xGQA: Cross-Lingual Visual Question Answering

Anonymous ACL submission

Abstract

Recent advances in multimodal *vision and language* modeling have predominantly focused on the English language, mostly due to the lack of multilingual multimodal datasets to steer modeling efforts. In this work, we address this gap and provide xGQA, a new multilingual evaluation benchmark for the visual question answering task. We extend the established English GQA dataset (Hudson and Manning, 2019) to 7 typologically diverse languages, enabling us to detect and explore crucial challenges in cross-lingual visual question answering. We further propose new adapter-based approaches to adapt multimodal transformer-based models to become multilingual, and—vice versa—multilingual models to become multimodal. Our proposed methods outperform current state-of-the-art multilingual multimodal models (e.g., M^3P) in zero-shot cross-lingual settings, but the accuracy remains low across the board; a performance drop of around 38 accuracy points in target languages showcases the difficulty of zero-shot cross-lingual transfer for this task. Our results suggest that simple cross-lingual transfer of multimodal models yields latent multilingual multimodal misalignment, calling for more sophisticated methods for vision and multilingual language modeling. The xGQA dataset is available online at: [URL].

1 Introduction

Transformer-based architectures (Vaswani et al., 2017) have become ubiquitous in NLP (Devlin et al., 2019; Liu et al., 2019; Conneau et al., 2020, *inter alia*) and in computer vision (CV) (Carion et al., 2020; Dosovitskiy et al., 2021), offering unmatched task performance. Having a shared architecture for multiple modalities opened up possibilities for effective fusion of information, yielding impressive performance gains across various multimodal tasks such as image captioning, phrase grounding, visual question answering, referring expression comprehension and image-text retrieval (Lu et al., 2019; Tan and Bansal, 2019; Li et al., 2020b; Zhang et al., 2021; Ni et al., 2021; Kamath et al., 2021). Yet, progress in this area has been limited mostly to the English language, as the main multimodal datasets consist only of English text. Due to the scarcity of multilingual evaluation benchmarks, there has been limited development of models that tackle this joint problem.

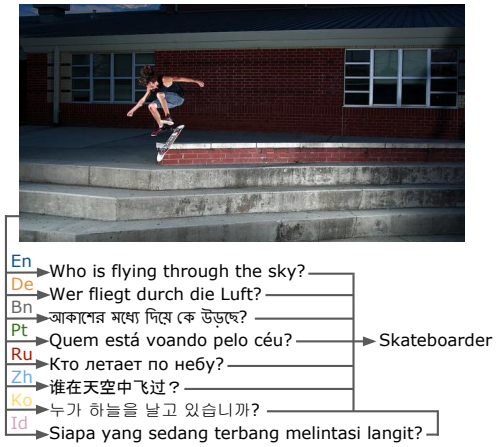


Figure 1: Example taken from the xGQA dataset with the same question uttered in 8 languages.

Aiming to address this gap, in this paper we propose **xGQA**, a multilingual evaluation benchmark for the visual question answering task, extending the monolingual English-only GQA dataset (Hudson and Manning, 2019). For xGQA we manually translate and adapt the balanced GQA test-dev set into 7 new languages from 7 language families, covering 5 distinct scripts; see Figure 1 and Table 1 later. In addition, we provide new fixed data splits to guide cross-lingual few-shot learning experiments, where only a small number of examples in the target language are utilized.

As pretraining is (i) notoriously computationally expensive for high-resource languages and (ii) only limited amounts of multilingual multimodal resources are available, we also propose computationally efficient adapter-based (Houlsby et al.,

2019) approaches as additional baselines for constructing multilingual multimodal models. In a nutshell, we extend multimodal models pretrained only on English text (Zhang et al., 2021) to become multilingual and—vice versa—multilingual models (Devlin et al., 2019) to become multimodal. To this end, we follow the approaches of Artetxe et al. (2020) and Pfeiffer et al. (2020b, 2021) and extend monolingual and multilingual models to new languages and scripts via learning new tokenizers and corresponding word-embedding matrices, as well as adapters for the target languages. To transfer the respective multilingual multimodal adapter-based models to the target task, we propose a novel *modality-specific split architecture*, which uses modality dependent adapter weights (see Figure 2 for an illustration of the architecture).

Our results clearly indicate that the proposed adapter-based architecture outperforms the recent state-of-the-art pretrained multilingual multimodal M^{3}P model (Ni et al., 2021) in zero-shot cross-lingual settings. However, the overall performance of zero-shot transfer remains low across the board, with an average drop of around 38 accuracy points across target languages. Using a small number of target language examples in a few-shot setup considerably improves performance for all approaches, but cross-lingual transfer performance still lags substantially behind source language performance. This demonstrates the inherent difficulty of the task, even though the corresponding questions are arguably simple, containing only 8.5 words on average (see Figure 1).

Contributions. 1) We propose the first evaluation benchmark for cross-lingual visual question answering, covering 7 diverse target languages; **2)** we propose novel adapter-based approaches for the creation of multilingual multimodal models; **3)** we systematically benchmark state-of-the-art and new multilingual multimodal models in zero-shot and few-shot learning setups, demonstrating the difficulty of the proposed task and serving as strong reference points for future work; **4)** we provide a thorough analysis of the different approaches, highlighting the aspects and question types that lead to the most common model failures, again motivating future work in this domain.

