Adapting Grounded Visual Question Answering Models to Low Resource Languages

Ying Wang, Jonas Pfeiffer, Nicolas Carion, Yann LeCun, Aishwarya Kamath
Center for Data Science, New York University
{yw3076, aish}@nyu.edu

Abstract

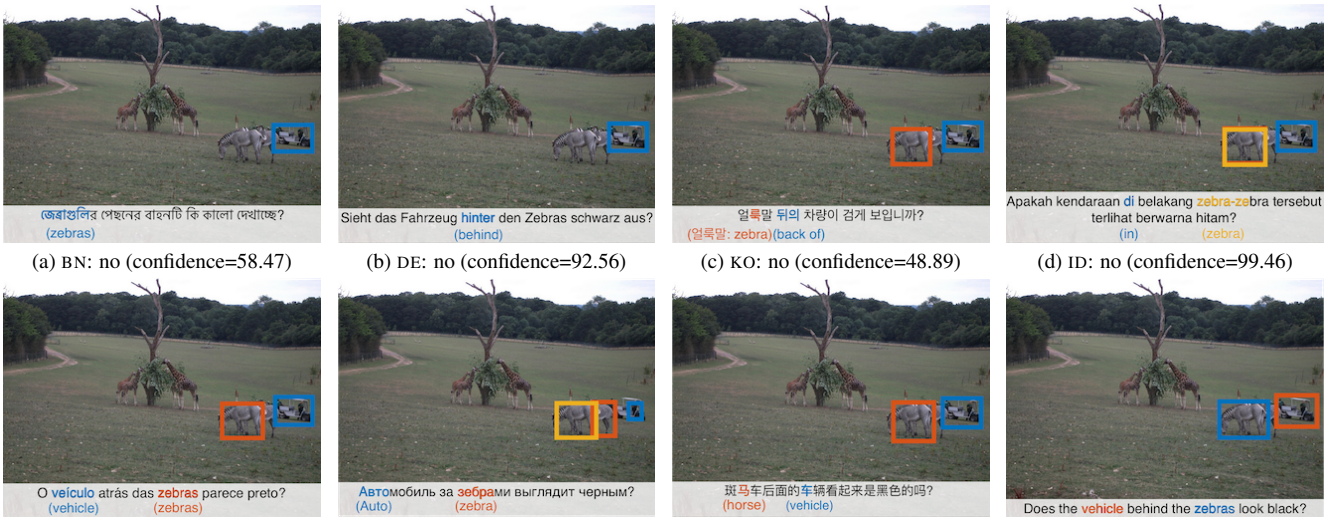
While huge progress has been made on a variety of vision and language tasks in recent years, most major advances have been restricted to the English language due to the scarcity of relevant training and evaluation datasets in other languages. A popular approach to address this gap, has been to utilize machine-translated multi-modal datasets or multi-lingual text-only datasets for pre-training. This approach not only fails to exploit existing pre-trained state-of-the-art English multi-modal models, but also is not a viable solution for low-resource languages where translation quality is not as reliable. Therefore, we propose xMDETR*, a multi-lingual grounded vision-language model based on the state-of-the-art model MDETR, by adapting it to new languages without machine-translated data, while also keeping most of the pre-trained weights frozen.* xMDETR *leverages mono-lingual pre-trained MDETR to achieve results competitive to state of the art on xGQA, a standard multilingual VQA benchmark. It is also interpretable, providing bounding boxes for key phrases in the multi-lingual questions. Our method utilizes several architectural as well as data-driven techniques such as training a new embedding space with a Masked Language Modeling (MLM) objective, code-switching, and adapters for efficient and modular training. We also explore contrastive losses to enforce the bridging of multi-modal and multi-lingual representations on multi-lingual multi-modal data, when available. We evaluate* xMDETR *on xGQA in both zero-shot and few-shot settings, improving results on Portuguese, Indonesian and Bengali, while remaining competitive on other languages.*

1. Introduction

Over the past few years, large-scaled pre-trained Vision-Language Models (VLMs) have emerged as a promising approach for tackling the challenging task of visual question answering (VQA). The dominant approach is to use transformer-based models pre-trained on large-scale web-scraped image-text datasets, capable of capturing complex semantic relationships between visual and textual modalities. Performance on the VQA benchmarks is evaluated after fine-tuning on large-scale visual question-answering datasets [1,9,11,30,32]. GQA [11] is a popular visual question answering (VQA) dataset consisting of 22M English questions about real images from the Visual Genome [13] dataset and has powered the development of state-of-the-art question answering models [12,33,38]. However, this kind of large-scale VQA dataset is missing in other languages, hindering the advancement of multi-lingual VQA models. Previous attempts at building multi-lingual VQA systems often utilize machine-translated datasets and/or a combination of textual datasets of multiple languages to train a transformer-based multi-modal model [20,28,37,39].

While these methods have shown promising results, they still face significant challenges. One major issue is the quality of machine translation, which is still heavily reliant on the availability and quality of resources in the target language. As a result, this approach may still be disadvantageous for low-resource languages where such resources may be scarce or non-existent. Further, training the model from scratch on the limited datasets available in the target language does not fully exploit existing large-scale pre-trained English VQA models and training from scratch can also be computationally expensive.

