-best language from up to 0.62 in English to 0.36 in Russian for `LLaVA 1.6 7B`. In Bengali, where the models have the lowest average accuracy of 0.19, all models besides `GPT 4V`, which achieves 0.44, perform worst by far. The best-performing model in English and the best-performing model on average over all languages are the `InternVL v1.2` and `InternVL v1.1` models. Notably, despite their (estimated) much larger size, `GPT 4V` and `Gemini Pro V` are among the worst-performing models in English. After manually inspecting the results, we found the reason for this to be that the models did not respond in a single word but with a brief sentence, which is considered a false answer according to the applied metric (see Appendix A.2 and Section 8.2).

MaXM The average accuracy of the models for Hindi (0.22), Hebrew (0.19), Romanian (0.27), Thai (0.25), and Chinese (0.24) is much lower than for English (0.51) and French (0.35). It is also worth pointing out that most models, regardless of their size, perform remarkably worse in languages other than English (and French). In contrast, on xGQA, which is also a VQA dataset, the differences between the languages are much more minor. This is likely due to the difference between the two datasets, i.e., that xGQA has multilingual questions but only English answers, while MaXM has multilingual questions and expects the answers in the respective language, too. We further underline this in our language fidelity analysis in Section 6.3.

XVNLI English accuracy is the best for most models, with an average of 0.58, whereas Arabic accuracy is the worst, with an average of 0.43. The performance drop from English to the other languages, i.e., Spanish (0.51), French (0.52), and

[2]While the architecture of `GPT 4V` is not known, it is likely different from the mBlip models' architecture, which employs Q-Formers, rarely used in state-of-the-art LMMs.

Russian, with average accuracy scores of 0.51, 0.52, and 0.52, is less substantial. Note that XVNLI is an NLI dataset, i.e., the random baseline is at $\frac{1}{3}$. All models surpass this baseline in all languages, except for CogVLM in Arabic (0.26) and French (0.27). The best-performing model is GPT4 V with an average accuracy across all languages of 0.68, followed by LLaVA 1.6 34B and InternVL V1.2+ with average scores of 0.59 and 0.58, respectively.

MaRVL The dataset's random baseline is 0.5, which is often only slightly surpassed by most models, especially for Swahili and Tamil languages, with an average accuracy of 0.53 and 0.54, respectively. Notably, only 8 of 18 models perform best in English, with an average accuracy of 0.61. For the other models, the English performance is surpassed by Chinese, Indonesian, or Turkish, with an average accuracy of 0.60, 0.60, and 0.59, respectively. GPT-4V is on par with LLaVA 1.6 34B despite the latter having much fewer parameters.

M5-VGR As with MaRVL, this dataset's random baseline is at 0.5. Only one of 18 models, i.e., InternVL V1.2+, could surpass or reach this baseline in all languages. As expected, most models performed best in English, German, or Russian, with average accuracies of 0.73, 0.68, and 0.69, respectively. They performed worst in low-resource languages such as Amharic, Berber, Bengali, Hausa, or Zulu, with an average accuracy of 0.53, 0.49, 0.55, and 0.52, respectively. Only three models, i.e., Gemini Pro V, mBlip mT0, and GPT 4V, consistently and significantly surpass the random baseline in all languages except for Berber. The only languages where the average performance is significantly higher than the 0.5 random baseline are English (0.73), German (0.68), Russian (0.69), and Thai (0.62). The average scores of the other languages range from 0.49 in Berber to 0.57 in Hindi.

M5-VLOD The dataset's random baseline is 0.2 since the models need to find the outlier within five images. Only GPT 4V and Gemini Pro V significantly surpassed that baseline in all languages, with an average accuracy of 0.42 and 0.36, respectively. They achieve the best scores in English with an average accuracy of 0.70 (GPT 4V) and 0.52 (Gemini Pro V. However, in Berber, both models only achieve scores around the random baseline. All other models do not surpass the random baseline in all languages, including English, by more than 0.1, with average scores between 0.08 (CogVLM) and 0.23 (InternVL V1.2+) This highlights the challenge introduced by our dataset and the performance gap between proprietary and open-source models.

xFlickrCO The majority of models perform best in English, often with a significant margin in average chrF++, i.e., 24.93 in English and 12.49 in non-English languages. Other languages where the models perform comparably well are German and Spanish, with average chrF++ scores of 19.95 and 19.55, respectively. Interestingly, all models perform worse in non-Latin script languages, i.e., Russian (9.70), Chinese (4.53), and Japanese (4.05). Unexpectedly, the proprietary models GPT 4V and Gemini Pro V are surpassed by mBliP BloomZ, mBliP mT0, and InternVL V1.2+, which are much smaller open-source models. Even in English, most open-source models outperform the proprietary models.

XM3600 Note that due to limited resources, we evaluated GPT 4V only on a subset of 12 of 36 languages. Most models perform best in English (27.14 average chrF++) by a large margin, followed by other Latin scripts in high-resource languages such as French (23.65), Spanish (23.52), or Dutch (21.01). On average, the models perform worst on non-Latin script languages like Korean (3.50), Telugu (4.79), and Bengali (5.11). However, although the chrF++ metric claims to be script and language-independent, the low scores in high-resource languages like Chinese (3.95) and Japanese (5.13) make the metric questionable. While detailed analysis is out of the scope of this work, in future work, we will investigate this issue further (see Section 8.1).

6 Aggregated Result Analyses

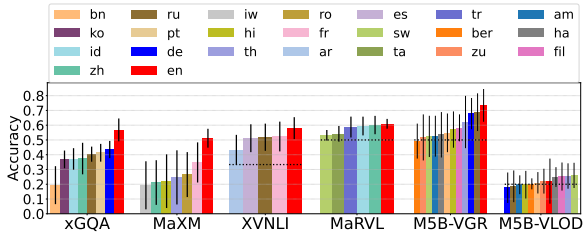
6.1 Performance per Language

Figure 5 shows the average performances aggregated by language[3] or language taxonomy classes (Joshi et al., 2020). These taxonomy classes indicated how well a respective language is represented and considered within the research field of NLP based on papers published at CL conferences. High-resource languages such as English or German are in Class 5, whereas low-resource

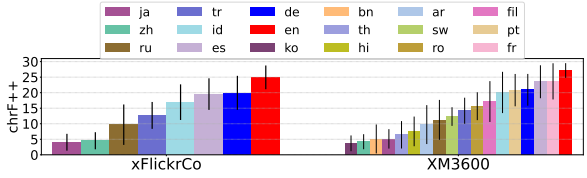
[3]We do not show all 36 languages of XM3600 for better readability.

languages such as Berber are in Class 0. For details about the languages and their taxonomy classes, please refer to Table 7.

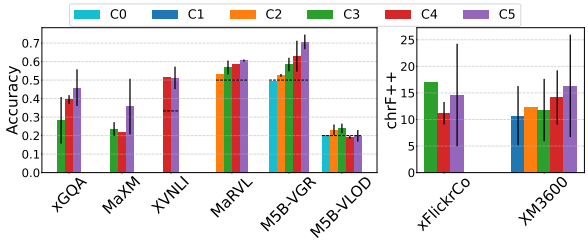
As can be observed from Figure 5a and Figure 5b, the models perform best in English, followed by other European languages across all datasets. Our newly presented M5-VLOD dataset is an exception, where the average performance for all languages is around the random baseline, indicating the challenge it implies. As expected, the models consistently perform worse on low-resource languages than on high-resource languages on all datasets. This is also displayed in Figure 5c, where it can be observed that the average performance decreases with the language taxonomy class. Note that this is not precisely true for xFlickrCO and XVNLI because the average on Class-5 languages is lowered by outliers, as indicated by the large error bars. In contrast, the models performed comparably well in only one Class 3 or 4 language, respectively.



(a) Performance on VQA, VGR, and VNLI datasets aggregated by language.



(b) Performance on image captioning datasets aggregated by language.



(c) Performance on datasets aggregated by language taxonomy class as introduced by Joshi et al. (2020).

Figure 5: Models' performances on all datasets aggregated by language or language taxonomy classes.

6.2 Performance vs. Model Parameters

In Figure 6, we plot the English and non-English average performance on the employed datasets versus the models' sizes in multiple regression plots. Note that, on the x-axes, we indicated the unknown sizes of `GPT 4V` and `Gemini Pro V` by "*???*", which are estimated to be of magnitudes larger than all other models evaluated in this benchmark hence should be much further right. However, we did not do so to improve the readability of the plots.

In the figures, we can make several observations: Firstly, the average English performance is higher than the non-English performance for all models on all datasets. Secondly, the markers, which represent the average performance of a specific model on a dataset, show that the largest model does not always perform best and that the difference between smaller and larger models is often neglectable. The same finding is shown by the relatively flat slope of the regression lines. However, for the M5-VLOD and VGR datasets, the regression line for the average English scores is steeper, meaning that larger models perform considerably better than the smaller models. Since this work introduces the datasets and M5-VLOD even introduces a novel task, it can be concluded that larger models can better generalize to unseen data.

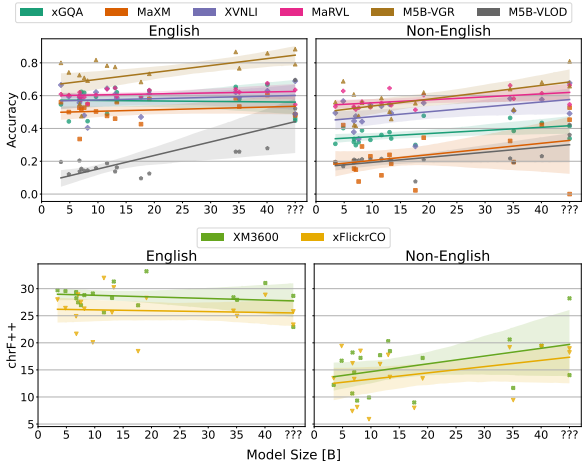


Figure 6: Regression plots showing the English and average non-English performance versus model size on different datasets. On the x-axis, we indicated the unknown sizes of `GPT 4V` and `Gemini Pro V` by "???".

6.3 Language Fidelity Analysis

Inspired by Holtermann et al. (2024), we report the results of a language fidelity analysis, which assesses how often a model responds in the requested language on average. For this, we used GlotLIDv3 (Kargaran et al., 2023) to predict the language based on the output text of the respective models. Since it is hard to predict the language of a word or a multi-word expression due to ambigu-

ity, we selected the xFlickrCO dataset, where the expected response of a model is an image caption, i.e., a sentence, in one of eight languages. As it can be observed from Table 2, all models achieve (almost) perfect fidelity in English where, whereas for Japanese, Russian, and Turkish, the average fidelity drops to two-thirds. Interestingly, the small-sized `mBLIP` models have almost perfect fidelity in all languages, (slightly) surpassing larger models like `InternVL V1.2+` and `GPT 4V`.

Table 2: Language fidelity results on the xFlickrCO dataset.

Model	Language								
	zh	en	de	id	ja	ru	es	tr	Avg.
BakLLaVA	.00	1.0	.39	.06	.00	.00	.44	.00	.24
Yi-VL 6B	.14	1.0	.20	.00	.20	.01	.57	.00	.28
Qwen-VL	.95	.99	.18	.11	.15	.08	.15	.07	.33
Yi-VL 34B	.43	1.0	.79	.45	.58	.22	.25	.33	.51
CogVLM	.44	.95	.74	.76	.38	.43	.82	.54	.63
LLaVA 1.5 13B	.88	1.0	.75	.55	.90	.26	.75	.40	.69
LLaVA 1.5 7B	.83	1.0	.96	.83	.09	.22	.97	.67	.70
MiniCPM-V	.21	1.0	.93	.79	.89	.96	.91	.68	.80
LLaVA 1.6 7B	.99	.99	.66	.91	.59	.88	.91	.89	.85
InternVL V1.1	.96	1.0	.93	.78	.88	.89	.97	.66	.89
OmniLMM 12B	.63	1.0	.95	.92	.83	.92	.98	.88	.89
Gemini Pro	.95	.95	.95	.88	.91	.96	.97	.96	.94
LLaVA 1.6 13B	1.0	1.0	.90	.96	.91	.87	.97	.93	.94
LLaVA 1.6 34B	.88	1.0	.99	.99	.86	.99	.99	.99	.96
GPT 4V	.97	1.0	1.0	.98	.88	.99	.99	1.0	.98
InternVL V1.2+	.99	1.0	1.0	.95	.97	.99	.99	.96	.98
mBliP BloomZ	.96	1.0	1.0	.99	.99	1.0	1.0	.99	.99
mBliP mT0	.96	1.0	1.0	.99	.99	1.0	1.0	1.0	.99
Avg.	.73	.99	.79	.72	.67	.65	.81	.66	.75

While the language fidelity of a model focuses on the generated text, we argue that the fidelity is also an indicator of the model's general language capabilities. To prove this hypothesis, we computed Pearson correlation coefficients between the reported fidelity and the models' performance on the datasets for the xFlickrCO languages. As shown in Table 17, there is a positive moderate or high correlation between the average fidelity and the average score for most datasets. However, for xGQA and M5-VLOD, there is only a minor positive average correlation.