2 Background and Related Work

Multilingual Language Models. Pretrained multilingual transformer-based LMs such as mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020) adopt the same pretraining regime as their respective monolingual counterparts: BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019). They are pretrained via self-supervised masked language modelling objective (MLM) on concatenated text corpora of more than 100 languages, where text is tokenized using WordPiece, SentencePiece or BytePair encodings. These multilingual models have been shown to work surprisingly well for cross-lingual tasks, despite the fact that they do not rely on direct cross-lingual supervision (e.g., parallel data, translation dictionaries; Pires et al., 2019a; Wu and Dredze, 2019; Artetxe et al., 2020; Hu et al., 2020; K et al., 2020; Rust et al., 2021).

Vision and Language Models. Most transformer-based multimodal models (Lu et al., 2019; Tan and Bansal, 2019; Chen et al., 2020; Li et al., 2020a; Gan et al., 2020; Li et al., 2020b; Bugliarello et al., 2020; Ni et al., 2021, *inter alia*) jointly encode text tokens and image region features by preprocessing images using object detection models—such as Faster R-CNN (Ren et al., 2015)—to extract features for regions of interest (RoI) (Anderson et al., 2018). The image region features are passed through an affine layer, which learns to project the region features to the joint embedding space of the multimodal transformer. The bounding box coordinates of the RoI act as positional embeddings for the visual features. As such, they undergo an affine transformation to the embedding space and are combined with their respective image region representation. The position-aware image region embeddings get passed into the transformer. The multi-head attention then attends over all text and image inputs at every layer, learning a joint representation of both modalities. On the other hand, Kamath et al. (2021) avoid using object detectors as a black-box for pre-extracting these region features and instead make it a central part of the multimodal transformer architecture. Training the object detector end-to-end with the multimodal transformer adds flexibility and better representation capacity.

Similar to MLM, multimodal transformer-based models are trained with self-supervised objectives such as masked feature regression, masked object detection, masked attribute detection, and contrastive losses such as cross-modality matching (Tan and Bansal, 2019). Typically, image captioning datasets are used for pretraining such as COCO (Lin et al., 2014), Flickr30k (Plummer et al., 2015),

Conceptual Captions (CC) (Sharma et al., 2018), and SBU (Ordonez et al., 2011). Similar to unimodal language models, the [CLS] token is used as a contextual representation for classification tasks.

Multilingual multimodal models have also been proposed recently: M^{3}P (Ni et al., 2021) is trained on the Wikipedias of 50 different languages and the English multimodal CC dataset. In order to align tokens of languages other than English with image representations, M^{3}P utilizes a code-switching mechanism, where words of the English CC examples are randomly replaced with words from corresponding bilingual dictionaries. In UC^2, Zhou et al. (2021) augment English multimodal datasets with other languages via machine translation and propose masked region-to-token modeling and visual translation language modeling.[1]

Adapters (Rebuffi et al., 2017; Houlsby et al., 2019) have been introduced as a more efficient fine-tuning strategy for transfer learning in NLP and CV. Instead of fine-tuning all the weights of a pretrained model on the target task, small feed-forward layers are introduced at each layer of the pretrained model. During task fine-tuning, only the adapter weights are updated, while the pretrained parameters remain fixed/frozen. Adapters have been shown to work well for machine translation (Bapna and Firat, 2019; Philip et al., 2020) and cross-lingual transfer (Pfeiffer et al., 2020b, 2021; Üstün et al., 2020).

Datasets. Pretraining and fine-tuning data for multilingual multimodal models is typically based on (multimodal information from) Wikipedia (**WikiCaps**, **WIT**, Schamoni et al., 2018; Srinivasan et al., 2021), or on available downstream task data. **Multi30k** (Elliott et al., 2016) is a multilingual image captioning dataset for retrieval-type questions, covering English, German, French, and Czech; **GEM** (Su et al., 2021) covers image and video retrieval tasks across 20 and 30 different languages, respectively; **HowTo100M** (Huang et al., 2021) is a multilingual and multimodal pretraining dataset for image and video retrieval; **MultiSubs** (Wang et al., 2021) focuses on fill-in-the-blank tasks and lexical translation, covering English, Spanish, German, Portuguese, and French. In contemporary work Liu et al. (2021) propose **MaRVL**, a binary multilingual question answering dataset similar to NLVR2 (Suhr et al., 2019), spanning 5 typologically diverse languages (Chinese, Tamil, Swahili, Indonesian, and Turkish).

Previous datasets predominantly focus on (arguably simpler) retrieval-type tasks, only cover a small set of similar languages (e.g., Multi30k, MultiSubs), or only cover binary questions. In contrast, we propose the first multilingual visual question answering dataset, which covers a typologically more diverse set of languages.

3 xGQA

The original English GQA dataset (Hudson and Manning, 2019) was constructed by leveraging Visual Genome scene graphs (Krishna et al., 2017). An English question engine that utilizes *content* (i.e. information about objects, attributes, and relations provided) and *structure* (a linguistic grammar that couples hundreds of structural patterns and detailed lexical semantic resources) was used to generate over 22 million diverse questions, which are visually grounded in the image scene graphs.

Each question is associated with additional metadata such as **structural types**: (1) *verify* for yes/no questions (e.g. "Do you see any cats?"), (2) *query* for all open questions (e.g. "Who is wearing jeans?"), (3) *choose* for questions that present two alternatives to choose from (e.g. "Is it red or blue?"), (4) *logical* which involve logical inference (e.g. "Is the field soft and snowy"), and (5) *compare* for comparison questions between two or more objects (e.g. "Are all the animals zebras?"). For further details regarding the metadata, we refer the reader to Hudson and Manning (2019).