To quantify the gap in performance of multi-lingual VQA systems, an evaluation benchmark called xGQA [22], based on the GQA dataset [11] was recently proposed. It extends the original English GQA dataset to 7 typologically diverse languages and provides few-shot splits to test the zero-shot and few-shot cross-lingual transfer ability of multi-lingual VQA models. Crucially, the amount of multi-lingual *training* data is very limited, only provided to test models in few shot settings. In this paper, we propose an efficient method to transfer a state-of-the art grounded mono-lingual vision-language model, MDETR [12], to new languages that are included in xGQA, while holding most of the parameters frozen. We test four strategies to minimize the gap between new languages and English, includ-



(a) BN: no (confidence=58.47) (b) DE: no (confidence=92.56) (c) KO: no (confidence=48.89) (d) ID: no (confidence=99.46)

(e) PT: no (confidence=70.35) (f) RU: no (confidence=68.12) (g) ZH: no (confidence=93.34) (h) EN: no (confidence=99.92)

Figure 1. An example taken from the test split of xGQA. The input query is "does the vehicle behind the zebras look black?" and the ground-truth answer is "no". xMDETR predicts bounding boxes with predicted spans in the input query and thus is more interpretable than methods that are only capable of answering prediction. Our method not only generates the correct answer but also successfully detects the main objects mentioned in the input for nearly all languages (except for the wrong span associated with the bounding box for Bengali).

ing training a new embedding space with Masked Language Modeling (MLM) objective, code-switched training which augments the English GQA with code-switched text in target languages and inserting adapters to the text encoder to allow for more flexibility without updating all weights. Finally, we employ contrastive training to further boost the performance on some relatively high-resource languages for which we have image-text datasets available.

The effectiveness of our proposed methods is evaluated in both zero-shot and few-shot settings and compared with other models on the xGQA benchmark. Our results are competitive with state of the art approaches, without using translated data. Since we build upon MDETR, a grounded VQA model, xMDETR is able to produce bounding boxes along with the alignment to relevant object words in the question in target languages. These bounding boxes provide some insights into the reasoning process of the model, making it more interpretable compared to other existing multilingual VQA systems.

Our contributions are as follows:

- We propose a novel method to leverage a state-of-the-art grounded VQA model, and adapt it to low-resource languages without relying on translated data.
- We report results on xGQA in both zero-shot and few-shot settings, while providing interpretable predictions. On Bengali, Indonesian and Portuguese we outperform existing approaches and are competitive on other languages.

2. Related Work

The major challenge in cross-lingual VQA is the scarcity of resources in target languages. The GQA dataset [11] for English VQA has 113K images and 22M questions, powering the training of multi-modal models in this field. In contrast, many other languages currently lack similar high-quality image-text datasets at a comparable scale. In this section, we briefly review previous works proposing datasets and approaches to evaluate and mitigate this gap.

xGQA [22] is a multi-lingual evaluation benchmark for the visual question answering (VQA) task, extending the test-dev set of the English GQA dataset [11] to 7 typologically diverse languages (German DE, Portuguese PT, Russian RU, Indonesian ID, Bengali BN, Korean KO, and Chinese ZH) by manually translating English questions. It provides a few-shot training split of 1, 5, 10, 20, 25 images, a development set of 50 images and a test set of 300 images. It is then integrated into a multi-lingual multi-modal benchmark called IGLUE [3] that covers more vision-language tasks. The release of this benchmark encourages research in this field and provides a valuable metric when comparing different multi-lingual VQA models.

Multilingual datasets Since xGQA only provides limited data for evaluation purposes, additional multilingual datasets are necessary for cross-lingual transfer. Large multilingual text datasets are usually created by web crawl, which allows us to leverage the vast amount of information available on the internet. OSCAR [21], or Open Super-large Crawled ALMAnaCH coRpus, is one successful example of this approach. It covers over 200 languages and is obtained

Language	ISO	OSCAR size	# speakers
Bengali	BN	5.8G	230M
Korean	KO	12G	77M
Indonesian	ID	16G	43M
Portuguese	PT	64G	250M
German	DE	145G	95M
Chinese	ZH	249G	1.2B
Russian	RU	568G	150M

Table 1. Languages in xGQA sorted ascendingly by size (in bytes) of corresponding OSCAR corpus. There is a significant imbalance of resources for different languages.

by classifying and filtering the Common Crawl dataset by languages. The statistics for different languages in OSCAR are shown in Table 1. Compared to unimodal datasets, multilingual multimodal datasets are much harder to obtain and are usually created based on existing monolingual multimodal datasets. For example, COCO-CN [16] enriches the original MSCOCO dataset [18] by 27K Chinese captions through manual annotations; Multi30K [2, 6, 7] translates the Flickr30K dataset [26] into German, French and Czech, resulting in around 30K new captions per language.

Machine-translated data has been utilized to create VQA datasets in target languages. Once the multi-lingual VQA datasets are available, the model can be trained using similar objectives as mono-lingual vision-language models. For example, UC^2 [39] translates captions from the Conceptual Captions [31] using Microsoft Azure Translation API into five different languages (German, French, Czech, Japanese, and Chinese) and obtains a new dataset consisting of 3.3M images with captions in six languages (including English). This machine-translated dataset is also used in subsequent works such as CCLM [37]. TD-MML [28] further translates the captions in Conceptual Captions [31] into all 19 languages in IGLUE [3] using M2M-100-large model [8]. However, this approach introduces biases when selecting the languages for translation and the resulting model performs worse on languages not included in the translated dataset. Furthermore, the effectiveness of the translation remains subject to the availability of resources for the target language, thereby placing this method at a potential disadvantage for languages with limited resources.