7 Conclusion

We introduced M5, a diverse benchmark in which we evaluated 18 Large Multimodal Models (LMMs) with varying sizes across five visio-linguistic tasks in eight datasets comprising 41 unique languages. Further, we presented two novel datasets – M5-VGR and M5-VLOD – which focus on underrepresented languages and depict culturally diverse scenes. With M5-VLOD, we introduce a new visio-linguistic outlier detection task in which only proprietary models achieve reasonable scores. Our experiments revealed that model size does not always correlate with better performance, especially in non-English languages, underscoring the importance of diverse, multilingual training data and robust architectures. Performance disparities were prominent between high-resource languages like English and low-resource languages across all datasets and models, highlighting ongoing challenges in achieving globally equitable multilingual AI. With M5, we aim to impel the development of more inclusive models suitable for diverse languages and cultures.

8 Limitations

This section outlines several limitations of our current study that will be addressed in future work.

8.1 Metrics for Multilingual Image Captioning

Our benchmark and current research generally lack robust metrics for evaluating multilingual image captioning, especially for non-Latin script languages. The issue, which is the same for machine translation tasks, arises because of the nature of most metrics, such as chrF (Popović, 2017), CIDEr (Vedantam et al., 2015), ROUGE (Lin, 2004), BLUE (Papineni et al., 2002), or METEOR (Banerjee and Lavie, 2005), which are based on comparing word or character n-grams between the source and target sequence. For non-Latin scripts, tokenization or segmentation can be challenging because it might not contain spaces or punctuation, or the characters are logographic. Hence, their usability or effectiveness is doubtful in such scenarios because the metrics rely on tokenization.

Other metrics, such as BERTScore (Zhang et al., 2020), CLIPScore (Hessel et al., 2021), or COMET (Rei et al., 2020), do not rely on the captions' surface forms but on their token or sentence embeddings. However, they suffer from other issues: They require strong multilingual or cross-lingual encoder models capable of computing embeddings for many languages, which itself is a challenging task. Further, the scores computed with these metrics are often not calibrated across languages and thus not directly comparable between different languages.

A promising currently popular solution might be the use of robust multilingual state-of-the-art LLMs such as GPT 4o[4], Claude 3 Opus[5], or Gem-

[4] https://openai.com/index/hello-gpt-4o/
[5] https://www.anthropic.com/news/

ini 1.5 Ultra[6] as a judge (Zheng et al., 2024). However, this would require more computational and financial resources and, most importantly, more investigation.

8.2 VQA Metrics for Generative Models

The problem when employing and evaluating generative language models on question-answering tasks is that the models can generally output arbitrary token sequences. However, the gold label answers are limited and often comprise only a short phrase, a single word, or even a binary label. Hence, mapping the predicted answers to their gold labels is not straightforward, and the difficulty drastically increases in multilingual scenarios. The relaxed accuracy metric employed in this study (see Section A.1) has been found to occasionally incorrectly classify correct answers, leading to false negatives, especially in open vocabulary visual question answering (VQA). One way to address this issue is to leverage strong state-of-the-art LLMs as judges, as described above, to enhance the accuracy of the evaluations.

8.3 Influence of Prompting

Another limitation of this and most, if not all, other current studies is grounded in the model prompting. Since different models might react differently to specific prompting styles, and we only employ a single prompt per dataset for all models[7] (see Figure 7), the results might not be optimal. This issue has been partially addressed by Ahuja et al. (2023a) but is out of the scope of this work.

8.4 "Outdated" Models

Since the pace of current research in NLP, CV, and multimodal machine learning is swift, the models employed in our benchmarking exercise might be considered slightly outdated. Note that we considered models released until March 2024. Since then, numerous improved LMMs based on state-of-the-art LLMs, such as Llama3[8] and novel image encoders techniques such as NaVIT (Dehghani et al., 2024), have been publicly released. Because this was foreseeable, we designed our benchmark to be easily extendable with newer models, which we will include in future work.

`claude-3-family`

[6] https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/

[7] We do apply the model-specific prompt or chat templates, though.

[8] https://ai.meta.com/blog/meta-llama-3

8.5 Small M5 Datasets

This work introduced two datasets, M5-VGR and M5-VLOD, which comprise about 115 samples for each of the 12 languages. Compared to other datasets, they can be considered small. We will increase their sizes in future work to obtain more robust and generalizable results.

8.6 Missing multimodal and Multilingual Datasets

Currently, the M5 Benchmark comprises 5 text-image tasks, i.e., VQA, VGR, VNLI, and image captioning, thus missing other suitable tasks like multimodal and multilingual summarization. Further, other multimodal multilingual VQA and VGR datasets have emerged while writing this paper. We will include both new tasks and new datasets in future versions of the M5.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 Technical Report: A Highly Capable Language Model Locally On Your Phone. *ArXiv*, 2404.14219.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 Technical Report. *ArXiv*, 2303.08774.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023a. MEGA: Multilingual Evaluation of Generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore.

Sanchit Ahuja, Divyanshu Aggarwal, Varun Gumma, Ishaan Watts, Ashutosh Sathe, Millicent Ochieng, Rishav Hada, Prachi Jain, Maxamed Axmed, Kalika Bali, et al. 2023b. MEGAVERSE: Benchmarking Large Language Models Across Languages, Modalities, Models and Tasks. *arXiv preprint arXiv:2311.07463*.

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.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: A Visual Language Model for Few-Shot Learning. *Advances in neural information processing systems*, 35:23716–23736.

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: A Family of Highly Capable Multimodal Models. *ArXiv*, 2312.11805.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. 2023. OpenFlamingo: An Open-Source Framework for Training Large Autoregressive Vision-Language Models. *arXiv preprint arXiv:2308.01390*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional AI: Harmlessness From AI Feedback. *ArXiv*, 2212.08073.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan.

BIG bench authors. 2023. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Emanuele Bugliarello, Fangyu Liu, Jonas Pfeiffer, Siva Reddy, Desmond Elliott, Edoardo Maria Ponti, and Ivan Vulić. 2022. IGLUE: A Benchmark for Transfer Learning across Modalities, Tasks, and Languages. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 2370–2392.

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.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2023. InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks. *arXiv preprint arXiv:2312.14238*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality.

Mostafa Dehghani, Basil Mustafa, Josip Djolonga, Jonathan Heek, Matthias Minderer, Mathilde Caron, Andreas Steiner, Joan Puigcerver, Robert Geirhos, Ibrahim M Alabdulmohsin, et al. 2024. Patch n' Pack: NaViT, a Vision Transformer for any Aspect Ratio and Resolution. *Advances in Neural Information Processing Systems*, 36.

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.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. 2023. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. *arXiv preprint arXiv:2306.13394*.

William Gaviria Rojas, Sudnya Diamos, Keertan Kini, David Kanter, Vijay Janapa Reddi, and Cody Coleman. 2022. The Dollar Street Dataset: Images Representing the Geographic and Socioeconomic Diversity of the World. *Advances in Neural Information Processing Systems*, 35:12979–12990.

Zhang Ge, Du Xinrun, Chen Bei, Liang Yiming, Luo Tongxu, Zheng Tianyu, Zhu Kang, Cheng Yuyang, Xu Chunpu, Guo Shuyue, Zhang Haoran, Qu Xingwei, Wang Junjie, Yuan Ruibin, Li Yizhi, Wang Zekun, Liu Yudong, Tsai Yu-Hsuan, Zhang Fengji, Lin Chenghua, Huang Wenhao, Chen Wenhu, and Fu Jie. 2024. CMMMU: A Chinese Massive Multi-discipline Multimodal Understanding Benchmark. *arXiv preprint arXiv:2401.20847*.

Gregor Geigle, Abhay Jain, Radu Timofte, and Goran Glavavs. 2023. mBLIP: Efficient Bootstrapping of Multilingual Vision-LLMs. *ArXiv*, 2307.06930.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, et al. 2023. Textbooks Are All You Need. *ArXiv*, 2306.11644.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring Massive Multitask Language Understanding. *arXiv preprint arXiv:2009.03300*.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic.

Carolin Holtermann, Paul Röttger, Timm Dill, and Anne Lauscher. 2024. Evaluating the Elementary Multilingual Capabilities of Large Language Models with MultiQ. *arXiv preprint arXiv:2403.03814*.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. 2024. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*.

Drew A Hudson and Christopher D Manning. 2019. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, Long Beach, CA, USA.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7B. *ArXiv*, 2310.06825.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online.

Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schütze. 2023. GlotLID: Language Identification for Low-Resource Languages. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, et al. 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision (IJCV)*, 123(1):32–73.

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. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic Evaluation of Language Models. *arXiv preprint arXiv:2211.09110*.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision (ECCV)*, pages 740–755, Zurich, Switzerland.

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.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved Baselines with Visual Instruction Tuning. *ArXiv*, 2310.03744.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916, New Orleans, LT, USA.

OpenAI. 2023. GPT-4 Vision System Card.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA.

Jonas Pfeiffer, Gregor Geigle, Aishwarya Kamath, Jan-Martin Steitz, Stefan Roth, Ivan Vulić, and Iryna Gurevych. 2022. xGQA: Cross-Lingual Visual Question Answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2497–2511, Dublin, Ireland.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A Neural Framework for MT Evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online.

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, et al. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *ArXiv*, 2307.09288.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based Image Description Evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4566–4575, Salt Lake City, UT, USA.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. 2023. CogVLM: Visual Expert for Pretrained Language Models. *ArXiv*, 2311.03079.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From Image Descriptions to Visual Denotations: New Similarity Metrics for Semantic Inference Over Event Descriptions. *Transactions of the Association for Computational Linguistics (TACL)*, 2:67–78.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. 2023. RLHF-V: Towards Trustworthy MLLMs via Behavior Alignment from Fine-Grained Correctional Human Feedback. *arXiv preprint arXiv:2312.00849*.

Liu Yuan, Duan Haodong, Zhang Yuanhan, Li Bo, Zhang Songyang, Zhao Wangbo, Yuan Yike, Wang Jiaqi, He Conghui, Liu Ziwei, Chen Kai, and Lin Dahua. 2023. MMBench: Is Your Multi-modal Model an All-around Player? *arXiv:2307.06281*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2023. MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. *arXiv preprint arXiv:2311.16502*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*, Online.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems*, 36.

A Experimental Setup Details

This section details the employed metrics, prompts, and generation hyperparameters.

Note that we ran all experiments on A6000 (50GB) and A100 (80GB) GPUs. The largest evaluated model (40B) fits on an A100.

A.1 Metrics

Following Geigle et al. (2023), we report a relaxed accuracy metric for the xGQA, MaXM, XVNLI, and MaRVL datasets due to the generative nature of the considered models. More specifically, we post-process the generated answers by, e.g., lowercasing, stripping, or removing punctuation. We then consider the processed generated answer correct if it matches the gold answer or starts or ends with the gold answer. Further, we allow synonyms for boolean and numerical values. Examples can be found in Table A.2.

Inspired by Ahuja et al. (2023b), we report the chrF++ (Popović, 2017) metric for the xFlickrCo and XM3600 datasets.

A.2 Relaxed Accuracy Metric

Table 3: Examples of generated answers considered correct or incorrect in the relaxed accuracy metric used to measure the performance on the xGQA, MaXM, MaRVL, XVNLI, M5-VGR, and M5-VLOD datasets. For more details, please refer to our GitHub repository.

Generated Answer	Gold Answer	Considered Correct
{Yes, 1, True}	true	yes
{No, 0, False}	false	yes
A car.	car	yes
Yes, it is correct.	yes	yes
It is not correct, no.	no	yes
The color of the leaf is green.	green	yes
There are three birds.	three birds	yes
Five	5	yes
{yes, true}	entailment	yes
{no, false}	contradiction	yes
maybe	neutral	yes
There are three birds in the image.	three birds	no
There are three birds.	3	no
three birds	3	no
three birds	3 birds	no

A.3 Prompts

Figure 7 presents the dataset-specific textual prompts we used for all models in this benchmark. Note that this does not include model-specific prompt templates, image placeholders, special tags, or symbols, only the ”raw” textual prompt, which is then embedded in the template as required by the respective model. The placeholders `{QUESTION}`, `{LANGUAGE}`, or `{HYPOTHESIS}` are replaced by the sample specific text. The prompts are partially inspired by Geigle et al. (2023) or Bugliarello et al. (2022).

A.4 Hyperparameters

This section briefly reports hyperparameters used within our experiments for better reproducibility.

A.4.1 Generation Parameters

We used the same generation hyperparameters to generate responses with all the employed open-source models on all datasets (see Table 4). Those are inspired by the default parameters in the “transformers”

xGQA

```
Question: {QUESTION} Short answer in English:
```

MaXM

```
Question: {QUESTION} Short answer in {LANGUAGE}:
```

MaRVL

```
Based on the two images, is it correct to say “{HYPOTHESIS}”? Yes or no? One word answer in
English:
```

XVNLI

```
Is it guaranteed true that “{HYPOTHESIS}”? Yes, no, or maybe? One word answer in English:
```

M5-VGR

```
Based on the two images, is it correct to say “{HYPOTHESIS}”? Yes or no? One word answer in
English:
```

M5-VLOD

```
Based on the 5 images ordered from top-left to bottom-right, which image does not match the
hypothesis “{HYPOTHESIS}”? Choose one from [A, B, C, D, E] and only output a single letter:
```

xFlickrCo

```
Brief caption in {LANGUAGE}:
```

XM3600

```
Brief caption in {LANGUAGE}:
```

Figure 7: Prompts employed for the different datasets.

library[9]. Because for CogVLM, beam search is not supported, we set “num_beams” to 1. For GPT 4V and Gemini Pro V, we use the default parameters of the respective Python clients.