Dataset Design. The principal objective when devising xGQA was to create a genuinely typologically diverse multimodal and multilingual evaluation benchmark for visual question answering. We utilize the balanced[2] test-dev set of GQA, which consists of 12,578 questions about 398 images.[3] Due to the defined structural patterns, the formulation of the questions is simple, with an average length of 8.5 words.[4] The resulting xGQA dataset

[1] The model weights of UC^2 were not released by the time of experimentation.

[2] To reduce biases in the conditional answer distribution Hudson and Manning (2019) utilize the structural metadata to downsample and create balanced datasets that are more robust against shortcuts and guesses.

[3] We chose to translate the test-dev set of GQA, as the labels for test-std are not released.

[4] For this reason, we chose to hire university students that are currently conducting their (Computer Science or Computational Linguistics) studies in English and are all fluent English speakers to translate the question into their native language. They were paid above the minimum hourly wage of the country of their respective university.

Language	iso	Family	Script	Speakers
English	en	IE:Germanic	Latin	400M
German	de	IE:Germanic	Latin	95M
Portuguese	pt	IE:Romance	Latin	250M
Russian	ru	IE:Slavic	Cyrillic	150M
Indonesian	id	Austronesian	Latin	43M
Bengali	bn	IE:Iranian	Bengali	230M
Korean	ko	Koreanic	Korean	77M
Chinese	zh	Sino-Tibetan	Chinese	1.2B

Table 1: Languages covered by xGQA. IE stands for Indo-European.

Set	Test	Dev	Train					
#Img	300	50	1	5	10	20	25	48
#Ques	9666	1422	27	155	317	594	704	1490

Table 2: Few-shot dataset sizes. The GQA test-dev set is split into new development, test sets, and training splits of different sizes. We maintain the distribution of structural types in each split.

covers translations in 7 languages, each representing a distinct language family, and contains examples written in 5 different scripts (see Table 1).

Few-Shot Data Splits. In order to conduct cross-lingual few-shot learning experiments, we provide new data splits of different sizes. We split on images and add all questions associated with the image to the respective set. The development and test sets consist of 50 and 300 images, respectively. The training splits consist of 1, 5, 10, 20, 25, and 48 images, see Table 2. We ensure that the distribution of structural types within each set is maintained.

xGQA is the first truly typologically diverse multilingual multimodal benchmark, unlocking new experimentation and analysis opportunities in cross-lingual zero-shot and few-shot scenarios. While the questions in xGQA are intuitive and easy for humans to solve, we later show that current state-of-the-art models still have difficulty with transfer.

4 Baselines

To analyze the performance and current gaps on xGQA, we first evaluate the recently proposed M^{3}P model, which has been pretrained on multilingual and multimodal data. However, pretraining is computationally expensive and only limited amounts of multilingual multimodal resources are available. Therefore, we further propose new and more efficient approaches that (1) extend state-of-the-art multilingual language models to the multimodal domain and (2) provide multilingual capabilities to state-of-the-art multimodal models.

Unless noted otherwise, we follow the predominant fine-tuning strategy for GQA; a prediction head is placed on top of the output of a pretrained transformer. All possible 1853 answers of the GQA task are mapped to a class label. The question associated with an image together with the position-aware region features are passed as input to the transformer, supervised using a cross-entropy loss.[5]

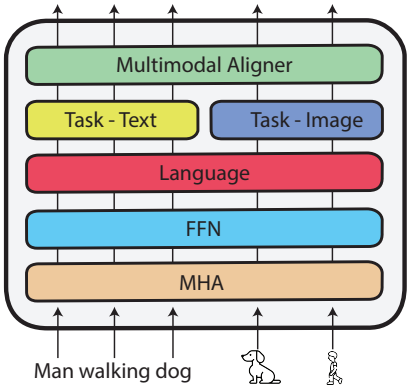


Figure 2: Architecture of an adapter-based multilingual multimodal model. Text and image inputs share the weights of the multi-head attention (MHA) and feed-forward (FFN) layers, as well as the *language* and *multimodal align* adapters. Each modality is passed through a modality specific *task* adapter, the outputs of which are concatenated.

4.1 Multimodal → Multilingual

OSCAR+^{Emb}. To extend a monolingual transformer LM to a multilingual domain, Artetxe et al. (2020) fine-tune a new word-embedding layer in the target language. Inspired by this idea, we now describe how we extend the current state-of-the-art monolingual multimodal transformer model OSCAR+ (Zhang et al., 2021) to learn new embeddings for the target languages.

In the *language-extension* phase, we replace the embedding matrix of OSCAR+ with a randomly initialized embedding matrix.[6] The transformer weights are frozen while only the newly introduced embeddings are fine-tuned on unlabeled text data of the target language with the MLM objective.

In the *target-task* phase, the original OSCAR+ model is fine-tuned on the English training data of GQA, where the transformer layers are fine-tuned, but the embedding layer is frozen. During inference, the embedding layer is replaced with the target language's embedding layer.

[5]For instance, we use this strategy to fine-tune all parameters of M^{3}P on the GQA training data.

[6]Following Pfeiffer et al. (2021), we copy the embeddings of lexically overlapping tokens (if such tokens exist) from the original embedding space to the new embedding space, as it typically works better than fully random initialization.

OSCAR+^{Ada}. We extend this by adding adapters.

In the *language-extension* phase we follow Pfeiffer et al. (2021) in order to extend the model to the target languages. Similar to OSCAR+^{Emb}, we train a new embedding layer. We further add *language* adapters at every transformer layer. Given that OSCAR+ is trained on English text, we follow Pfeiffer et al. (2020b) when training English *language* adapter modules, without replacing the embedding matrix. The transformer weights are frozen while only the *newly* introduced embeddings and *language* adapter weights are fine-tuned on unlabeled text data of the language.