Code Switch is a broadly used technique in multilingual language modeling as a cost-efficient strategy to augment datasets for low-resource languages [27]. Concretely, it replaces a proportion of words in the original English text with its translation in the target language from a bilingual dictionary. Compared to machine translation, word-to-word translation requires fewer computational resources and is feasible for almost all languages in the world. However, the disadvantage is also apparent: it does not take into account any syntactical differences between English and the target language, which can result in lower-quality translation results, especially for languages having structures that vary greatly from English. It is first applied in multi-lingual multi-modal training by M^{3}P [20] and shows promising results. In M^{3}P, the authors pre-train a transformer by altering a multi-lingual text stream from Wikipedia, a mono-lingual multi-modal stream and a multi-modal code-switched stream from Conceptual Captions [31], and then fine-tune the model using Multi30K [2, 6, 7] that includes German, French and Czech, and extension of MSCOCO that contains more Japanese [36] and Chinese captions [16]. We include its results on xGQA in Table 2 for comparison.

Contrastive learning is a type of self-supervised learning widely adopted in image-text representation learning. The fundamental idea is to train a model to recognize whether two input samples are similar or dissimilar, by comparing them in a learned latent space. InfoNCE loss [35] is a popular loss function used in contrastive learning. It aims to maximize the mutual information between a set of positive pairs while minimizing the mutual information between positive and negative pairs. This approach has been widely adopted by mono-lingual VLMs at the image text level [29], as well as at the object-phrase level [5, 12, 15]. In the multi-lingual multi-modal setting, CCLM [37] applies contrastive learning to maximize the mutual information of learned embeddings from (i) paired image and caption and (ii) parallel sentence pairs of English and target languages.

$$l(x, y) = \frac{1}{N} \sum_i -log \frac{exp(x_i^T y_i/\tau)}{\sum_k exp(x_i^T y_k/\tau)} \quad (1)$$

where τ is a learnable temperature parameter. By symmetry, we have

$$l(y, x) = \frac{1}{N} \sum_i -log \frac{exp(y_i^T x_i/\tau)}{\sum_k exp(y_i^T x_k/\tau)} \quad (2)$$

The contrastive loss is usually defined as the average of the above two terms.

$$l_c(x, y) = \frac{1}{2}(l(x, y) + l(y, x)) \quad (3)$$

Concretely, this approach tries to learn latents x and y in a shared embedding space where x_i and y_i represent either paired image and text, or paired English text and non-English text.

Adapters facilitate transfer learning of large pre-trained models and have proven to be effective for cross-lingual transfer [22–24]. During fine-tuning, only weights of bottleneck layers introduced within each layer of the pre-trained transformer are trained, while the rest of the parameters remain frozen. It is more efficient than fine-tuning all parameters and more flexible than only updating the embedding layer in the pre-trained model. In xGQA [22], adapter-based methods are proposed to transfer a mono-lingual

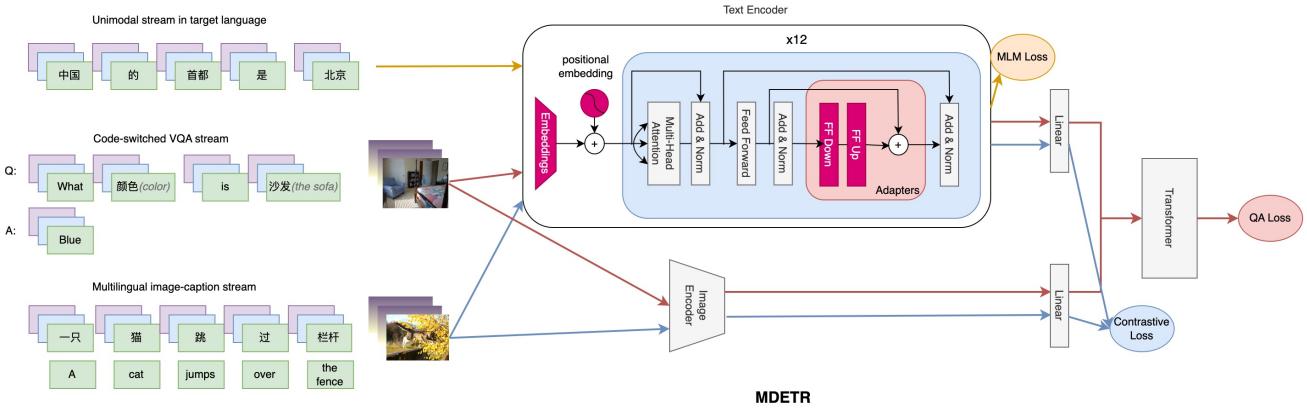


Figure 2. There are two or three streams of data in the proposed cross-lingual transfer. During training, only the embedding layer and adapters (pink) in the text encoders are updated, while all the rest weights are frozen (grey). The first stream only includes textual data in the target language and is fed to the text encoder to compute the MLM loss. The second stream consists of images from GQA and corresponding code-switched questions, fed into pre-trained MDETR for QA loss. For languages with existing image-caption datasets (such as German and Chinese), we have an additional data stream to compute the contrastive loss.

multi-modal model (OSCAR+ [17]) to a multi-lingual setting and multi-lingual models (mBERT [4]) to a multi-modal setting. The former model is called *OSCAR+*^{ada} and the latter is called *mBERT*^{ada}, compared in Table 2.