Table 4: Generation hyperparameters to generate responses with all the employed models on all datasets.

Parameter	Value
num_beams	2
do_sample	True
max_new_tokens	50
temperature	1.0
top_k	50
top_p	0.95

A.4.2 Image Order for Multi-Image Datasets

Most models employed in our dataset only support a single image per prompt. For datasets where a sample comprises more than one image, i.e., for MaRVL, M5-VGR, and M5-VLOD, we use the following strategy: We first stack the images horizontally with a gutter of 10 pixels, provide them as a single image in the prompt, and generate the response. Then, we do the same again but stack the images vertically. For

[9] https://huggingface.co/docs/transformers/en/main_classes/text_generation

M5-VLOD, we also create a stacked image with two columns and three rows. The reported scores are the average of all variants.

B Dataset Details

B.1 M5-VGR and M5-VLOD Details

B.1.1 M5-VGR Examples

ID: a06bcb6431fb4a7cb4335bbfb02e2047
Topic: paper
Language: Amharic
Hypothesis:
ሁለቱም ምስሎች ጠረጴዛ ላይ የተቀመጡ ወረቀቶች ያሳያል።
Machine Translation:
Both images show papers sitting on a table.
Answer: True

Figure 8: Amharic M5-VGR Sample.

ID: 90059df935e843f3a59b3b86c36cee96
Topic: tooth paste
Language: Bengali
Hypothesis:
বাম ছবতিে দুটটি টুথপস্টে টউিব দখোয় এবং একটটি টুথপস্টে টউিব ডানদকিে দখোনে।া হয়ছে
Machine Translation:
The left picture shows two toothpaste tubes and one toothpaste tube is shown on the right
Answer: True

Figure 9: Bengali M5-VGR Sample.

ID: acc84d8170bc4cf492af35235731f437
Topic: water outlet
Language: Berber
Hypothesis:
ⵙⵏⴰ ⵜⵍⵍⴰⵅⵉ ⴰⵏ ⵙⵙⵅⵍⵍⵏ ⴰⵙⵅⵙⵙ ⵏ ⵍⴰⵎⴰⵏ ⵏ ⵜⵣⵎⵎ.
Machine Translation:
ⵙⵏⴰ ⵜⵍⵍⴰⵅⵉ ⴰⵏ ⵙⵙⵅⵍⵍⵏ ⴰⵙⵅⵙⵙ ⵏ ⵍⴰⵎⴰⵏ ⵏ ⵜⵣⵎⵎ.
Answer: False

Figure 10: Berber M5-VGR Sample.

ID: 7b86a95366dd424e8d24597936e89434
Topic: paper
Language: English
Hypothesis:
The first image shows green paper in a printer, and the second image shows yellow paper on a wooden floor.
Machine Translation:
The first image shows green paper in a printer, and the second image shows yellow paper on a wooden floor.
Answer: False

Figure 11: English M5-VGR Sample.

ID: cc82590e83a846cb9edbebcf753055e6
Topic: water outlet
Language: Filipino
Hypothesis:
Ang pinagmumulan ng tubig ay marumi at nagkalat sa parehong larawan.
Machine Translation:
The source of the water is dirty and scattered in the same picture.
Answer: False

Figure 12: Filipino M5-VGR Sample.

ID: e0740bcbc03c406cb099eaa5c2040eda
Topic: table with food
Language: German
Hypothesis:
Das erste Bild zeigt eine Frau, die am Tisch Weintrauben isst, während das zweite Bild Essen für drei Personen zeigt.
Machine Translation:
The first image shows a woman eating grapes at the table, while the second image shows food for three people.
Answer: False

Figure 13: German M5-VGR Sample.

ID: 1668e4ad23d247909860a3d32eb2dba2
Topic: lock on front door
Language: Hausa
Hypothesis:
Dukka hotunan biyu kofar ɗaki ne wanda aka rufe da kwaɗon rufe ɗaki
Machine Translation:
Both pictures are a closed room with a closed door.
Answer: True

Figure 14: Hausa M5-VGR Sample.

ID: ba3a8016212e4ba58c2f8adeaa3a42ba
Topic: shower
Language: Hindi
Hypothesis:
दोनों तस्वीरें स्नान घर की है।
Machine Translation:
Both pictures are of the bath house.
Answer: False

Figure 15: Hindi M5-VGR Sample.

ID: 1610f5e020a9435f9e773ef424033e73
Topic: shower
Language: Russian
Hypothesis:
На первом изображении в душевой стены желтые, а на втором изображении в душевой стены красные.
Machine Translation:
In the first image, the shower walls are yellow, and in the second image in the shower walls, they are red.
Answer: False

Figure 16: Russian M5-VGR Sample.

ID: eb2af5af22c2418ea83c7d148c125687
Topic: wardrobe
Language: Swahili
Hypothesis:
Katika picha zote mbili kuna kabati la nguo.
Machine Translation:
In both pictures there is a dresser cupboard.
Answer: False

Figure 17: Swahili M5-VGR Sample.

ID: 53ecad00e365421b8cfc9c220468e9ca
Topic: washing clothes/cleaning
Language: Thai
Hypothesis:
ทั้งสองภาพเป็นภาพคนกำลังซักผ้า
Machine Translation:
Both images are of people doing laundry.
Answer: False

Figure 18: Thai M5-VGR Sample.

ID: c50c1001121a4454aed3b1884ff04167
Topic: guest bed
Language: Zulu
Hypothesis:
Isithombe sokuqala yigumbi elicocwe kahle elinezingubo zokulala ezimhlophe kanti isithombe sesibili yigumbi elingacocwe kahle elidlala ezingane.
Machine Translation:
The first picture is a well-cleaned room with white bedding and the second picture is a poorly cleaned room that plays with children.
Answer: False

Figure 19: Zulu M5-VGR Sample.

B.1.2 M5-VLOD Examples

ID: f07848da8e4544a8a34d6c3e8141e88c
Topic: source of cool
Language: Amharic
Hypothesis:
ሁሉም ምስሎች እራሳችን ለማቀዝቀዝ የምንጠቀምበት መሳሪያን ያሳያል።
Machine Translation:
All images show a device that we use to cool ourselves.
Outlier: 5

Figure 20: Amharic M5-VLOD Sample. The images are ordered from top-left to bottom-right.

ID: 11a37d8036f841d8ba028a501cf856c2
Topic: bedroom
Language: Bengali
Hypothesis:
সব ইমজে বডে রুম বছিানা কন্ট্যাটনিস
Machine Translation:
Bed room bed contins in all images
Outlier: 2

Figure 21: Bengali M5-VLOD Sample. The images are ordered from top-left to bottom-right.

ID: c5149f4ac81e439ea4be741e1f2e722d
Topic: cooking utensils
Language: Berber
Hypothesis:
+ΣШИₒЖΣI ₀Λ ⊙⊙ȣЖЧI+ ЯΣΧₒI I ΣΘₒИИI I ȣ+ҀΣ.
Machine Translation:
+ΣШИₒЖΣI ₀Λ ⊙⊙ȣЖЧI+ ЯΣΧₒI I ΣΘₒИИI I ȣ+ҀΣ.
Outlier: 4

Figure 22: Berber M5-VLOD Sample. The images are ordered from top-left to bottom-right.

ID: 843fac7edeff4fb4a2edc7c3ad1db388
Topic: drainage
Language: English
Hypothesis:
All images show a drain or drainage in a metal, ceramic surface or outside the house.
Machine Translation:
All images show a drain or drainage in a metal, ceramic surface or outside the house.
Outlier: 2

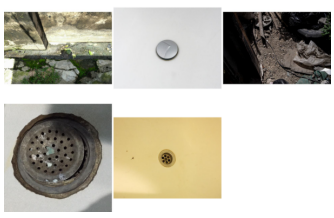
Figure 23: English M5-VLOD Sample. The images are ordered from top-left to bottom-right.

ID: 9af172b955dd4bb29e4d1c8601d504b2
Topic: armchair
Language: Filipino
Hypothesis:
Ang mga upuan sa mga larawan ay may armchair.
Machine Translation:
The chairs in the pictures have armchairs.
Outlier: 5

Figure 24: Filipino M5-VLOD Sample. The images are ordered from top-left to bottom-right.

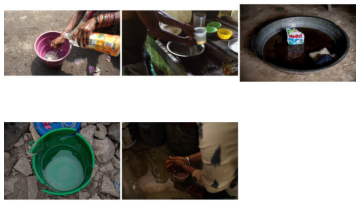
ID: 8f4008857b4c4bfab8135d40a9419219
Topic: plate of food
Language: German
Hypothesis:
Die Bilder zeigen Teller mit Essen, das gegessen wird.
Machine Translation:
The pictures show plates of food being eaten.
Outlier: 4

Figure 25: German M5-VLOD Sample. The images are ordered from top-left to bottom-right.



ID: 97ba6f364e38430eb779c56ad24cf89c
Topic: drainage
Language: Hausa
Hypothesis:
Dukka hotunan akwai hanyyoin magudanar ruwa na waje da cikin gida.
Machine Translation:
All images are available on the exterior and exterior of the house.
Outlier: 3

Figure 26: Hausa M5-VLOD Sample. The images are ordered from top-left to bottom-right.



ID: efa9c9642f4545849405e080a666ee56
Topic: hand washing
Language: Hindi
Hypothesis:
इन सभी छवियों में हैंडवाश करते हुए या हैंडवाश की चीज़ शामिल है
Machine Translation:
All of these images include handwashing or handwashing things
Outlier: 4

Figure 27: Hindi M5-VLOD Sample. The images are ordered from top-left to bottom-right.

ID: da46b1729e8b4871bd1c401d48fa4715
Topic: dish washing brush/cloth
Language: Russian
Hypothesis:
На изображении показан неупакованный предмет для чистки поверхностей.
Machine Translation:
The image shows an unwrapped surface cleaning item.
Outlier: 2

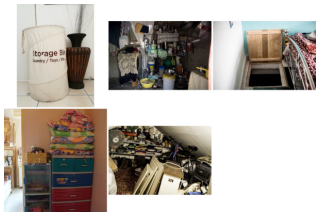
Figure 28: Russian M5-VLOD Sample. The images are ordered from top-left to bottom-right.

ID: a0f64574c38f45b888b18da6032d5547
Topic: bathroom privacy
Language: Swahili
Hypothesis:
Picha zote zinaonyesha faragha ya bafuni.
Machine Translation:
All photos show the privacy of the bathroom.
Outlier: 5

Figure 29: Swahili M5-VLOD Sample. The images are ordered from top-left to bottom-right.

ID: 47d9120f8ff541d19aeb988cab28d62b
Topic: dish racks
Language: Thai
Hypothesis:
ทุกภาพเป็นภาพที่วางจานแบบต่าง ๆ
Machine Translation:
Every picture is a picture of a different type of plate place.
Outlier: 1

Figure 30: Thai M5-VLOD Sample. The images are ordered from top-left to bottom-right.



ID: a8e7cc284e8c4d4794f5a811d09df92e
Topic: storage room
Language: Zulu
Hypothesis:
Izithombe zibonisa amagumbi agcwele izinto zasekhaya ezingasetyenziswa.
Machine Translation:
Pictures show rooms full of usable household items.
Outlier: 3

Figure 31: Zulu M5-VLOD Sample. The images are ordered from top-left to bottom-right.

B.1.3 Topics

Table 5: Number of images tagged with a certain topic in the M5-VGR (A) and M5-VLOD (B) datasets.