For the *target-task* phase, we propose a novel modality-split architecture (see Figure 2) inspired by the cross-lingual transfer method of Pfeiffer et al. (2020b). At each transformer layer, text and image representations are passed through the pretrained multi-head attention (MHA) and feed-forward (FFN) layers. Both image and text representations are also passed through the pre-trained *language* adapters. Each modality is then passed through modality-specific *text* and *image task* adapters and next through a shared *multimodal alignment* adapter.[7] We follow Pfeiffer et al. (2020b), freezing transformer, embedding and *language* adapter weights during training, thus fine-tuning only the *task* and *multimodal aligner* adapter weights, together with the prediction head. At inference time, the embedding layer and the *language* adapters are replaced with the target language weights.

4.2 Multilingual → Multimodal

mBERT^{Ada}. For experiments where we extend a multilingual model to become multimodal, we utilize mBERT (Devlin et al., 2019).

Given that mBERT is able to represent many different languages, it is not necessary to learn new embedding layers for the target languages in the *language-extension* phase. Instead, we utilize the mBERT-compatible *language* adapters available on AdapterHub.ml (Pfeiffer et al., 2020a).[8]

For the *target-task* phase, we follow OSCAR+ for the image representation layer, where image features are combined with their respective positional information and passed through an affine transformation layer. We experiment with the same adapter architecture from Figure 2, as described for OSCAR+^{Ada}. We again freeze transformer, embedding and *language* adapter weights during training. However, in contrast to OSCAR+*, we randomly initialize and fine-tune the affine image transformation layer. We also fine-tune the *task*, *multimodal aligner* adapter weights, and prediction head, all on the GQA task. At inference time, the embedding layer and the *language* adapters are replaced with the corresponding target language weights.

5 Experimental Setup

5.1 Language-Extension Phase

For OSCAR+^{Emb} and OSCAR+^{Ada}, we follow the general setups proposed by Pfeiffer et al. (2020b, 2021). We train a new word-piece tokenizer for each target language with a vocabulary size of 30k. We fine-tune the randomly initialized embedding layer, and (for OSCAR+^{Ada}) adapter layers for 100k update steps with a batch size of 64 and a learning rate of $1e{-}4$. For mBERT^{Ada}, we utilize the language adapters from AdapterHub.ml.

5.2 Fine-tuning on GQA

We follow the standard setup proposed by Li et al. (2020b), passing the representation of the [CLS] token through a prediction head. We fine-tune the respective models using a cross-entropy loss with labels being all possible answers in the GQA dataset. Following prior work (Li et al., 2020b), we use a batch size of 192 and train for 5 epochs on the unbalanced GQA training portion.

M^{3}P. We fine-tune all weights of the pretrained model with a learning rate of $3e{-}5$.

OSCAR+^{Emb}, OSCAR+^{Ada}, and mBERT^{Ada}. We use the pretrained weights and image region features provided by Zhang et al. (2021). However, we do not pass the object attribute labels as inputs to the model. The object attribute labels are in English and utilizing them in cross-lingual scenarios is non-trivial.[9] We leave this for future work.

For the OSCAR+^{Emb} setting, we fine-tune the transformer weights and the prediction head and freeze the embedding layer, using a learning rate

[7] We have compared multiple different architectures as illustrated in Figure 6 in the Appendix, finding this setup to perform best. We present results of the alternative architectures also in the Appendix.

[8] While all xGQA languages already have readily available language adapters on AdapterHub, any hypothetical extension of experiments to languages without such adapters would involve training their dedicated language adapters, e.g., following the procedure of Pfeiffer et al. (2020b).

[9] The replaced tokenizer and embedding representations of the target language potentially do not adequately represent English terms, resulting in a misalignment between the question (in the target language) and the object attributes (in English).

model	en	de	pt	ru	id	bn	ko	zh	mean
M3P	58.43 ±1.4	23.93 ±3.2	24.37 ±4.0	20.37 ±3.4	**22.57 ±6.1**	**15.83 ±3.6**	16.90 ±3.8	18.60 ±1.0	20.37
OSCAR+$^{\text{Emb}}$	**62.23 ±0.3**	17.35 ±1.0	19.25 ±0.4	10.52 ±4.0	18.26 ±0.4	14.93 ±2.0	17.10 ±1.8	16.41 ±3.2	16.26
OSCAR+$^{\text{Ada}}$	60.30 ±0.4	18.91 ±0.8	27.02 ±2.3	17.50 ±1.2	18.77 ±0.3	15.42 ±2.0	15.28 ±2.7	14.96 ±2.1	18.27
mBERT$^{\text{Ada}}$	56.25 ±0.5	**29.76 ±2.3**	**30.37 ±1.8**	**24.42 ±1.1**	19.15 ±2.8	15.12 ±1.9	**19.09 ±0.9**	**24.86 ±1.8**	**23.25**

Table 3: Zero-shot transfer results when transferring from English GQA. Average accuracy and standard deviation are reported. Best results are highlighted in **bold**; *mean* scores are not averaged over the source language (English).

of $3e{-}5$. For the OSCAR+^{Ada} and mBERT^{Ada} settings, we add adapter layers as described in §4.1 and illustrated in Figure 2. We freeze all pretrained weights–including embeddings, transformer layers, and language adapters–and only fine-tune the newly introduced adapters and the prediction head. For mBERT^{Ada}, we also add and train the affine image transformation layer. We fine-tune the adapter-based models with a learning rate of $1e{-}4$.