3. Method

Inspired by the success of MDETR on English grounded visual question answering, we extend the model to new languages by replacing the embedding layer, inserting adapters in the text encoder, and then training the newly added parameters with masked language modeling, code-switched QA and contrastive objectives. In this section, we first briefly review the details of MDETR framework and then describe our strategies for cross-lingual transfer.

3.1. Background

MDETR [12] is an end-to-end text-modulated detector with fine-grained multi-modal understanding capabilities. It is trained to predict bounding boxes, along with the alignment to object phrases in the query text. It uses a pre-trained convolutional backbone (ResNet-101 [10] or EfficientNet [34]) to encode image inputs and a pre-trained transformer-based language model (RoBERTa [19]) to encode text inputs, followed by a projection to a shared embedding space. The concatenation of the two modalities is fed into a transformer encoder-decoder frameworkliju. To extend it for visual question answering, MDETR uses five additional heads that are specialized for question types defined in GQA annotations, which are REL, OBJ, GLOBAL, CAT and ATTR. It is pre-trained on 1.3M image-text pairs obtained by combining Flickr30k [26], MS COCO [18], and Visual Genome (VG) [13] datasets, and fine-tuned on the GQA [11] dataset for visual question answering.

3.2. xMDETR

We first replace the tokenizer of MDETR with that of XLM-R [14], which is the state-of-the-art multi-lingual language model and covers all languages in xGQA. This also requires that the token embedding layer is also replaced, because the vocabulary size changes according to the tokenizer. We experiment with the following strategies to train the embedding layer and the adapters, while keeping the rest of the pre-trained weights frozen.

Masked Language Modeling (MLM). MLM randomly masks 15% of tokens in the input texts and aims to predict the masked tokens. It only requires a corpus of text in target languages and is widely used for pre-training language models [4, 19]. Since the goal of cross-lingual VQA is to understand questions in new languages, we adopt the MLM objective to enable MDETR to acquire knowledge of the target languages.

Code-Switched Training. Due to the complexity of VQA, MLM alone is insufficient to transfer the alignment between images and English to a new language. Given that VQA data is limited for non-English languages, we train the question-answering (QA) loss in MDETR with code-switched data. Specifically, we randomly select a proportion of words in the original queries in the GQA dataset and replace them with their corresponding translation using a bilingual dictionary. The model first predicts the question type given the original image and the code-switched text, and then predicts an answer from the relevant answer-type head as in MDETR [12]. The QA loss is the cross-entropy loss of the ground-truth answer and the predicted answer.

Adapters. Instead of fine-tuning the entire text encoder, we insert adapters at every transformer layer in the text encoder of MDETR and only update these adapters, as well

as the embedding layers, during training. Compared to updating the embedding layer alone, this approach enables the model better adapt to the new language with modest computational costs compared to full fine-tuning.

Contrastive Learning. While a high-quality non-English VQA dataset is currently lacking in the literature, certain languages have image-caption datasets that could potentially enhance the model's multi-modal understanding. To exploit these existing image-description datasets, we add additional training for German and Chinese using contrastive loss.

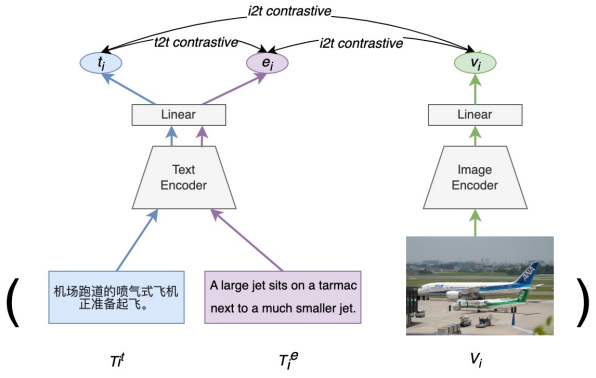


Figure 3. Illustration of proposed contrastive learning. Note that the Chinese caption means "jet planes on the airport runway are preparing to take off", which is different from the English caption. Since they describe the same image, we assume they convey similar information and thus should have close embeddings.

For each image V_i, we have an English caption T_i^e and another caption in the target language T_i^t (which is not necessarily the translation of the English caption). The images and corresponding captions are fed into the image encoder and text encoder respectively and then projected to a shared embedding space. The resulting embeddings for V_i, T_i^e and T_i^t are denoted as v_i, e_i, t_i. We take the average of the InfoNCE loss of images and English captions $l_c(v, e)$ and the InfoNCE loss of images and non-English captions $l_c(v, t)$, and denote it as the contrastive loss between images and texts l_{i2t}. Namely, we have the following using our definition of contrastive loss in Equation 3.

$$l_{i2t} = \frac{1}{2}(l_c(v, e) + l_c(v, t)) \tag{4}$$

We also compute the InfoNCE loss of English and non-English captions l_{t2t}, where we treat the caption pair describing the same image as the positive.

$$l_{t2t} = l_c(e, t) \tag{5}$$

The final contrastive loss is the average of l_{i2t} and l_{t2t}.

$$l = \frac{1}{2}(l_{i2t} + l_{t2t}) \tag{6}$$

4. Experimental Setup

4.1. Zero-Shot Cross-Lingual Transfer

We build on MDETR with EfficientNet-B5 as the backbone, which was shown to achieve the best results for visual question answering on GQA. The tokenizer is XLM-R-base from HuggingFace.[1] We train the following configurations for 100K steps with an effective batch size of 32 and a learning rate of 1e-4.