Topic	Language																							
	Amharic		Berber		Bengali		German		English		Filipino		Hausa		Hindi		Russian		Swahili		Thai		Zulu	
	A	B	A	B	A	B	A	B	A	B	A	B	A	B	A	B	A	B	A	B	A	B	A	B
armchair	1	2	1	1	1	1	1	1	1	2	3	1	1	1	1	1	2	1	3	1	1	1	1	1
backyard	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	3	1
bathroom privacy	1	1	3	3	1	1	2	1	1	1	3	4	1	1	1	1	1	1	2	1	1	1	1	1
bathroom/toilet	1	2	3	1	1	2	1	3	2	1	1	1	2	1	1	1	1	1	3	3	1	1	1	1
bed	1	1	1	2	2	1	1	1	1	3	1	2	4	1	1	1	2	2	1	1	4	1	1	1
bedroom	2	4	1	2	2	2	1	1	1	2	1	1	3	1	1	2	1	1	2	1	1	1	1	1
books	2	2	1	1	1	2	1	1	1	1	2	1	1	1	1	1	2	1	1	1	1	1	1	1
ceiling	1	2	1	1	2	1	1	1	2	2	1	1	1	4	2	1	2	1	2	2	2	1	1	2
children room	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	1	1	2	1	1	1
cleaning equipment	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1
cooking pots	1	1	2	2	2	1	1	1	1	1	2	1	2	2	1	1	2	1	1	2	1	2	1	1
cooking utensils	1	1	3	2	1	3	1	1	1	1	1	1	1	3	2	1	1	1	1	1	2	1	1	1
couch	1	1	1	1	1	1	1	2	2	1	3	3	3	1	1	1	2	1	2	1	1	1	3	1
cups/mugs/glasses	1	1	1	1	1	1	1	1	3	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1
cutlery	1	1	1	1	1	1	1	1	2	1	3	1	1	1	1	2	1	1	1	1	1	2	1	1
dish racks	1	1	1	2	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	3	2	1	1
dish washing brush/cloth	2	1	1	1	3	1	1	1	1	3	1	1	1	1	3	2	1	2	1	1	1	1	1	1
dish washing soap	1	1	1	1	1	1	1	1	3	1	1	1	1	3	1	2	2	2	2	2	1	2	1	1
drainage	1	2	1	1	1	1	1	1	1	2	1	2	1	1	1	1	1	1	1	1	1	1	1	1
drinking water	3	4	2	2	1	2	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	4	2	
drying	3	1	1	1	5	1	1	1	1	1	2	2	1	2	3	1	1	1	1	1	1	1	1	1
everyday shoes	1	2	1	2	2	1	3	1	1	1	2	1	2	3	1	1	1	2	1	2	2	1	2	2
family	2	2	4	1	2	1	3	2	1	1	2	1	3	3	1	1	1	2	1	1	2	2	2	2
floor	1	1	3	1	2	1	1	1	1	1	1	1	1	1	1	1	2	1	1	2	1	1	1	1
front door	2	1	4	1	1	1	1	1	1	1	1	1	1	3	2	1	1	1	4	1	3	2	1	1
grains	2	1	1	1	2	1	2	1	1	1	1	1	1	1	1	2	2	2	2	1	1	1	1	1
guest bed	3	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	2	2	1	1	1	1	1	1
hair brush/comb	1	1	1	1	3	1	2	3	1	1	1	1	2	2	1	1	2	1	1	2	1	1	3	1
hand back	1	2	1	1	1	1	1	1	1	1	2	2	3	1	2	1	2	1	1	1	1	2	1	2
hand palm	1	1	3	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	2	2	1	1
hand washing	2	1	3	2	1	1	1	5	1	1	1	4	2	1	1	3	2	2	2	1	2	1	2	3
home	1	1	3	2	2	1	2	1	1	2	1	4	1	1	5	1	1	2	1	2	2	2	1	2
jewelry	1	1	1	1	1	2	1	1	1	2	1	1	1	2	1	1	1	1	2	1	1	1	1	1
kitchen	2	1	1	2	1	1	1	1	4	1	2	2	1	1	1	1	2	2	1	2	2	2	1	1
kitchen sink	1	2	2	2	1	1	4	2	1	2	2	1	1	2	1	2	1	1	1	3	1	1	3	3
light source in kitchen	1	1	2	1	2	3	2	2	3	2	1	1	1	1	3	1	1	1	2	1	1	1	1	1
light source in livingroom	1	2	2	2	1	1	1	1	2	1	1	2	1	1	1	1	2	1	1	1	1	1	1	1
living room	1	1	1	1	1	1	1	2	2	1	3	1	1	1	1	1	2	2	1	1	2	1	1	1
lock on front door	1	1	1	1	1	4	1	1	3	1	1	1	2	1	1	3	1	2	1	1	2	3	1	1
make up	1	1	1	2	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	2	1
meat or fish	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1
medication	1	1	1	1	1	2	1	1	1	2	1	1	2	1	1	1	2	1	1	1	1	1	1	1
most loved item	1	1	1	1	1	2	3	1	2	2	3	3	2	1	2	2	2	2	2	0	1	1	4	4
most loved toy	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	2	1	1
nicest shoes	1	1	1	1	1	2	2	2	2	1	1	1	1	1	2	1	1	1	1	2	2	2	1	1
oven	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	2	2	1	1	1	1	1	1
paper	2	1	1	2	1	2	2	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
pen/pencils	1	1	1	2	1	2	1	1	1	1	1	1	1	1	1	1	1	2	3	1	1	1	1	1
phone	2	2	1	1	2	1	1	1	2	3	1	3	2	1	2	1	2	1	1	3	1	1	2	1
place where eating dinner	1	2	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	2	1	1	2	2	1	1
plate of food	2	1	1	4	1	1	2	3	1	1	2	1	3	1	1	1	1	2	1	1	1	1	1	1
plates	2	1	1	1	2	2	1	1	1	1	1	3	1	1	1	1	1	2	1	1	2	1	1	1
play area	1	1	2	2	1	1	1	1	1	1	2	1	1	2	1	1	1	1	1	1	1	1	2	1
power outlet	1	2	1	1	1	1	3	4	2	1	1	1	1	1	3	1	1	1	1	1	1	4	2	1
refrigerator	1	1	1	1	2	1	4	4	1	3	1	3	1	1	1	3	2	1	1	1	1	1	1	1
roof	2	1	1	1	1	3	2	2	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1
shampoo	1	2	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	2	2	1	1	2	1	1
shower	1	1	1	2	1	1	1	1	1	2	1	2	1	1	1	1	2	2	1	1	1	1	1	1
sitting area	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
soap for hands and body	1	1	2	2	2	2	1	2	1	1	1	1	2	1	2	1	2	1	1	1	2	2	1	1
social drink	1	1	1	1	1	1	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2
sofa	1	1	1	1	1	1	1	1	1	2	1	1	2	4	1	1	1	1	1	1	2	1	1	1
source of cool	1	1	1	1	2	1	1	1	2	1	1	1	2	1	2	2	1	1	1	1	1	1	1	2
spices	1	2	1	1	2	1	1	3	3	3	2	1	1	1	2	1	2	2	1	2	1	2	1	3
storage room	1	2	1	1	1	2	1	1	1	1	5	1	2	1	1	2	1	2	1	1	1	1	1	1
stove/hob	2	1	1	2	1	1	1	1	1	3	1	1	1	5	1	2	1	1	2	3	2	1	1	4
street detail	4	1	1	1	1	1	1	1	1	3	2	1	1	2	1	2	2	1	2	1	2	1	1	1
street view	1	1	1	2	1	4	2	1	1	1	1	1	2	1	3	1	1	2	1	2	2	1	2	2
switch on/off	2	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	2	3	1	1
table with food	2	4	1	1	1	1	2	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
teeth	1	1	1	1	1	1	1	1	1	2	1	2	3	1	1	1	2	2	2	2	1	2	1	1
toilet	1	2	1	2	1	1	1	1	1	1	2	1	1	1	1	1	2	1	1	2	1	1	1	1
toilet paper	3	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	3	2
tooth paste	2	1	1	1	1	2	2	3	2	2	2	1	1	1	4	1	2	2	3	1	1	1	2	3
toothbrush	1	2	1	1	1	1	3	2	1	1	2	1	1	2	1	1	1	2	1	3	1	1	3	3
toys	2	1	2	1	3	5	1	1	1	3	1	2	2	1	1	4	2	3	1	1	2	3	1	1
trash/waste	1	1	1	1	1	1	1	1	3	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1
tv	1	1	1	2	3	2	1	1	2	2	1	1	1	1	1	1	2	1	1	1	1	2	4	6
vegetables	1	2	2	1	2	1	1	1	3	1	1	1	1	1	1	1	2	2	1	1	1	1	1	1
wall	2	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
wall clock	2	1	1	2	1	1	1	1	1	1	1	4	1	1	1	2	2	1	1	1	1	2	1	1
wall decoration	1	2	1	1	1	1	2	1	1	1	1	1	2	2	1	1	1	2	2	0	2	1	1	1
wall inside	1	2	1	1	1	1	2	1	1	1	1	2	2	1	1	1	1	2	1	1	1	1	1	1
wardrobe	1	3	2	1	2	1	1	2	1	1	2	1	1	1	2	2	1	1	2	2	2	2	1	1
washing clothes/cleaning	1	1	1	2	1	1	1	1	1	2	1	1	1	1	3	1	1	1	4	4	1	3	1	1
washing detergent	2	1	1	1	1	2	1	1	1	1	1	2	1	1	1	2	1	1	2	2	1	2	1	1
water outlet	2	1	3	2	1	1	1	2	2	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1

B.2 Dataset Language Details

Table 6: Language support of the datasets considered in this work. More details one the languages are reported in Table 7.

Language	Script	MaXM	xGQA	XNLVI	MaRVL	M5-VLOD	M5-VGR	xFlickrCO	XM3600
Amharic	Ethiopic	no	no	no	no	yes	yes	no	no
Arabic	Arabic	no	no	yes	no	no	no	no	yes
Bengali	Bengali	no	yes	no	no	yes	yes	no	yes
Berber	Tifinagh	no	no	no	no	yes	yes	no	no
Chinese	Hanzi	yes	yes	no	yes	no	no	yes	yes
Croatian	Latin	no	no	no	no	no	no	no	yes
Czech	Latin	no	no	no	no	no	no	no	yes
Danish	Latin	no	no	no	no	no	no	no	yes
Dutch	Latin	no	no	no	no	no	no	no	yes
English	Latin	yes	yes	yes	no	yes	yes	yes	yes
Filipino	Latin	no	no	no	no	yes	yes	no	yes
Finnish	Latin	no	no	no	no	no	no	no	yes
French	Latin	yes	no	yes	no	no	no	no	yes
German	Latin	no	yes	no	no	yes	yes	yes	yes
Greek	Greek	no	no	no	no	no	no	no	yes
Hausa	Latin	no	no	no	no	yes	yes	no	no
Hebrew	Hebrew	yes	no	no	no	no	no	no	yes
Hindi	Devanagari	yes	no	no	no	yes	yes	no	yes
Hungarian	Latin	no	no	no	no	no	no	no	yes
Indonesian	Latin	no	yes	no	yes	no	no	yes	yes
Italian	Latin	no	no	no	no	no	no	no	yes
Japanese	Japanese	no	no	no	no	no	no	yes	yes
Korean	Hangul	no	yes	no	no	no	no	no	yes
Maori	Latin	no	no	no	no	no	no	no	yes
Norwegian	Latin	no	no	no	no	no	no	no	yes
Persian	Perso-Arabic	no	no	no	no	no	no	no	yes
Polish	Latin	no	no	no	no	no	no	no	yes
Portuguese	Latin	no	yes	no	no	no	no	no	yes
Quechua	Latin	no	no	no	no	no	no	no	yes
Romanian	Latin	yes	no	no	no	no	no	no	yes
Russian	Cyrillic	no	yes	yes	no	yes	yes	yes	yes
Spanish	Latin	no	no	yes	no	no	no	yes	yes
Swahili	Latin	no	no	no	yes	yes	yes	no	yes
Swedish	Latin	no	no	no	no	no	no	no	yes
Tamil	Tamil	no	no	no	yes	no	no	no	no
Telugu	Telugu	no	no	no	no	no	no	no	yes
Thai	Thai	yes	no	no	no	yes	yes	no	yes
Turkish	Latin	no	no	no	yes	no	no	yes	yes
Ukrainian	Cyrillic	no	no	no	no	no	no	no	yes
Vietnamese	Latin	no	no	no	no	no	no	no	yes
Zulu	Latin	no	no	no	no	yes	yes	no	no
Unique Languages		7	8	5	5	12	12	8	36
Unique Scripts		4	5	3	3	7	7	4	12

B.3 Language Details

Table 7: Details and statistics of languages comprised in the datasets of this benchmark. The continent and subregion columns refer to the content or subregion where the respective language is mostly spoken. The number of speakers is an estimate of the number of L1 and L2 speakers based on different public sources such as Wikipedia[10], Ethnologue [11], and Statista[12]. The "Taxonomy" column indicates the taxonomy class of the language based on Joshi et al. (2020).

Language	ISO 639	Lang. Family	Script	Continent	Subregion	Taxonomy	Speakers / 10^6
Arabic	ar	Afro-Asiatic	Arabic	Afrika & Asia	North Africa & Middle East	5	630.00
Chinese	zh	Sino-Tibetan	Hanzi	Asia	Northeastern Asia	5	1330.00
English	en	Indo-European	Latin	America	North America	5	1457.00
French	fr	Indo-European	Latin	Europe	Western Europe	5	310.00
German	de	Indo-European	Latin	Europe	Western Europe	5	175.00
Japanese	ja	Japonic	Japanese	Asia	Northeastern Asia	5	128.00
Spanish	es	Indo-European	Latin	Europe	Southern Europe	5	600.00
Croatian	hr	Indo-European	Latin	Europe	Central & Eastern Europe	4	6.80
Czech	cs	Indo-European	Latin	Europe	Central & Eastern Europe	4	11.00
Dutch	nl	Indo-European	Latin	Europe	Western Europe	4	30.00
Finnish	fi	Uralic	Latin	Europe	Northern Europe	4	5.80
Hindi	hi	Indo-European	Devanagari	Asia	Central & South Asia	4	600.00
Hungarian	hu	Uralic	Latin	Europe	Central & Eastern Europe	4	17.00
Italian	it	Indo-European	Latin	Europe	Southern Europe	4	68.00
Korean	ko	Koreanic	Hangul	Asia	Northeastern Asia	4	82.00
Persian	fa	Indo-European	Perso-Arabic	Asia	Middle East	4	130.00
Polish	pl	Indo-European	Latin	Europe	Central & Eastern Europe	4	41.00
Portuguese	pt	Indo-European	Latin	Europe & America	Southern Europe & South America	4	360.00
Russian	ru	Indo-European	Cyrillic	Asia	Central Asia	4	260.00
Swedish	sv	Indo-European	Latin	Europe	Northern Europe	4	13.00
Turkish	tr	Turkic	Latin	Asia	Middle East	4	90.00
Vietnamese	vi	Austroasiatic	Latin	Asia	Southeastern Asia	4	85.00
Bengali	bn	Indo-European	Bengali	Asia	Central & South Asia	3	270.00
Danish	da	Indo-European	Latin	Europe	Western Europe	3	6.00
Filipino	fil	Austronesian	Latin	Asia	Southeastern Asia	3	83.00
Greek	el	Indo-European	Greek	Europe	Central & Eastern Europe	3	13.50
Hebrew	he & iw	Afro-Asiatic	Hebrew	Asia	Middle East	3	9.00
Indonesian	id	Austronesian	Latin	Asia	Southeastern Asia	3	300.00
Romanian	ro	Indo-European	Latin	Europe	Central & Eastern Europe	3	28.50
Tamil	ta	Dravidian	Tamil	Asia	Central & South Asia	3	86.00
Thai	th	Kra-Dai	Thai	Asia	Southeastern Asia	3	80.00
Ukrainian	uk	Indo-European	Cyrillic	Europe	Central & Eastern Europe	3	32.80
Amharic	am	Afro-Asiatic	Ethiopic	Africa	Eastern Africa	2	57.00
Hausa	ha	Afro-Asiatic	Latin	Africa	Western Africa	2	79.00
Swahili	sw	Niger-Congo	Latin	Africa	Eastern Africa	2	73.00
Zulu	zu	Niger-Congo	Latin	Africa	Southern Africa	2	28.00
Maori	mi	Austronesian	Latin	Australia & Oceania	Australia & Oceania	1	0.19
Norwegian	no	Indo-European	Latin	Europe	Northern Europe	1	4.32
Quechua	quz	Quechuan	Latin	America	South America	1	9.00
Telugu	te	Dravidian	Telugu	Asia	Central & South Asia	1	96.00
Berber	ber	Afro-Asiatic	Tifinagh	Africa	Northern Africa	0	26.20