5.3 Zero-Shot Cross-Lingual Transfer

For zero-shot cross-lingual evaluation, we utilize the model fine-tuned on the GQA training data and evaluate on the multilingual xGQA test data. The model checkpoint that performed best on the English GQA validation data is selected for transfer.

M^{3}P. As the model is pre-trained to cover a large variety of languages, no additional steps are required for cross-lingual transfer.

OSCAR+^{Emb}. We replace the English embedding layer with the target-language embedding layer.

OSCAR+^{Ada}. We replace the English embedding and language adapter layers with the embedding and adapters layers of the target language.

mBERT^{Ada}. We replace the language adapter layers with the adapters layers of the target language.

5.4 Few-Shot Cross-Lingual Transfer

For few-shot cross-lingual scenarios we follow Lauscher et al. (2020) and start from the same fine-tuned model as for zero-shot transfer (see §5.3). We then fine-tune the same parts of the model as when training on the English training data as in §5.2, but on the small portions of multimodal data available in the target language. We train on the different data splits, consisting of 1, 5, 10, 15, 20, 25, and 48 images (see Table 2). We experiment with training for a different number of epochs (5, 10) using different learning rates ($1e{-}5$ and $5e{-}5$ for M^{3}P and OSCAR+^{Emb}, and $5e{-}5$ and $1e{-}4$ for OSCAR+^{Ada} and mBERT^{Ada}). We find that training for longer and with a larger learning rate performed best for all settings.

6 Results and Discussion

The main results are presented in Table 3 (zero-shot experiments) and in Table 4 (few-shot).

6.1 Zero-Shot Cross-Lingual Transfer

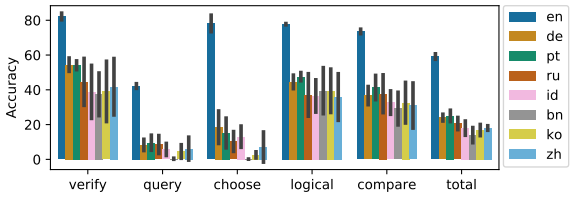
One of our core findings is that multimodal zero-shot cross-lingual transfer is extremely difficult; we witness an average drop in accuracy of more than 38 points on the target languages of the xGQA dataset compared to English GQA scores (e.g., compare the results with M^{3}P).

While, as expected, OSCAR+ achieves the best accuracy on the English test set, the massively multilingual models—M^{3}P and mBERT—perform considerably better in cross-lingual transfer.[10] This indicates, that joint multilingual pretraining is important and a simple multilingual adapter-based or embedding-based extension of monolingual models achieves inferior cross-lingual performance.

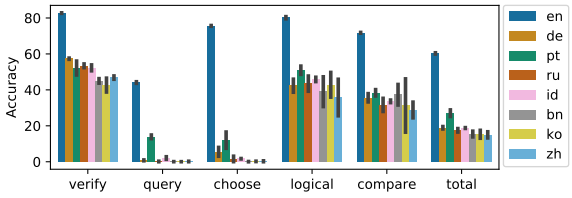
While the pretraining method M^{3}P achieves better accuracy on the English test set, the adapter-based multimodal extension of mBERT outperforms M^{3}P in cross-lingual transfer. We hypothesize that, when fine-tuning all transformer weights on monolingual multimodal data, the cross-lingual alignment breaks within M^{3}P. However, this does not happen in adapter-based settings, as the multilingual weights are frozen and thus remain intact.

Analysis of Structural Question Types. Figure 3 depicts our analysis of the structural question types in zero-shot experiments. We observe large drops

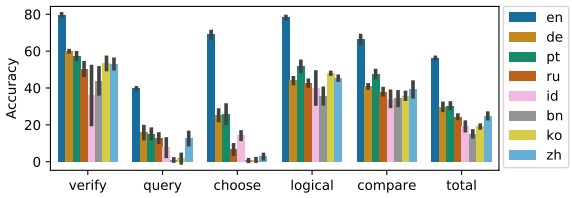
[10] The superior accuracy of OSCAR+ on the English test set is expected as the model was pretrained on large English multimodal data. We find that fine-tuning all transformer weights (OSCAR+^{Emb}) achieves slightly better results than only training adapter weights (OSCAR+^{Ada}). Our slightly lower scores compared to results by Zhang et al. (2021) can be explained by us (1) not fine-tuning the embedding layer, and (2) not utilizing the attribute labels. Further, previous works that focus only on English add the official *validation* set to the *training* set, use the official *test-dev* set as their development set, and report their test scores of the official GQA test benchmark *test-std* for which labels are not available. Our scores follow the training splits, where we use the official *test-dev* set as the final test-set we report our results on, as described in dataset construction.