a Randomly initialize the embedding layer and fine-tune the embedding layer with MLM loss using the 2019 release of OSCAR dataset [21] from HuggingFace.[2]

b Randomly initialize the embedding layer but copy the embedding for shared tokens (overlapping words in the vocabularies of RoBERTa-base and XLM-R-base) from MDETR's embedding space. Finetune the embedding layer with MLM loss.

c Adopt the same initialization strategy as b. Fine-tune the embedding layer with both MLM loss and QA loss. To construct bilingual dictionaries for code-switched training, we first gather ground-truth bilingual dictionaries from MUSE,[3] and then use Google Translate to obtain translations for those words that are not present in MUSE but are included in the annotations of the GQA training split, to guarantee comprehensive code switching during training.

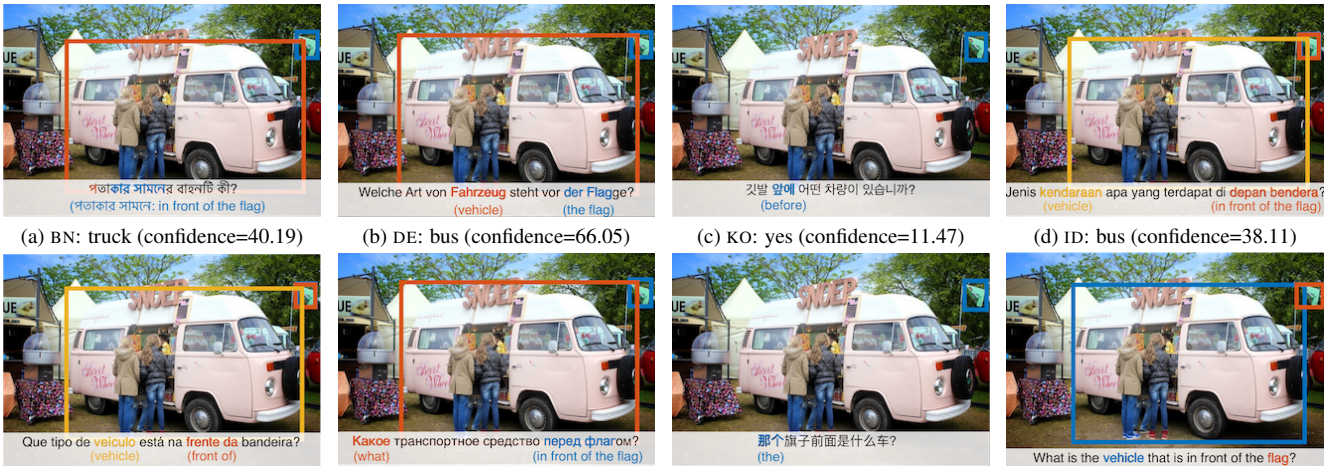
d Insert adapters into the text encoder using AdapterHub [23]. Only the embedding layer and adapters are updated, while other parameters remain frozen to decrease training time. Other settings are the same as c.

e After training d, add extra 10 epochs of contrastive learning for Chinese (with MSCOCO-CN [16]) and German (with Multi-30k [7]). A larger batch size of 32 (with gradient accumulation over 2 steps) and a lower learning rate of 1e-6 show better performance. To prevent the model from forgetting the VQA task, we alternate between contrastive loss and QA loss.

4.2. Few-Shot Cross-Lingual Transfer

We follow the same setup as the xGQA paper [22] for few-shot cross-lingual transfer. Based on the best configuration in 4.1, we continue fine-tuning the active parameters in the previous training while freezing the rest. Since all models are fine-tuned for 10 epochs in xGQA, we also perform 10 epochs of few-shot learning with QA loss. We experiment with different learning rates {1e-5, 1e-4, 2.5e-4, 5e-4,1e-3, 2e-3} and find 1e-3 leads to the highest validation accuracy.

[1] https://huggingface.co/xlm-roberta-base
[2] https://huggingface.co/datasets/oscar
[3] https://github.com/facebookresearch/MUSE



(a) BN: truck (confidence=40.19) (b) DE: bus (confidence=66.05) (c) KO: yes (confidence=11.47) (d) ID: bus (confidence=38.11)

(e) PT: bus (confidence=43.44) (f) RU: yes (confidence=70.86) (g) ZH: no (confidence=81.01) (h) EN: van (confidence=96.36)

Figure 4. Failure cases based on the 48-shot prediction example taken from the test split of xGQA. The input query is "which kind of vehicle is in front of the flag?" and the ground-truth answer is "van". MDTER shows impressive performance on English input as it correctly identifies the "van" and "vehicle" mentioned in the text and predicts the correct class "van" with high confidence. Bengali, German, Indonesian, and Portuguese models also correctly detect major objects, but some of them predict the wrong span of text and their predictions are not as accurate as in English. Russian model correctly detects both objects but fails to make the correct prediction because it misclassifies the question type. Korean and Chinese models also misclassify the question type and their detection is less accurate.