[3] https://en.wikipedia.org/wiki/List_of_languages_by_total_number_of_speakers

[4] https://www.ethnologue.com/

[5] https://www.statista.com/statistics/266808/the-most-spoken-languages-worldwide/

C Model Details

Table 8: Architectural details of the LMMs evaluated in this study. The columns LM, VM, and ML are "**L**anguage **M**odel", "**V**ision **M**odel", and "**M**apping **M**odules", respectively, and show the number of parameters of the particular module. "|Total|" shows all parameters of the model. Note that we report friedly names of the models which are enriched with hyperlinks pointing to the respective Huggingface repositories (when viewed digitally). For Gemini Pro Vision and GPT-4 Vision, we used the `gemini-1.0-pro-vision` and `gpt-4-1106-vision-preview` variants, respectively.

Model	LM	VM	MM	|Total|	|LM|	|VM|	|MM|
MiniCPM-V [27; 50]	MiniCPM-2B	SigLIP 400M	MLP	3.43B	3.01B	397.75M	29.51M
mBliP mT0 [22]	Flan-T5-XL	EVA01 CLIP-ViT-g	QFormer	4.84B	3.74B	985.95M	106.71M
Yi-VL 6B [5]	Yi-6B-Chat	CLIP-ViT-H-14	MLP	6.71B	5.80B	631.75M	22.04M
LLaVA 1.6 7B [38]	Vicuna-7B-v1.5	CLIP-ViT-L	MLP	6.76B	6.61B	303.51M	20.98M
LLaVA 1.5 7B [39]	Vicuna-7B-v1.5	CLIP-ViT-L	MLP	7.06B	6.74B	303.51M	20.98M
BakLLaVA [39]	Mistral 7B v0.1	CLIP-ViT-L	MLP	7.57B	7.24B	303.51M	20.98M
mBliP BloomZ [22]	BloomZ 7B	EVA01 CLIP-ViT-g	QFormer	8.16B	7.07B	985.95M	108.29M
Qwen-VL [9]	Qwen-7B	CLIP-VIT-bigG	CrossAttn	9.66B	7.10B	1.94B	80.00M
OmniLMM 12B [50]	Zephyr 7B β	EVA02 CLIP ViT-E	MLP	11.61B	7.24B	4.28B	93.36M
LLaVA 1.6 13B [38]	Vicuna-13B-v1.5	CLIP-ViT-L	MLP	13.05B	12.85B	303.51M	31.47M
LLaVA 1.5 13B [39]	Vicuna-13B-v1.5	CLIP-ViT-L	MLP	13.35B	13.02B	303.51M	31.47M
CogVLM [48]	Vicuna-7B-v1.5	EVA02 CLIP ViT-E	CrossAttn	17.64B	6.74B	4.28B	6.62B
InternVL V1.1 [15]	Llama-2-13B	InternViT 6B	MLP	19.11B	13.12B	5.91B	91.79M
LLaVA 1.6 34B [38]	Nous-Hermes-2-Yi-34B	CLIP-ViT-L	MLP	34.45B	33.93B	303.51M	58.73M
Yi-VL 34B [5]	Yi-34B-Chat	CLIP-ViT-H	MLP	35.08B	33.93B	631.75M	60.60M
InternVL V1.2+ [15]	Nous-Hermes-2-Yi-34B	InternViT-6B V1-2	MLP	40.07B	34.39B	5.54B	143.17M
Gemini Pro Vision [7]	?	?	?	?	?	?	?
GPT-4 Vision [40]	?	?	?	?	?	?	?

D Results Details

D.1 General Results

D.1.1 xGQA

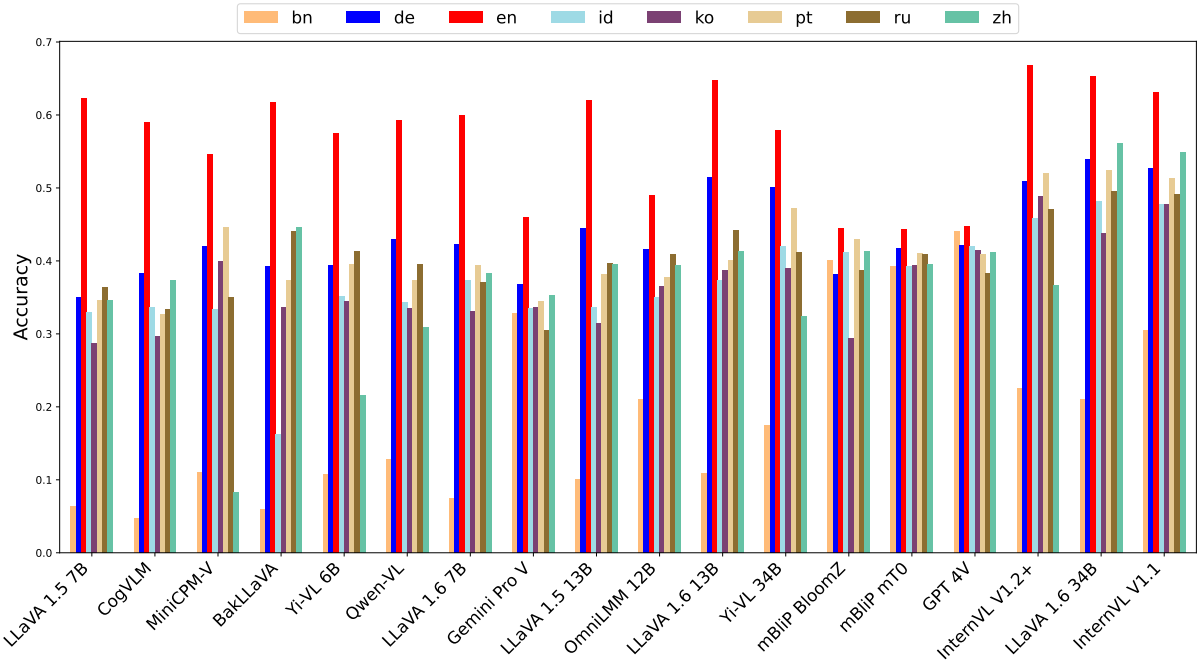


Figure 32: A bar plot showing the average accuracy per language and model on the xGQA dataset. The models on the x-Axis are ordered by their average score across all languages so that the best performing model is on the right and the worst is on the left.

Table 9: The average accuracy per language and model on the xGQA dataset. The column "NEA" stands for the average of Non-English languages.

Model	Language								
	bn	de	en	id	ko	pt	ru	zh	NEA
LLaVA 1.5 7B	0.06	0.35	0.62	0.33	0.29	0.35	0.36	0.35	0.30
CogVLM	0.05	0.38	0.59	0.34	0.30	0.33	0.33	0.37	0.30
MiniCPM-V	0.11	0.42	0.55	0.33	0.40	0.45	0.35	0.08	0.31
BakLLaVA	0.06	0.39	0.62	0.16	0.34	0.37	0.44	0.45	0.32
Yi-VL 6B	0.11	0.39	0.57	0.35	0.34	0.39	0.41	0.22	0.32
Qwen-VL	0.13	0.43	0.59	0.34	0.34	0.37	0.39	0.31	0.33
LLaVA 1.6 7B	0.07	0.42	0.60	0.37	0.33	0.39	0.37	0.38	0.34
Gemini Pro V	0.33	0.37	0.46	0.34	0.34	0.34	0.31	0.35	0.34
LLaVA 1.5 13B	0.10	0.44	0.62	0.34	0.31	0.38	0.40	0.40	0.34
OmniLMM 12B	0.21	0.42	0.49	0.35	0.37	0.38	0.41	0.39	0.36
LLaVA 1.6 13B	0.11	0.52	0.65	0.37	0.39	0.40	0.44	0.41	0.38
Yi-VL 34B	0.18	0.50	0.58	0.42	0.39	0.47	0.41	0.32	0.38
mBliP BloomZ	0.40	0.38	0.44	0.41	0.29	0.43	0.39	0.41	0.39
mBliP mT0	0.39	0.42	0.44	0.39	0.39	0.41	0.41	0.40	0.40
GPT 4V	0.44	0.42	0.45	0.42	0.41	0.41	0.38	0.41	0.41
InternVL V1.2+	0.22	0.51	0.67	0.46	0.49	0.52	0.47	0.37	0.43
LLaVA 1.6 34B	0.21	0.54	0.65	0.48	0.44	0.52	0.50	0.56	0.46
InternVL V1.1	0.31	0.53	0.63	0.48	0.48	0.51	0.49	0.55	0.48
Average	0.19	0.43	0.57	0.37	0.37	0.41	0.40	0.37	0.37

D.1.2 MaXM

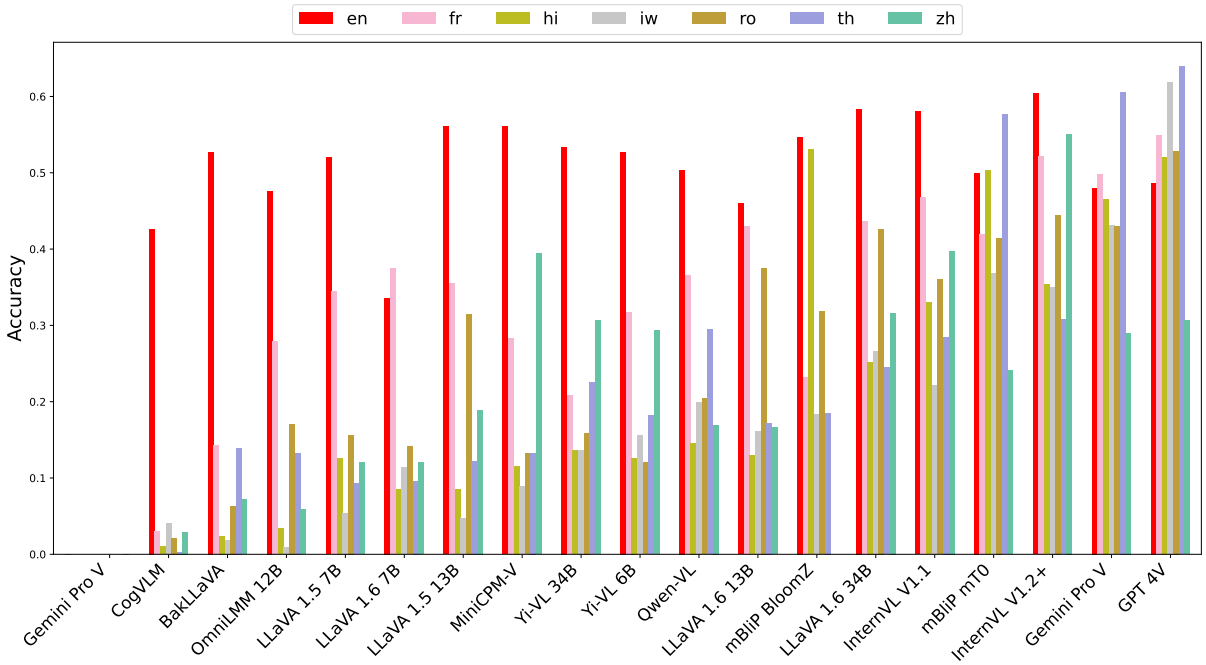


Figure 33: A bar plot showing the average accuracy per language and model on the MaXM dataset. The models on the x-Axis are ordered by their average score across all languages so that the best performing model is on the right and the worst is on the left.

Table 10: The average accuracy per language and model on the MaXM dataset. The column "NEA" stands for the average of Non-English languages.