(a) M3P

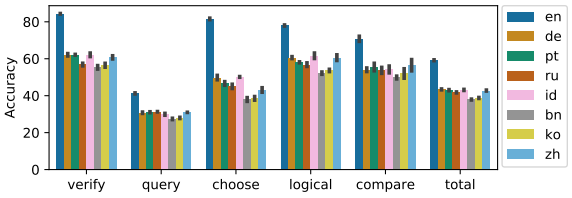


(b) OSCAR+$^{\text{Ada}}$

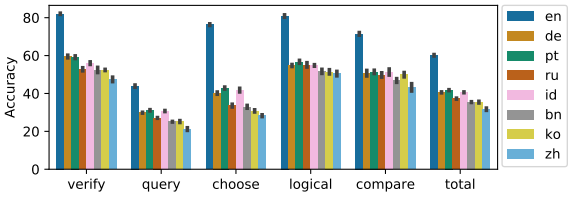


(c) mBERT$^{\text{Ada}}$

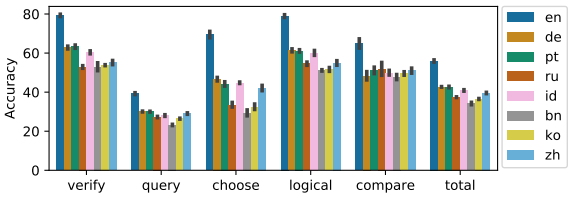
Figure 3: Zero-shot accuracy across different languages and structural question types from xGQA.



(a) M3P



(b) OSCAR+$^{\text{Ada}}$



(c) mBERT$^{\text{Ada}}$

Figure 4: Few-shot accuracy (with 48 images, see Table 2) across different languages and structural question types from xGQA.

in accuracy especially for *query* and *choose* type questions. *Query* type questions are free-form and thus semantically the most difficult to answer, even in the source language (English). This explains the overall low accuracy across all approaches in zero-shot settings for this question type.

This is in stark contrast with the *choose*-type questions, which the models perform very well on in the source language. However, we report a substantial accuracy drop in zero-shot cross-lingual transfer. This decrease is most likely due to the nature of the question formulation and the modelling implementation. *Choose*-type questions are formulated such that the answer to the question is a word or phrase which appears in the question, i.e. "Is it red or blue?". The label classes, and consequently the prediction head, are constructed as a set of all answers appearing in the dataset. This means that the model learns a distributed representation of each answer in its final layer. Consequently, in cross-lingual transfer, the model is required to automatically align the question's options "red" or "blue" (translated in their respective language), with their English latent representation of the model's prediction head. The very low results in this category indicate that this cross-lingual word alignment breaks in zero-shot scenarios.

Overall, zero-shot transfer with our proposed multimodal adapter-based extension of mBERT (mBERT^{Ada}) achieves the best accuracy, with almost 3 points increase over M^{3}P and almost 5 points increase over OSCAR+. However, the overall accuracy of all approaches remains low in comparison to the results in English. This indicates that zero-shot multimodal cross-lingual transfer is extremely difficult, most likely due to the misalignment issue between visual and cross-lingual internal representations. To investigate this conjecture further, we run similar tests in few-shot setups, which should potentially mitigate the misalignment issue observed in zero-shot setups.

6.2 Few-Shot Cross-Lingual Transfer

The main results of few-shot experiments are provided in Table 4, while the plot illustrating the impact of different amounts of training data is shown in Figure 5. One crucial finding is that as expected, utilizing an increasing amount of data instances in the target language consistently improves accuracy for all methods. This culminates in an improvement of up to 20 accuracy points when specializing the model with only 48 images in the target language. This indicates that a small number of target-language examples supports the models in

Lang	Model	# Training Images						
		0	1	5	10	20	25	48
de	M3P	24.78	31.49	**39.31**	**41.05**	**42.22**	**42.54**	**43.16**
	OSCAR+^{Emb}	17.49	17.84	29.09	34.48	37.35	38.45	41.08
	OSCAR+^{Ada}	17.84	21.40	31.26	35.84	37.92	38.46	40.58
	mBERT^{Ada}	**32.41**	**33.87**	37.44	39.15	40.65	41.63	42.71
pt	M3P	26.73	32.98	37.23	**39.07**	**40.92**	**41.05**	43.06
	OSCAR+^{Emb}	19.36	22.55	32.42	36.37	39.01	40.15	**43.27**
	OSCAR+^{Ada}	24.58	29.61	34.73	37.46	38.82	39.70	41.75
	mBERT^{Ada}	**31.45**	**33.27**	**37.31**	38.88	40.51	41.03	42.62
ru	M3P	24.29	**32.32**	**36.71**	**38.53**	**39.94**	**40.13**	**41.85**
	OSCAR+^{Emb}	7.98	17.32	23.72	28.21	32.15	32.87	36.84
	OSCAR+^{Ada}	16.38	19.74	27.42	30.17	33.22	34.21	37.28
	mBERT^{Ada}	**25.51**	26.47	31.69	32.47	34.93	35.53	37.42
id	M3P	18.74	31.37	**37.24**	**38.65**	**41.07**	**42.00**	**43.12**
	OSCAR+^{Emb}	17.89	21.09	29.76	33.59	36.69	37.31	40.51
	OSCAR+^{Ada}	18.52	23.94	31.45	34.60	37.26	37.97	40.60
	mBERT^{Ada}	**19.77**	**31.99**	34.49	36.26	39.15	39.81	40.88
bn	M3P	**17.59**	17.33	**26.94**	**31.09**	**34.58**	**35.27**	**37.96**
	OSCAR+^{Emb}	13.35	**17.40**	21.67	26.61	31.94	32.78	36.97
	OSCAR+^{Ada}	13.96	15.60	22.35	27.20	31.25	31.81	35.45
	mBERT^{Ada}	13.38	11.33	23.10	26.55	31.60	32.26	34.18
ko	M3P	19.70	**22.94**	**32.28**	**35.50**	**37.72**	**37.84**	**38.61**
	OSCAR+^{Emb}	15.11	16.43	19.99	24.78	29.48	30.43	35.59
	OSCAR+^{Ada}	12.25	15.48	20.73	25.97	31.37	32.20	35.41
	mBERT^{Ada}	**19.92**	17.71	27.83	31.27	34.44	35.03	36.51
zh	M3P	19.66	**27.76**	**36.15**	**38.21**	**40.48**	**40.53**	**42.55**
	OSCAR+^{Emb}	12.66	14.77	19.17	22.13	27.97	29.08	33.24
	OSCAR+^{Ada}	13.20	15.12	19.67	22.74	26.81	28.19	31.69
	mBERT^{Ada}	**26.16**	23.47	32.93	35.82	38.22	37.89	39.57