5. Results and Discussion

5.1. Quantitative Results

Not surprisingly, the combination of all techniques mentioned in Section 4.1 gives the best performance, as shown in Table 3. Note that the same technique may have different effects on different languages due to the syntactic difference. For example, copying embeddings (b) for shared tokens from MDETR leads to lower accuracy for Bengali but improves the performance of all other languages. Since this technique is proven to be useful in [22,25], we keep applying this technique for all languages in the following experiments. The code-switch strategies (c) largely improve the model as it generates synthetic VQA data for target languages. This approach is especially effective for languages with similar sentence structures as English. Portuguese and Indonesian follow the same subject-verb-object sentence format as English, which we believe is the reason why they achieve much higher accuracy compared to other languages.

Our zero-shot and few-shot results compared with other methods in xGQA are summarized in Table 2. Our model outperforms other models in Bengali, Indonesian, and Portuguese, and achieves comparable performance with the rest of the languages. However, there is still a performance gap between MDETR in English and the target languages.

We further analyze the performance on different question types in GQA, as shown in Figure 5. We see distinct performance gaps on different question types. For "object" and "global" questions, xMDETR achieves high accuracy for all languages. This is probably due to the fact that these two questions require less precise comprehension of the question. For example, answering whether an object A exists in the image only requires correctly understanding what A represents in the target language and detecting it in the given image. Similarly, answering questions about global properties, like weather, only requires understanding of some keywords in the target language. Thus, the code-switched strategy should be effective for these two types of questions. In contrast, "relation" questions require a deeper understanding of the question, and thus our model performs worse in this type of question and we see a large gap compared to its performance in English. Few-shot learning significantly improves the prediction on this type of question because it exposes the model to high-quality data in the target language. Few-shot learning also greatly improves question-type classification and minimizes the performance gap between the target language and English.

5.2. Qualitative Error Analysis

We identify some common sources of error.

1. Multiple acceptable answers & limits introduced due to the tokenizer's vocabulary. In xGQA, there is only one ground-truth label and accuracy is calculated by examining if the prediction is exactly the same as the ground truth. However, considering that all synonyms of the word are also correct, there can be more than one acceptable answer given the image and the question. We find that our model often predicts reasonable answers, however not matching the exact word

Language	Model	# Training Images					
		0	5	10	20	25	48
BN	M^{3}P	17.59	26.94	31.09	34.58	35.27	37.96
	OSCAR+^{emb}	13.35	21.67	26.61	31.94	32.78	36.97
	OSCAR+^{ada}	13.96	22.35	27.20	31.25	31.81	35.45
	mBERT^{ada}	13.38	23.10	26.55	31.60	32.26	34.18
	xMDETR	**24.08**	**28.32**	**32.08**	**36.54**	**38.10**	**39.50**
KO	M^{3}P	19.70	**32.28**	**35.50**	**37.72**	**37.84**	**38.61**
	OSCAR+^{emb}	15.11	19.99	24.78	29.48	30.43	35.59
	OSCAR+^{ada}	12.25	20.73	25.97	31.37	32.20	35.41
	mBERT^{ada}	**19.92**	27.83	31.27	34.44	35.03	36.51
	xMDETR	**19.92**	21.52	26.33	31.01	33.01	36.66
ID	M^{3}P	18.74	**37.24**	38.65	41.07	42.00	43.12
	OSCAR+^{emb}	17.89	29.76	33.59	36.69	37.31	40.51
	OSCAR+^{ada}	18.52	31.45	34.60	37.26	37.97	40.60
	mBERT^{ada}	19.77	34.49	36.26	39.15	39.81	40.88
	xMDETR	**30.96**	37.10	**40.29**	**42.91**	**44.91**	**47.75**
PT	M^{3}P	26.73	37.23	39.07	40.92	41.05	43.06
	OSCAR+^{emb}	19.36	32.42	36.37	39.01	40.15	43.27
	OSCAR+^{ada}	24.58	34.73	37.46	38.82	39.70	41.75
	mBERT^{ada}	31.45	37.31	38.88	40.51	41.03	42.62
	xMDETR	**39.74**	**40.19**	**43.38**	**45.30**	**45.90**	**50.07**
DE	M^{3}P	24.78	**39.31**	**41.05**	**42.22**	**42.54**	**43.16**
	OSCAR+^{emb}	17.49	29.09	34.48	37.35	38.45	41.08
	OSCAR+^{ada}	17.84	31.26	35.84	37.92	38.46	40.58
	mBERT^{ada}	**32.41**	37.44	39.15	40.65	41.63	42.71
	xMDETR	20.73	23.55	29.95	34.23	36.03	42.95
	xMDETR +	23.46	26.64	32.04	37.77	39.30	42.98
ZH	M^{3}P	19.66	**36.15**	**38.21**	**40.48**	40.53	42.55
	OSCAR+^{emb}	12.66	19.17	22.13	27.97	29.08	33.24
	OSCAR+^{ada}	13.20	19.67	22.74	26.81	28.19	31.69
	mBERT^{ada}	**26.16**	32.93	35.82	38.22	37.89	39.57
	xMDETR	21.00	23.31	29.30	33.23	35.00	41.38
	xMDETR +	23.92	28.43	33.15	38.74	**40.68**	**43.38**
RU	M^{3}P	24.29	**36.71**	**38.53**	**39.94**	**40.13**	41.85
	OSCAR+^{emb}	7.98	23.72	28.21	32.15	32.87	36.84
	OSCAR+^{ada}	16.38	27.42	30.17	33.22	34.21	37.28
	mBERT^{ada}	**25.51**	31.69	32.47	34.93	35.53	37.42
	xMDETR	23.13	28.36	31.40	36.94	37.61	**42.51**

Table 2. Test accuracy for zero-shot (the 0 column) and few-shot results on xGQA. **Bold** numbers are the highest and underlined numbers are the second highest in each column for each individual language.