Model	Language							
	en	fr	hi	iw	ro	th	zh	NEA
CogVLM	0.43	0.03	0.01	0.04	0.02	0.00	0.03	0.02
BakLLaVA	0.53	0.14	0.02	0.02	0.06	0.14	0.07	0.08
OmniLMM 12B	0.48	0.28	0.03	0.01	0.17	0.13	0.06	0.11
LLaVA 1.5 7B	0.52	0.34	0.13	0.05	0.16	0.09	0.12	0.15
LLaVA 1.6 7B	0.34	0.38	0.09	0.11	0.14	0.10	0.12	0.16
LLaVA 1.5 13B	0.56	0.35	0.09	0.05	0.32	0.12	0.19	0.19
MiniCPM-V	0.56	0.28	0.12	0.09	0.13	0.13	0.39	0.19
Yi-VL 34B	0.53	0.21	0.14	0.14	0.16	0.23	0.31	0.20
Yi-VL 6B	0.53	0.32	0.13	0.16	0.12	0.18	0.29	0.20
Qwen-VL	0.50	0.37	0.15	0.20	0.20	0.29	0.17	0.23
LLaVA 1.6 13B	0.46	0.43	0.13	0.16	0.38	0.17	0.17	0.24
mBliP BloomZ	0.55	0.23	0.53	0.18	0.32	0.19	0.42	0.31
LLaVA 1.6 34B	0.58	0.44	0.25	0.27	0.43	0.25	0.32	0.32
InternVL V1.1	0.58	0.47	0.33	0.22	0.36	0.28	0.40	0.34
mBliP mT0	0.50	0.42	0.50	0.37	0.41	0.58	0.24	0.42
InternVL V1.2+	0.60	0.52	0.35	0.35	0.44	0.31	0.55	0.42
Gemini Pro V	0.48	0.50	0.47	0.43	0.43	0.61	0.29	0.45
GPT 4V	0.49	0.55	0.52	0.62	0.53	0.64	0.31	0.53
Average	0.51	0.35	0.22	0.19	0.27	0.25	0.24	0.25

D.1.3 XVNLI

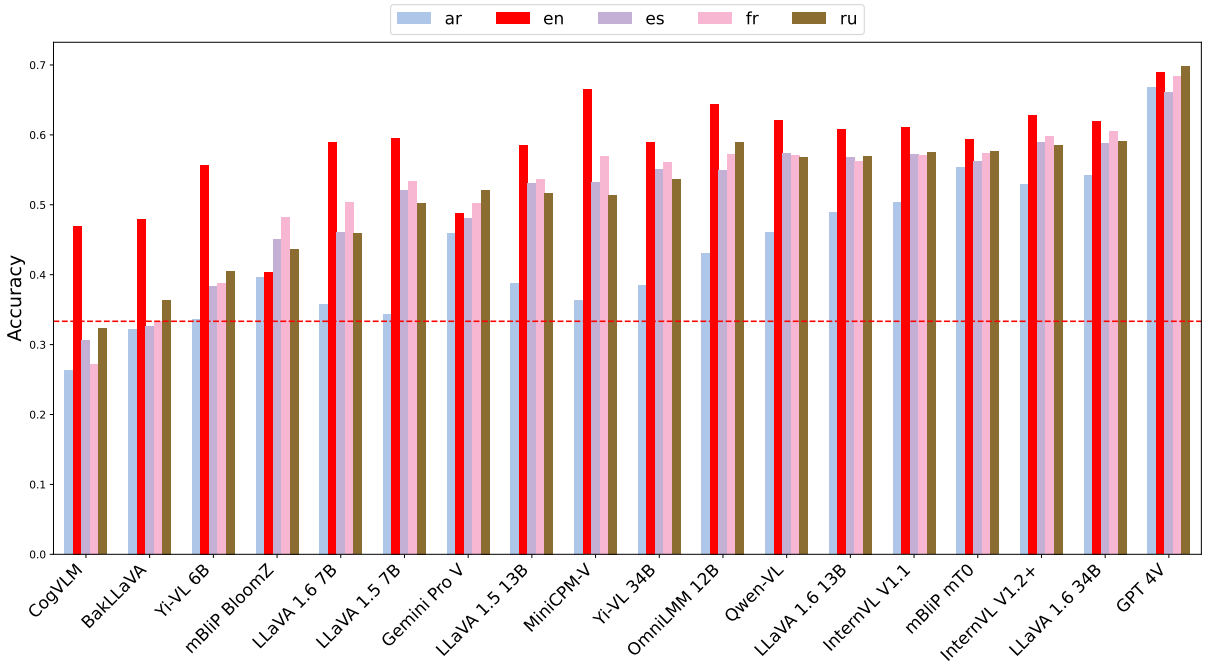


Figure 34: A bar plot showing the average accuracy per language and model on the XVNLI dataset. The models on the x-Axis are ordered by their average score across all languages so that the best performing model is on the right and the worst is on the left.

Table 11: The average accuracy per language and model on the XVNLI dataset. The column "NEA" stands for the average of Non-English languages.

Model	Language					
	ar	en	es	fr	ru	NEA
CogVLM	0.26	0.47	0.31	0.27	0.32	0.29
BakLLaVA	0.32	0.48	0.33	0.33	0.36	0.34
Yi-VL 6B	0.34	0.56	0.38	0.39	0.41	0.38
mBliP BloomZ	0.40	0.40	0.45	0.48	0.44	0.44
LLaVA 1.6 7B	0.36	0.59	0.46	0.50	0.46	0.45
LLaVA 1.5 7B	0.34	0.60	0.52	0.53	0.50	0.47
Gemini Pro V	0.46	0.49	0.48	0.50	0.52	0.49
LLaVA 1.5 13B	0.39	0.59	0.53	0.54	0.52	0.49
MiniCPM-V	0.36	0.66	0.53	0.57	0.51	0.49
Yi-VL 34B	0.39	0.59	0.55	0.56	0.54	0.51
OmniLMM 12B	0.43	0.64	0.55	0.57	0.59	0.54
Qwen-VL	0.46	0.62	0.57	0.57	0.57	0.54
LLaVA 1.6 13B	0.49	0.61	0.57	0.56	0.57	0.55
InternVL V1.1	0.50	0.61	0.57	0.57	0.57	0.56
mBliP mT0	0.55	0.59	0.56	0.57	0.58	0.57
InternVL V1.2+	0.53	0.63	0.59	0.60	0.59	0.58
LLaVA 1.6 34B	0.54	0.62	0.59	0.60	0.59	0.58
GPT 4V	0.67	0.69	0.66	0.68	0.70	0.68
Average	0.43	0.58	0.51	0.52	0.52	0.50

D.1.4 MaRVL

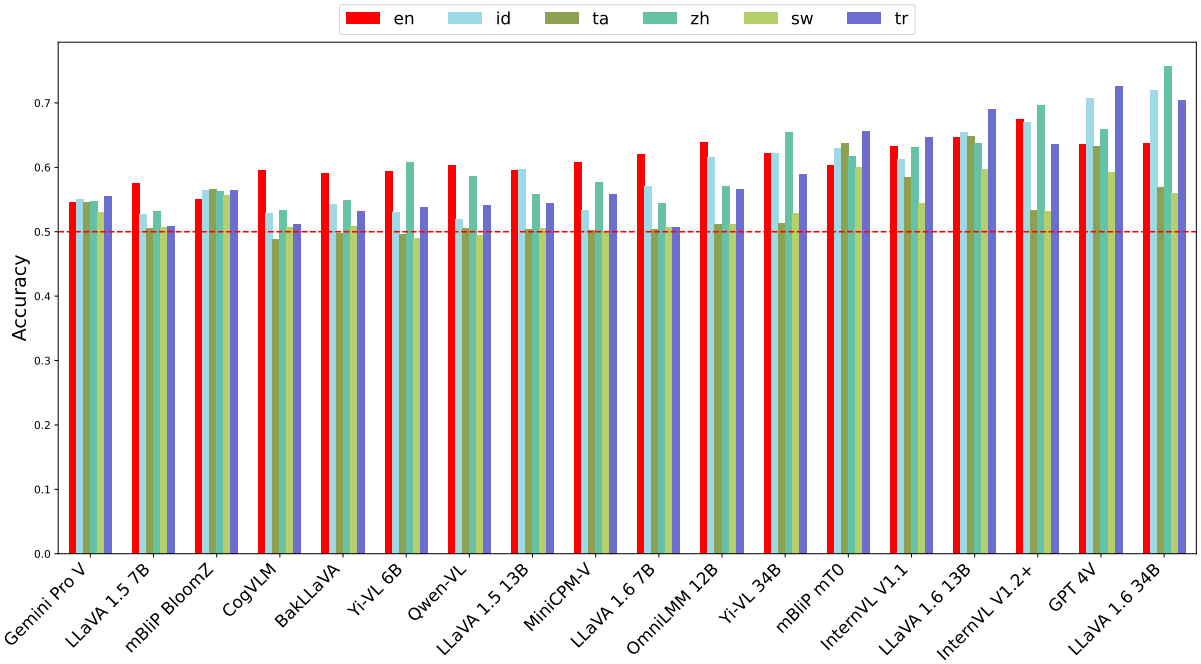


Figure 35: A bar plot showing the average accuracy per language and model on the MaRVL dataset. Note that MaRVL does not contain English data originally and we machine-translated English from the other languages and averaged the results. The models on the x-Axis are ordered by their average score across all languages so that the best performing model is on the right and the worst is on the left.

Table 12: The average accuracy per language and model on the MaRVL dataset. Note that MaRVL does not contain English data originally and we machine-translated English from the other languages and averaged the results. The column "NEA" stands for the average of Non-English languages.

Model	Language						
	en	id	sw	ta	tr	zh	NEA
CogVLM	0.60	0.53	0.51	0.49	0.51	0.53	0.51
LLaVA 1.5 7B	0.57	0.53	0.51	0.51	0.51	0.53	0.52
BakLLaVA	0.59	0.54	0.51	0.50	0.53	0.55	0.53
LLaVA 1.6 7B	0.62	0.57	0.51	0.50	0.51	0.54	0.53
Qwen-VL	0.60	0.52	0.50	0.50	0.54	0.59	0.53
Yi-VL 6B	0.59	0.53	0.49	0.50	0.54	0.61	0.53
MiniCPM-V	0.61	0.53	0.50	0.50	0.56	0.58	0.53
LLaVA 1.5 13B	0.60	0.60	0.51	0.50	0.54	0.56	0.54
Gemini Pro V	0.55	0.55	0.53	0.55	0.56	0.55	0.55
OmniLMM 12B	0.64	0.62	0.51	0.51	0.57	0.57	0.56
mBliP BloomZ	0.55	0.57	0.56	0.57	0.56	0.56	0.56
Yi-VL 34B	0.62	0.62	0.53	0.51	0.59	0.65	0.58
InternVL V1.1	0.63	0.61	0.54	0.58	0.65	0.63	0.60
InternVL V1.2+	0.68	0.67	0.53	0.53	0.64	0.70	0.61
mBliP mT0	0.60	0.63	0.60	0.64	0.66	0.62	0.63
LLaVA 1.6 13B	0.65	0.66	0.60	0.65	0.69	0.64	0.65
LLaVA 1.6 34B	0.64	0.72	0.56	0.57	0.70	0.76	0.66
GPT 4V	0.64	0.71	0.59	0.63	0.73	0.66	0.66
Average	0.61	0.60	0.53	0.54	0.59	0.60	0.57

D.1.5 M5-VGR

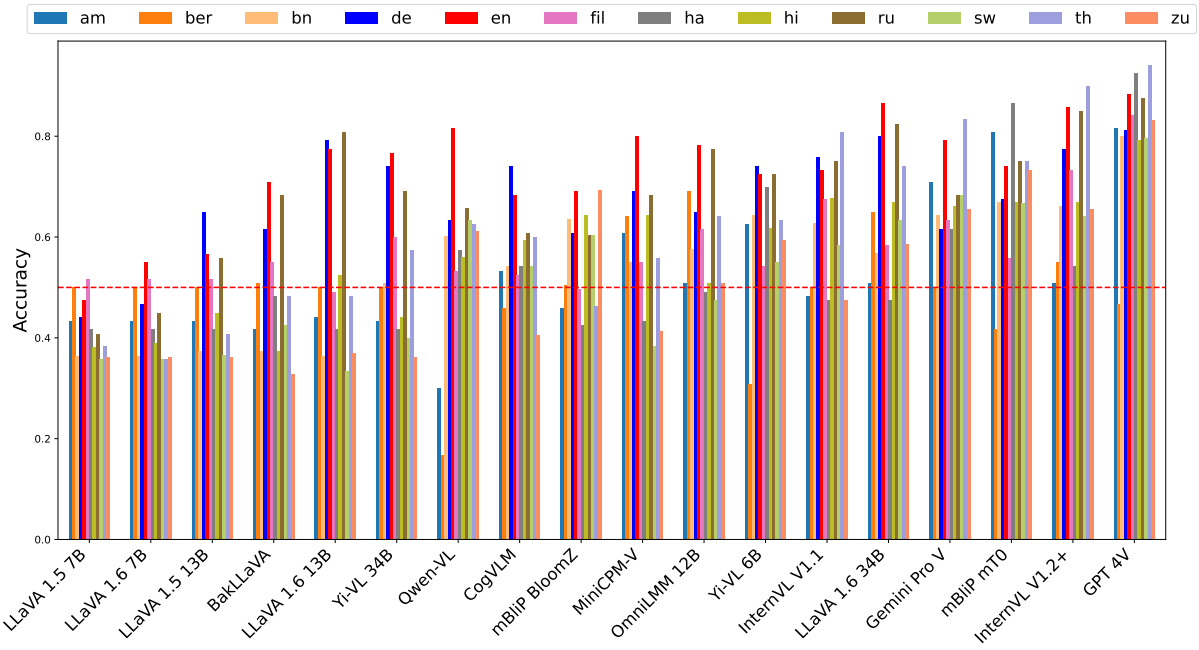


Figure 36: A bar plot showing the average accuracy per language and model on the M5-VGR dataset. The models on the x-Axis are ordered by their average score across all languages so that the best performing model is on the right and the worst is on the left.