Table 4: Average accuracy of few-shot results, utilizing different amounts of training data. *0* presents the best zero-shot results. These models are used as initialization for the subsequent few-shot experiments. **Bold** numbers indicate the best scores.

partially repairing its internal cross-lingual multimodal alignment. Interestingly, we find that with as little as 5 images, and their corresponding questions, M^{3}P begins to outperform mBERT^{Ada}—the best performing zero-shot model.

We again analyze the impact of few-shot learning on the accuracy across different structural question types, with the results depicted in Figure 4. The overall accuracy increases across all types compared to zero-shot scenarios (cf., Figure 3). However, the most pronounced gains are reported for *query* and *chose*-type questions, on which the model performed the worst in zero-shot setups. This implies the improved alignment between latent multimodal and multilingual representations, achieved via fine-tuning the model on a small amount of examples in the target language.

6.3 Language Transfer

We witness cross-lingual transfer capability patterns similar to those shown by previous work, where our models perform best on typologically close languages (Pires et al., 2019b; Lauscher et al., 2020). Our models transfer best to German (de) and Portuguese (pt), both being part of the Indo-European (IE) language family and also sharing

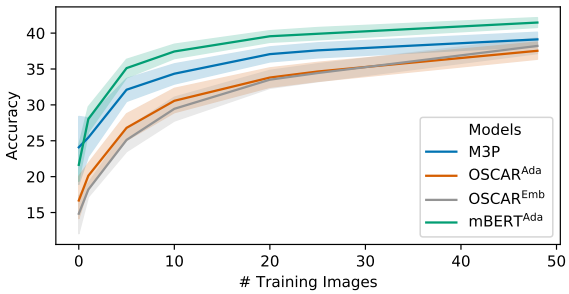


Figure 5: Few-shot accuracy with different training dataset sizes of the different approaches. Scores are averaged over all languages.

the same script (Latin) with the source language English (en). We see a small drop in accuracy for Russian (ru), Indonesian (id), and Chinese (zh) and a larger drop in accuracy for Bengali (bn) and Korean (ko). All of these languages are typologically different to the source language and in most cases do not share the same script. These differences highlight the importance of language diversity in cross-lingual transfer. Our benchmark thus enables experimentation and evaluation of multilingual multimodal models on a representative set of truly typologically diverse languages.

7 Conclusion

We have proposed xGQA, a first cross-lingual evaluation benchmark for the visual question answering task. xGQA extends the English GQA by 7 typologically diverse languages, covering 5 different scripts. As additional baselines, we have further proposed new adapter-based methods to extend unimodal multilingual models to become multimodal and—vice-versa—monolingual multimodal models to become multilingual. Our results have indicated that 1) efficient adapter-based methods slightly outperform the pretrained multilingual multimodal model M^{3}P in zero-shot scenarios, but 2) the overall zero-shot cross-lingual transfer yields harsh accuracy drops compared to the English performance for all models in comparison. Further, accuracy can be partially recovered via few-shot learning, where small amounts of training data are available in the target language. However, the large gaps remain, suggesting the inherent complexity of the cross-lingual task despite it being extremely intuitive and easy to solve by (bilingual) humans.

We hope that our dataset and error analysis will motivate future work on this task and, more broadly, in the exciting emerging domain of multilingual multimodal representation learning.

References

P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6077–6086.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Ankur Bapna and Orhan Firat. 2019. Simple, scalable adaptation for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1538–1548, Hong Kong, China. Association for Computational Linguistics.

Emanuele Bugliarello, Ryan Cotterell, Naoaki Okazaki, and Desmond Elliott. 2020. Multimodal pretraining unmasked: Unifying the vision and language berts. *arXiv preprint*, abs/2011.15124.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, volume 12346 of *Lecture Notes in Computer Science*, pages 213–229. Springer.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: universal image-text representation learning. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX*, volume 12375 of *Lecture Notes in Computer Science*, pages 104–120. Springer.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

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

Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. 2020. Large-scale adversarial training for vision-and-language representation learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, pages 4411–4421, Virtual. PMLR.

Po-Yao Huang, Mandela Patrick, Junjie Hu, Graham Neubig, Florian Metze, and Alexander Hauptmann. 2021. Multilingual multimodal pre-training for zero-shot cross-lingual transfer of vision-language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2443–2459, Online. Association for Computational Linguistics.

Drew A. Hudson and Christopher D. Manning. 2019. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6700–6709. Computer Vision Foundation / IEEE.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual BERT: an empirical study. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*, Addis Ababa, Ethiopia. OpenReview.net.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Ishan Misra, Gabriel Synnaeve, and Nicolas Carion. 2021. MDETR - modulated detection for end-to-end multimodal understanding. In *2021 IEEE International Conference on Computer Vision, ICCV 2021, Online, October 10-17, 2021*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.*, 123(1):32–73.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. 2020a. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 11336–11344. AAAI Press.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020b. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX*, volume 12375 of *Lecture Notes in Computer Science*, pages 121–137. Springer.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer.