Configuration added		BN	DE	ID	KO	PT	RU	ZH	AVG
a	random embedding	16.16	8.01	6.40	8.28	6.50	10.51	4.82	8.67
b	lexical embedding	8.60	16.35	11.07	11.62	15.85	12.69	12.59	12.68
c	+ code switch	18.41	19.91	25.12	17.06	31.74	19.98	19.15	21.62
d	+ adapter	**24.08**	20.73	**30.96**	**19.92**	**39.74**	**23.13**	21.00	**25.65**
e	+ contrastive	-	**23.46**	-	-	-	-	**23.92**	-

Table 3. Test accuracy for different configurations in Section 4.1. "Random embedding" represents MLM training with random initialization of the embedding; "lexical embedding" represents MLM training with initialization of embedding that reuses English embedding for shared tokens; "code switch" represents integrating the code-switch strategy; "adapter" represents inserting adapters into the text encoder ; "contrastive" represents additional multi-modal training with contrastive objectives. **Bold** numbers are the highest score for each language.

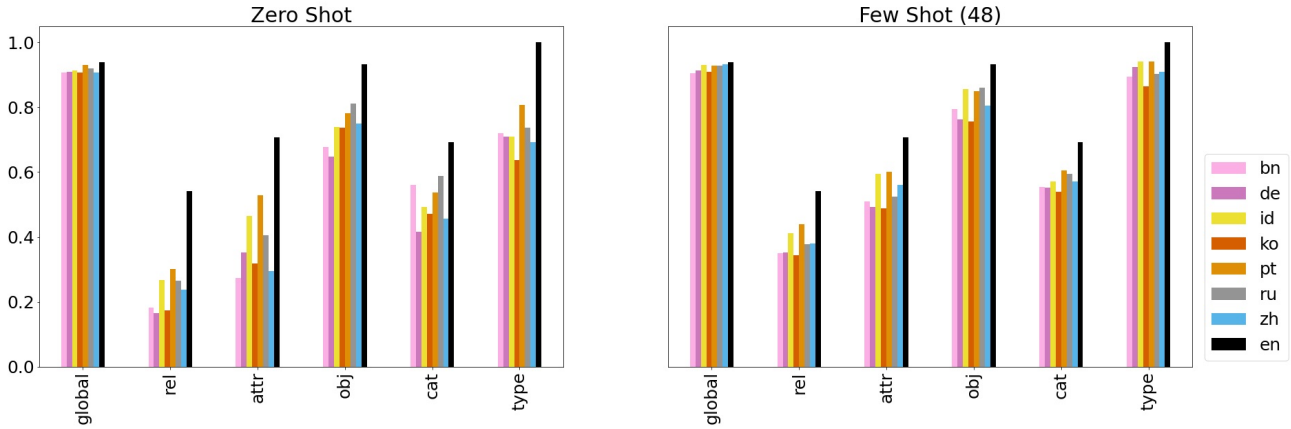


Figure 5. The accuracy of xMDETR on different question types by languages. The GQA dataset [11] defines five semantic question types: (1) global: overall properties of the scene; (2) relation (rel): subject or object of described relation; (3) attribute(attr): properties of the object; (4) object: the existence of the object (yes/no/unknown); (5) category (cat): class of the object. The last entry ("type") represents the accuracy of question type classification, which is the first step of MDETR predicting the answer to the given query. Besides the seven languages in xGQA, we also include English here as a benchmark. Since MDETR is already fine-tuned on English GQA, the English results on both plots represent the fully fine-tuned results.

(a) **Question** (DE): Welche Art von Spielzeug ist weich? (Which kind of toy is soft?); **GT**: stuffed toy; **Prediction**: teddy bear

(b) **Question** (ID): Jenis kendaraan apa yang terbuat dari logam? (Which kind of vehicle is metallic?); **GT**: truck; **Prediction**: car

Figure 6. Examples of xMDETR making reasonable predictions but not matching the ground-truth labels.

given by the ground truth in the dataset. For example, "teddy bear" should be a correct prediction because the "stuffed toy" in Figure 6a is indeed a teddy bear. Similarly, both "car" and "truck" are correct in the context of Figure 6b.

2. Wrong answer type classification. Since the first step of prediction is a question-type classification, a failure in this stage will directly cause the failure of the whole prediction, as illustrated in the Russian (4f), Korean (4c) and Chinese (4g) examples. Since it is a pure NLP classification task, providing more question data might alleviate this issue.

3. Challenge of comprehending the referring expression. Some questions (especially for the "relation" type) involve complex reasoning that requires a strong semantic understanding of the question. In Figure 4, the model implicitly performs the following logical steps (i) detect the flag object A (ii) locate the object B in front of A (iii) classify the object B. Techniques such as code-switching, while effective, seem not to be enough to fully transfer knowledge from English to the target language, at the level required for such multi-hop reasoning.