Table 13: The average accuracy per language and model on the M5-VGR dataset. The column "NEA" stands for the average of Non-English languages.

Model	Language												
	am	ber	bn	de	en	fil	ha	hi	ru	sw	th	zu	NEA
LLaVA 1.5 7B	0.43	0.50	0.36	0.44	0.47	0.52	0.42	0.38	0.41	0.36	0.38	0.36	0.42
LLaVA 1.6 7B	0.43	0.50	0.36	0.47	0.55	0.52	0.42	0.39	0.45	0.36	0.36	0.36	0.42
LLaVA 1.5 13B	0.43	0.50	0.37	0.65	0.57	0.52	0.42	0.45	0.56	0.37	0.41	0.36	0.46
BakLLaVA	0.42	0.51	0.37	0.62	0.71	0.55	0.48	0.37	0.68	0.42	0.48	0.33	0.48
LLaVA 1.6 13B	0.44	0.50	0.36	0.79	0.78	0.49	0.42	0.53	0.81	0.33	0.48	0.37	0.50
Yi-VL 34B	0.43	0.50	0.51	0.74	0.77	0.60	0.42	0.44	0.69	0.40	0.57	0.36	0.52
Qwen-VL	0.30	0.17	0.60	0.63	0.82	0.53	0.57	0.56	0.66	0.63	0.62	0.61	0.54
CogVLM	0.53	0.46	0.54	0.74	0.68	0.53	0.54	0.59	0.61	0.54	0.60	0.41	0.55
mBliP BloomZ	0.46	0.50	0.64	0.61	0.69	0.50	0.42	0.64	0.60	0.60	0.46	0.69	0.56
MiniCPM-V	0.61	0.64	0.55	0.69	0.80	0.55	0.43	0.64	0.68	0.38	0.56	0.41	0.56
OmniLMM 12B	0.51	0.69	0.58	0.65	0.78	0.62	0.49	0.51	0.78	0.47	0.64	0.51	0.59
Yi-VL 6B	0.62	0.31	0.64	0.74	0.72	0.54	0.70	0.62	0.72	0.55	0.63	0.59	0.61
InternVL V1.1	0.48	0.50	0.63	0.76	0.73	0.68	0.47	0.68	0.75	0.58	0.81	0.47	0.62
LLaVA 1.6 34B	0.51	0.65	0.57	0.80	0.87	0.58	0.47	0.67	0.82	0.63	0.74	0.59	0.64
Gemini Pro V	0.71	0.50	0.64	0.62	0.79	0.63	0.62	0.66	0.68	0.68	0.83	0.66	0.66
InternVL V1.2+	0.51	0.55	0.66	0.78	0.86	0.73	0.54	0.67	0.85	0.64	0.90	0.66	0.68
mBliP mT0	0.81	0.42	0.67	0.68	0.74	0.56	0.87	0.67	0.75	0.67	0.75	0.73	0.69
GPT 4V	0.82	0.47	0.80	0.81	0.88	0.84	0.93	0.79	0.88	0.80	0.94	0.83	0.81
Average	0.53	0.49	0.55	0.68	0.73	0.58	0.53	0.57	0.69	0.52	0.62	0.52	0.57

D.1.6 M5-VLOD

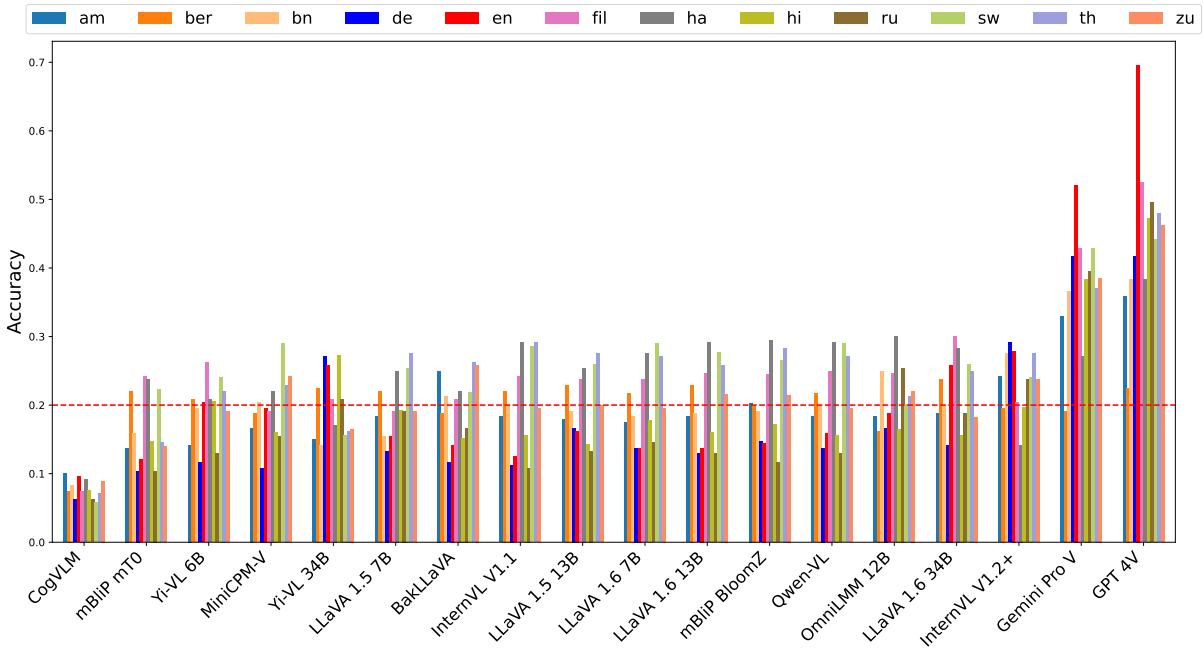


Figure 37: A bar plot showing the average accuracy per language and model on the M5-VLOD dataset. The models on the x-Axis are ordered by their average score across all languages so that the best performing model is on the right and the worst is on the left.

Table 14: The average accuracy per language and model on the M5-VLOD dataset. The column "NEA" stands for the average of Non-English languages.

Model	Language												
	am	ber	bn	de	en	fil	ha	hi	ru	sw	th	zu	NEA
CogVLM	0.10	0.07	0.08	0.06	0.10	0.07	0.09	0.08	0.06	0.06	0.07	0.09	0.08
mBliP mT0	0.14	0.22	0.16	0.10	0.12	0.24	0.24	0.15	0.10	0.22	0.15	0.14	0.17
Yi-VL 6B	0.14	0.21	0.20	0.12	0.20	0.26	0.21	0.21	0.13	0.24	0.22	0.19	0.19
Yi-VL 34B	0.15	0.22	0.14	0.27	0.26	0.21	0.17	0.27	0.21	0.16	0.16	0.17	0.19
MiniCPM-V	0.17	0.19	0.20	0.11	0.20	0.19	0.22	0.16	0.15	0.29	0.23	0.24	0.20
LLaVA 1.5 7B	0.18	0.22	0.15	0.13	0.15	0.19	0.25	0.19	0.19	0.25	0.27	0.19	0.20
BakLLaVA	0.25	0.19	0.21	0.12	0.14	0.21	0.22	0.15	0.17	0.22	0.26	0.26	0.20
LLaVA 1.5 13B	0.18	0.23	0.19	0.17	0.16	0.24	0.25	0.14	0.13	0.26	0.28	0.20	0.21
InternVL V1.1	0.18	0.22	0.20	0.11	0.12	0.24	0.29	0.16	0.11	0.29	0.29	0.19	0.21
LLaVA 1.6 7B	0.17	0.22	0.18	0.14	0.14	0.24	0.27	0.18	0.15	0.29	0.27	0.19	0.21
LLaVA 1.6 13B	0.18	0.23	0.19	0.13	0.14	0.25	0.29	0.16	0.13	0.28	0.26	0.22	0.21
Qwen-VL	0.18	0.22	0.20	0.14	0.16	0.25	0.29	0.16	0.13	0.29	0.27	0.19	0.21
mBliP BloomZ	0.20	0.20	0.19	0.15	0.14	0.24	0.29	0.17	0.12	0.26	0.28	0.21	0.21
OmniLMM 12B	0.18	0.16	0.25	0.17	0.19	0.25	0.30	0.17	0.25	0.20	0.21	0.22	0.21
LLaVA 1.6 34B	0.19	0.24	0.20	0.14	0.26	0.30	0.28	0.16	0.19	0.26	0.25	0.18	0.22
InternVL V1.2+	0.24	0.20	0.28	0.29	0.28	0.20	0.14	0.20	0.24	0.24	0.28	0.24	0.23
Gemini Pro V	0.33	0.19	0.37	0.42	0.52	0.43	0.27	0.38	0.40	0.43	0.37	0.39	0.36
GPT 4V	0.36	0.22	0.38	0.42	0.70	0.53	0.38	0.47	0.50	0.44	0.48	0.46	0.42
Average	0.20	0.20	0.21	0.18	0.22	0.25	0.25	0.20	0.19	0.26	0.26	0.22	0.22

D.1.7 xFlickrCO

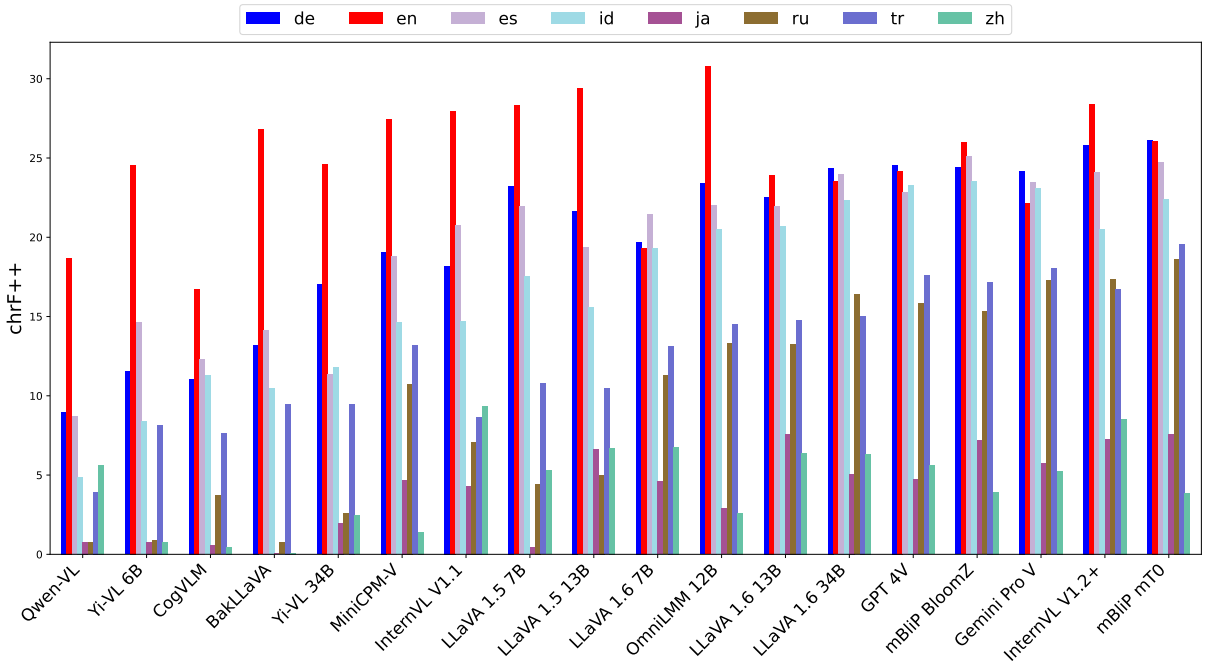


Figure 38: A bar plot showing the average chrF++ score per language and model on the xFlickrCO dataset. The models on the x-Axis are ordered by their average score across all languages so that the best performing model is on the right and the worst is on the left.

Table 15: The average chrF++ score per language and model on the xFlickrCO dataset. The column "NEA" stands for the average of Non-English languages.