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, EMNLP 2021, Online, November , 2021*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *arXiv preprint*, abs/1907.11692.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13–23.

Minheng Ni, Haoyang Huang, Lin Su, Edward Cui, Taroon Bharti, Lijuan Wang, Dongdong Zhang, and Nan Duan. 2021. M3P: learning universal representations via multitask multilingual multimodal pretraining. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 3977–3986. Computer Vision Foundation / IEEE.

Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. 2011. Im2text: Describing images using 1 million captioned photographs. In *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*, pages 1143–1151.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020a. AdapterHub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54, Online. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020b. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2021. UNKs Everywhere: Adapting Multilingual Language Models to New Scripts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Online, November , 2021*.

Jerin Philip, Alexandre Berard, Matthias Gallé, and Laurent Besacier. 2020. Monolingual adapters for zero-shot neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages

4465–4470, Online. Association for Computational Linguistics.

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

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

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2641–2649.

Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. 2017. Learning multiple visual domains with residual adapters. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 506–516.

Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 91–99.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics, ACL 2021, Online, August 1-6, 2021*. Association for Computational Linguistics.

Shigehiko Schamoni, Julian Hitschler, and Stefan Riezler. 2018. A dataset and reranking method for multimodal MT of user-generated image captions. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas, AMTA 2018, Boston, MA, USA, March 17-21, 2018 - Volume 1: Research Papers*, pages 140–153. Association for Machine Translation in the Americas.

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

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. WIT: wikipedia-based image text dataset for multimodal multilingual machine learning. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 2443–2449. ACM.

Lin Su, Nan Duan, Edward Cui, Lei Ji, Chenfei Wu, Huaishao Luo, Yongfei Liu, Ming Zhong, Taroon Bharti, and Arun Sacheti. 2021. GEM: A general evaluation benchmark for multimodal tasks. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, pages 2594–2603. Association for Computational Linguistics.

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

Hao Tan and Mohit Bansal. 2019. LXMERT: learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 5099–5110. Association for Computational Linguistics.

Ahmet Üstün, Arianna Bisazza, Gosse Bouma, and Gertjan van Noord. 2020. UDapter: Language adaptation for truly Universal Dependency parsing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2302–2315, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Josiah Wang, Pranava Madhyastha, Josiel Figueiredo, Chiraag Lala, and Lucia Specia. 2021. Multisubs: A large-scale multimodal and multilingual dataset. *arXiv preprint*.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. VinVL: Making Visual Representations Matter in Vision-Language Models. *arXiv preprint*.

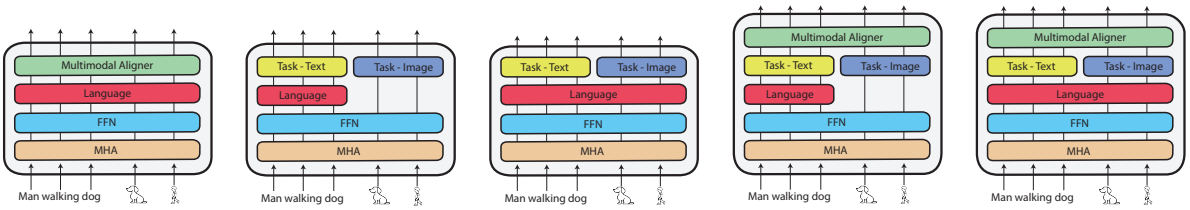
Mingyang Zhou, Luowei Zhou, Shuohang Wang, Yu Cheng, Linjie Li, Zhou Yu, and Jingjing Liu. 2021. UC2: universal cross-lingual cross-modal vision-and-language pre-training. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 4155–4165. Computer Vision Foundation / IEEE.

A Appendix

We experiment with different multimodal adapter architectures as illustrated in Figure 6. In initial experiments we find that splitting the modalities (settings 2-5) outperforms a joint adapter (setting 1). However, a joint "alignment" architectures (settings 4-5) outperform settings where we only use modality-specific adapters (settings 2-3). We more thoroughly investigate settings 4-5 and report scores in Table 5. Interestingly we find that when only using the language adapter for the textual inputs, cross-lingual accuracy drops for both OSCAR+ and mBERT; The difference is more pronounced for OSCAR+. We speculate that this is due to a latent misalignment of the representation spaces, partly due to the residual connection. Due to the better performance of setting 5, we have reported scores of this architecture in the main paper (as illustrated in Figure 2).

model	Setting	en	de	pt	ru	id	bn	ko	zh	mean
OSCAR+^{Ada}	4	60.21	18.60	25.48	8.22	17.79	10.47	9.97	12.54	14.72
OSCAR+^{Ada}	5	**60.30**	**18.91**	**27.02**	**17.50**	**18.77**	**15.42**	**15.28**	**14.96**	**18.27**
mBERT^{Ada}	4	**57.83**	27.86	28.88	22.87	**20.86**	14.74	18.30	24.39	22.56
mBERT^{Ada}	5	56.25	**29.76**	**30.37**	**24.42**	19.15	**15.12**	**19.09**	**24.86**	**23.25**

Table 5: Zero-shot transfer results on xGQA for the different adapter architecture settings (as illustrated in Figure 6) when transferring from English GQA. Average accuracy is reported. Best results for each language and model type are highlighted in **bold**; *mean* scores are not averaged over the source language (English).



(a) Setting 1 (b) Setting 2 (c) Setting 3 (d) Setting 4 (e) Setting 5

Figure 6: The different multimodal multilingual adapter architectures we experimented with. The best performing architecture was setting 5, which we present results for in the main paper.