6. Conclusion

In this paper, we extend the state-of-the-art mono-lingual multi-modal model, MDETR, to a multi-lingual setting. Our proposed method integrates multi-lingual unimodal training (MLM objective) and multi-lingual multi-modal training (QA objective with code-switch GQA data and contrastive objective with multi-lingual image-caption data) with adapters for accelerated training. Our approach doesn't use any machine-translated datasets and can be easily extended to more languages. By leveraging MDETR's powerful text-modulated detection, our model is able to provide bounding boxes with predicted alignment to key objects mentioned in the questions in target languages, which allows for better interpretability. Our model shows competitive results on the xGQA dataset. However, there is still a large gap between the performance of our model and human-level performance due to the complex nature of multilingual visual question-answering. Therefore, it is crucial to exercise caution and consider the limitations discussed in Section 5.2 when applying this model in real world.

Acknowledgements

This work was supported in part through the NYU IT High Performance Computing resources, services, and staff expertise. It is based upon work supported in part by the National Science Foundation under NSF Award 1922658.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 1

[2] Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. Findings of the third shared task on multimodal machine translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 304–323, 2018. 3

[3] Emanuele Bugliarello, Fangyu Liu, Jonas Pfeiffer, Siva Reddy, Desmond Elliott, Edoardo Maria Ponti, and Ivan Vulić. IGLUE: A benchmark for transfer learning across modalities, tasks, and languages. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2370–2392. PMLR, 17–23 Jul 2022. 2, 3

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. 4

[5] Zi-Yi Dou, Aishwarya Kamath, Zhe Gan, Pengchuan Zhang, Jianfeng Wang, Linjie Li, Zicheng Liu, Ce Liu, Yann LeCun, Nanyun Peng, et al. Coarse-to-fine vision-language pre-training with fusion in the backbone. *arXiv preprint arXiv:2206.07643*, 2022. 3

[6] Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. Findings of the second shared task on multimodal machine translation and multilingual image description. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 215–233, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. 3

[7] Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. Multi30k: Multilingual english-german image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74. Association for Computational Linguistics, 2016. 3, 5

[8] Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48, 2021. 3

[9] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 1

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 4

[11] Drew Hudson and Christopher Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. pages 6693–6702, 06 2019. 1, 2, 4, 8

[12] Aishwarya Kamath, Mannat Singh, Yann LeCun, Ishan Misra, Gabriel Synnaeve, and Nicolas Carion. Mdetr–modulated detection for end-to-end multi-modal understanding. *arXiv preprint arXiv:2104.12763*, 2021. 1, 3, 4

[13] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. 2016. 1, 4

[14] Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 4

[15] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022. 3

[16] Xirong Li, Xiaoxu Wang, Chaoxi Xu, Weiyu Lan, Qijie Wei, Gang Yang, and Jieping Xu. COCO-CN for cross-lingual image tagging, captioning and retrieval. *CoRR*, abs/1805.08661, 2018. 3, 5

[17] Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. *ECCV 2020*, 2020. 4

[18] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. 3, 4

[19] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. 4

[20] Minheng Ni, Haoyang Huang, Lin Su, Edward Cui, Taroon Bharti, Lijuan Wang, Dongdong Zhang, and Nan Duan. M3p: Learning universal representations via multitask multilingual multimodal pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3977–3986, June 2021. 1, 3

[21] Pedro Javier Ortiz Su'arez, Benoit Sagot, and Laurent Romary. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pages 9 – 16, Mannheim, 2019. Leibniz-Institut f'ur Deutsche Sprache. 2, 5

[22] Jonas Pfeiffer, Gregor Geigle, Aishwarya Kamath, Jan-Martin Steitz, Stefan Roth, Ivan Vulić, and Iryna Gurevych. xGQA: Cross-lingual visual question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2497–2511, Dublin, Ireland, May 2022. Association for Computational Linguistics. 1, 2, 3, 5, 6

[23] Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 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, Oct. 2020. Association for Computational Linguistics. 3, 5

[24] Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 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, Nov. 2020. Association for Computational Linguistics. 3

[25] Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. UNKs everywhere: Adapting multilingual language models to new scripts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10186–10203, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. 6

[26] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2641–2649, 2015. 3, 4

[27] Libo Qin, Minheng Ni, Yue Zhang, and Wanxiang Che. Cosda-ml: Multi-lingual code-switching data augmentation for zero-shot cross-lingual nlp, 2020. 3

[28] Chen Qiu, Dan Oneata, Emanuele Bugliarello, Stella Frank, and Desmond Elliott. Multilingual multimodal learning with machine translated text, 2022. 1, 3

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021. 3

[30] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VIII*, pages 146–162. Springer, 2022. 1

[31] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*, 2018. 3

[32] Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. A corpus of natural language for visual reasoning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 217–223, 2017. 1

[33] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019. 1

[34] Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR, 09–15 Jun 2019. 4

[35] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018. 3

[36] Yuya Yoshikawa, Yutaro Shigeto, and Akikazu Takeuchi. STAIR captions: Constructing a large-scale Japanese image caption dataset. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 417–421, Vancouver, Canada, July 2017. Association for Computational Linguistics. 3

[37] Yan Zeng, Wangchunshu Zhou, Ao Luo, and Xinsong Zhang. Cross-view language modeling: Towards unified cross-lingual cross-modal pre-training. *arXiv preprint arXiv:2206.00621*, 2022. 1, 3

[38] Zelin Zhao, Karan Samel, Binghong Chen, and Le Song. Proto: Program-guided transformer for program-guided tasks, 2021. 1

[39] Mingyang Zhou, Luowei Zhou, Shuohang Wang, Yu Cheng, Linjie Li, Zhou Yu, and Jingjing Liu. Uc2: Universal cross-lingual cross-modal vision-and-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4155–4165, June 2021. 1, 3