Model	Language								
	de	en	es	id	ja	ru	tr	zh	NEA
Qwen-VL	9.00	18.68	8.69	4.88	0.77	0.74	3.91	5.62	4.80
Yi-VL 6B	11.53	24.54	14.61	8.37	0.78	0.90	8.15	0.79	6.45
CogVLM	11.08	16.76	12.32	11.27	0.56	3.71	7.62	0.46	6.72
BakLLaVA	13.21	26.79	14.17	10.48	0.06	0.75	9.49	0.09	6.89
Yi-VL 34B	17.02	24.62	11.36	11.79	2.00	2.57	9.50	2.44	8.10
MiniCPM-V	19.05	27.43	18.81	14.62	4.69	10.73	13.18	1.40	11.78
InternVL V1.1	18.21	27.98	20.74	14.69	4.31	7.07	8.67	9.38	11.87
LLaVA 1.5 7B	23.22	28.32	21.95	17.58	0.44	4.45	10.77	5.29	11.96
LLaVA 1.5 13B	21.66	29.39	19.37	15.59	6.63	5.02	10.45	6.72	12.21
LLaVA 1.6 7B	19.70	19.31	21.48	19.32	4.60	11.27	13.14	6.78	13.75
OmniLMM 12B	23.39	30.76	22.05	20.50	2.89	13.29	14.55	2.59	14.18
LLaVA 1.6 13B	22.55	23.94	21.98	20.73	7.57	13.26	14.79	6.39	15.33
LLaVA 1.6 34B	24.38	23.52	23.98	22.36	5.08	16.40	15.05	6.34	16.23
GPT 4V	24.56	24.17	22.82	23.29	4.73	15.82	17.58	5.60	16.34
mBliP BloomZ	24.39	25.99	25.12	23.56	7.18	15.31	17.16	3.93	16.67
Gemini Pro V	24.17	22.13	23.50	23.10	5.75	17.28	18.03	5.24	16.73
InternVL V1.2+	25.81	28.41	24.13	20.48	7.25	17.34	16.73	8.54	17.18
mBliP mT0	26.10	26.07	24.74	22.41	7.56	18.64	19.58	3.87	17.56
Average	19.95	24.93	19.55	16.95	4.05	9.70	12.69	4.53	12.49

D.1.8 XM3600

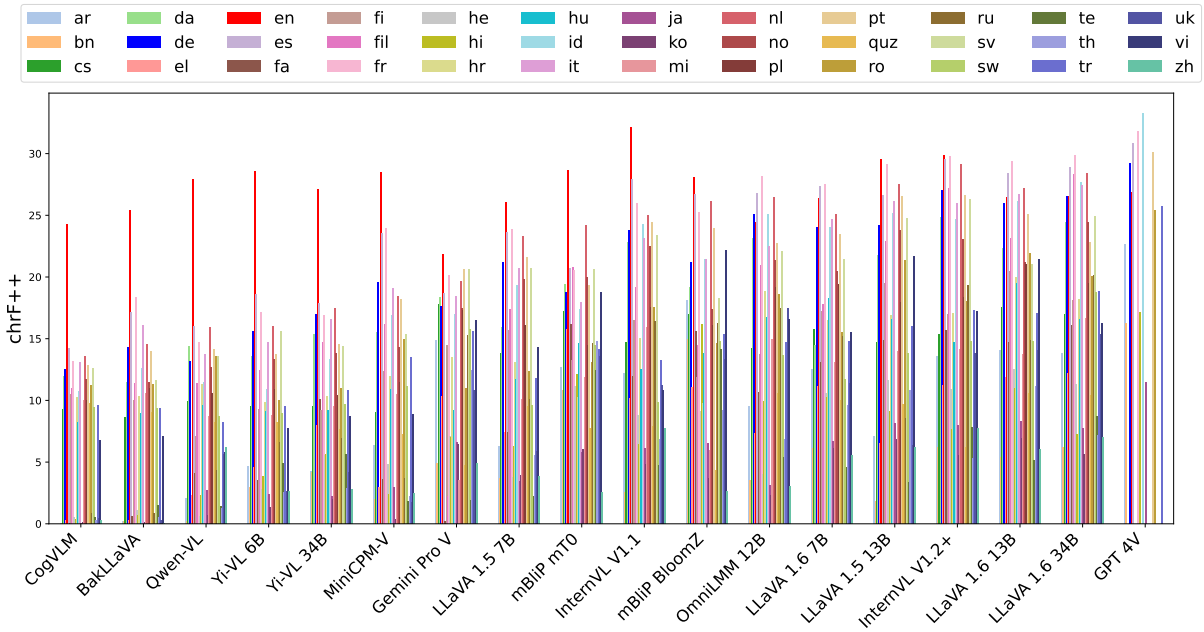


Figure 39: A bar plot showing the average chrF++ score per language and model on the XM3600 dataset. Due to resource restrictions, we evaluated GPT 4V only on a subset of languages. The models on the x-Axis are ordered by their average score across all languages so that the best performing model is on the right and the worst is on the left.

Table 16: The average chrF++ score per language and model on the XM3600 dataset. Due to resource restrictions, we evaluated GPT 4V only on a subset of languages. The column "NEA" stands for the average of Non-English languages.

Model	Language											
	ar	bn	cs	da	de	el	en	es	fa	fi	fil	fr
CogVLM	0.07	0.04	9.30	11.92	12.50	0.25	24.26	14.25	0.02	10.52	10.96	13.18
BakLLaVA	0.21	0.22	8.65	11.45	14.33	0.24	25.39	17.13	0.64	10.02	11.41	18.33
Qwen-VL	2.08	0.17	9.89	14.38	13.14	2.32	27.89	16.00	4.09	7.13	11.36	14.70
Yi-VL 6B	4.65	2.98	9.48	13.55	15.58	4.54	28.59	18.58	3.50	9.29	12.42	17.12
Yi-VL 34B	4.24	4.14	9.52	15.40	17.00	8.00	27.11	17.86	10.06	9.17	14.73	16.93
MiniCPM-V	6.38	1.96	9.05	15.52	19.60	2.98	28.53	23.54	3.57	12.33	16.19	23.98
Gemini Pro V	14.90	4.94	17.79	18.32	17.63	10.36	21.81	18.64	0.21	14.50	2.25	20.15
LLaVA 1.5 7B	6.30	3.71	13.80	15.93	21.18	7.42	26.02	23.60	7.45	15.67	17.38	23.83
mBliP mT0	12.68	10.79	17.20	19.43	18.74	15.76	28.68	20.71	16.19	13.26	20.79	20.52
InternVL V1.1	12.23	2.55	14.74	22.82	23.77	10.20	32.10	27.91	11.94	16.47	19.20	25.95
mBliP BloomZ	18.10	14.92	16.99	19.16	21.17	11.03	28.05	26.73	15.59	11.86	14.47	25.28
OmniLMM 12B	9.48	3.51	14.24	23.15	25.05	7.37	24.42	26.75	10.65	13.78	20.92	28.18
LLaVA 1.6 7B	12.52	6.13	15.79	14.50	24.06	11.11	26.41	27.37	13.07	17.23	17.76	27.48
LLaVA 1.5 13B	7.07	1.80	14.75	21.74	24.15	6.49	29.55	26.59	14.90	19.51	22.91	29.14
InternVL V1.2+	13.59	6.19	15.34	24.85	27.05	11.20	29.84	29.50	15.69	17.01	27.22	29.80
LLaVA 1.6 13B	14.07	5.42	17.51	22.30	25.95	11.90	26.42	28.39	14.72	20.44	23.14	29.42
LLaVA 1.6 34B	13.85	6.20	16.94	24.44	26.51	12.17	26.52	28.90	16.09	18.08	28.35	29.83
GPT 4V	22.67	16.27	-	-	29.24	-	26.89	30.86	-	-	-	31.82
Average	9.73	5.11	12.83	17.16	20.92	7.41	27.14	23.52	8.80	13.13	16.19	23.65

Model	Language											
	he	hi	hr	hu	id	it	ja	ko	mi	nl	no	pl
CogVLM	0.52	0.38	10.25	8.25	10.70	13.11	0.07	0.13	10.00	13.59	11.73	9.98
BakLLaVA	1.07	0.71	10.33	8.98	12.59	16.12	0.07	0.16	10.62	14.56	11.48	10.97
Qwen-VL	0.58	2.32	11.33	9.60	11.50	13.76	2.75	0.70	8.73	15.91	12.64	10.59
Yi-VL 6B	2.78	3.86	9.82	9.12	10.90	14.69	2.40	1.32	8.81	16.04	13.30	10.88
Yi-VL 34B	5.58	5.64	10.31	9.23	13.30	16.55	2.21	2.02	9.55	17.43	13.79	10.40
MiniCPM-V	4.86	2.36	11.96	10.91	16.94	19.06	2.92	0.39	10.49	18.47	14.27	11.51
Gemini Pro V	7.12	6.98	13.48	9.22	16.98	18.44	6.63	6.43	3.55	19.67	17.43	17.29
LLaVA 1.5 7B	3.76	6.29	13.05	11.69	19.33	20.73	3.48	3.93	10.10	23.30	19.79	16.10
mBliP mT0	11.16	12.08	10.26	14.59	17.39	17.92	5.79	6.00	11.88	24.20	19.97	14.49
InternVL V1.1	8.80	6.47	15.05	12.49	24.31	23.13	6.09	4.83	15.93	25.02	22.45	17.58
mBliP BloomZ	9.16	16.18	9.78	13.84	21.44	21.39	6.53	3.67	5.99	26.17	17.35	16.07
OmniLMM 12B	3.99	9.91	18.84	16.72	25.07	22.50	3.16	2.31	14.94	26.47	21.36	19.16
LLaVA 1.6 7B	10.61	10.26	16.52	18.26	24.05	24.71	6.66	6.09	13.12	25.07	20.49	19.38
LLaVA 1.5 13B	11.63	9.13	16.87	16.54	25.13	26.11	8.16	6.86	13.98	27.52	23.77	17.96
InternVL V1.2+	10.88	7.69	17.07	14.70	24.65	25.94	7.96	5.53	14.17	29.11	23.02	18.37
LLaVA 1.6 13B	12.54	11.00	19.99	19.52	26.15	26.66	8.27	6.95	13.73	27.15	21.19	21.03
LLaVA 1.6 34B	11.30	7.27	18.16	16.57	27.69	27.40	7.75	5.60	16.69	28.42	24.45	19.49
GPT 4V	-	17.16	-	-	33.24	-	11.46	-	-	-	-	-
Average	6.46	7.54	12.95	12.23	20.08	19.35	5.13	3.50	10.68	21.01	17.14	14.51

Model	Language												
	pt	quz	ro	ru	sv	sw	te	th	tr	uk	vi	zh	NEA
CogVLM	12.87	9.75	11.23	0.86	12.57	9.41	0.51	0.26	9.58	0.46	6.74	0.29	7.04
BakLLaVA	14.00	9.00	11.30	0.85	11.61	9.37	1.47	0.57	9.36	0.31	7.11	0.03	7.58
Qwen-VL	14.17	8.25	13.60	4.30	13.59	8.75	1.44	1.28	8.26	5.66	5.76	6.20	8.20
Yi-VL 6B	13.77	8.25	10.04	6.57	15.64	8.94	4.93	2.57	9.55	2.65	7.76	2.61	8.82
Yi-VL 34B	14.57	7.64	10.95	6.95	14.42	9.71	5.62	2.92	10.84	4.19	8.74	2.82	9.78
MiniCPM-V	18.21	7.21	14.94	3.69	15.36	11.16	1.83	2.24	13.47	1.74	8.88	2.46	10.30
Gemini Pro V	20.60	4.72	10.98	15.27	20.60	15.80	1.87	12.45	15.62	10.82	16.48	4.88	12.37
LLaVA 1.5 7B	21.57	9.55	12.38	10.08	20.68	9.59	2.23	5.51	11.78	5.84	14.34	3.87	12.44
mBliP mT0	19.35	7.70	13.05	14.63	20.66	14.45	12.42	14.76	14.13	13.60	18.73	2.59	14.80
InternVL V1.1	24.47	7.91	17.55	16.39	23.40	9.82	4.73	6.85	13.22	11.26	10.80	7.76	14.97
mBliP BloomZ	23.93	4.32	14.59	16.25	18.31	14.82	14.12	9.19	15.34	13.35	22.14	2.65	15.20
OmniLMM 12B	22.75	10.61	18.61	17.49	22.09	13.68	5.41	6.84	14.68	17.49	16.58	3.00	15.34
LLaVA 1.6 7B	23.42	10.04	15.55	15.18	21.42	11.69	4.60	9.62	14.81	11.40	15.54	5.58	15.46
LLaVA 1.5 13B	26.51	9.70	21.33	8.53	24.80	13.81	3.39	10.84	15.98	6.36	21.66	6.22	16.05
InternVL V1.2+	26.63	6.20	18.06	19.30	26.27	14.83	7.79	5.30	17.30	13.79	17.22	7.71	17.05
LLaVA 1.6 13B	25.07	10.60	21.96	14.86	21.01	14.80	5.18	11.11	17.03	14.03	21.44	6.02	17.44
LLaVA 1.6 34B	22.85	10.39	20.08	20.11	24.92	18.73	8.70	7.19	18.83	15.36	16.23	7.02	17.79
GPT 4V	30.13	-	25.41	-	-	-	-	-	25.70	-	-	-	24.91
Average	20.83	7.88	15.65	10.63	18.19	11.63	4.79	6.08	14.19	8.24	13.12	3.98	13.64

D.2 Language Fidelity Analysis

Table 17: Pearson correlation coefficients between language fidelity on xFlickrCO and Performance on other datasets.

Dataset	**Language**								
	Avg.	zh	en	de	id	ja	ru	es	tr
xFlickrCO	.91	.85	.65	0.86	.88	.91	.92	.90	.84
XM3600	.81	.74	.63	0.63	.69	.74	.76	.67	.82
MaXM	.55	.17	.43	-	-	-	-	-	-
XVNLI	.51	-	.46	-	-	-	.47	.20	-
MaRVL	.46	.21	.41	-	.50	-	-	-	.50
M5-VGR	.34	-	.11	0.15	-	-	.42	-	-
xGQA	.21	.35	.47	0.08	.37	-	-.04	-	-
M5-VLOD	.14	-	.44	0.20	-	-	.14	-	-