Unifying Cross-Lingual and Cross-Modal Modeling Towards Weakly Supervised Multilingual Vision-Language Pre-training

Zejun Li[1], **Zhihao Fan**[1], **JingJing Chen**[2], **Qi Zhang**[2],
Xuanjing Huang[2,3], **Zhongyu Wei**[1,4*]
[1]School of Data Science, Fudan University, China
[2]School of Computer Science, Fudan University, China
[3]Shanghai Collaborative Innovation Center of Intelligent Visual Computing, China
[4]Research Institute of Intelligent and Complex Systems, Fudan University, China
{zejunli20,fanzh18,chenjingjing,qz,xjhuang,zywei}@fudan.edu.cn

Abstract

Multilingual Vision-Language Pre-training (VLP) is a promising but challenging topic due to the lack of large-scale multilingual image-text pairs. Existing works address the problem by translating English data into other languages, which is intuitive and the generated data is usually limited in form and scale. In this paper, we explore a more practical and scalable setting: weakly supervised multilingual VLP with only English image-text pairs and multilingual text corpora. We argue that the universal multilingual representation learned from texts allows the cross-modal interaction learned in English to be transferable to other languages. To this end, we propose a framework to effectively unify cross-lingual and cross-modal pre-training. For unified modeling on different data, we design an architecture with flexible modules to learn different interactions. Moreover, two unified tasks are introduced to efficiently guide the unified cross-lingual cross-modal learning. Extensive experiments demonstrate that our pre-trained model learns universal multilingual multimodal representations, allowing effective cross-lingual transfer on multimodal tasks. Code and models are available at https://github.com/FudanDISC/weakly-supervised-mVLP.

1 Introduction

In recent years, self-supervised pre-training technology has been studied extensively in various fields. The pre-trained models are able to encode generalized contextual representations for texts (Devlin et al., 2019; Liu et al., 2019; Lewis et al., 2020), images (Bao et al., 2021a; He et al., 2022), and image-text pairs (Chen et al., 2020; Li et al., 2020, 2021a), further facilitating the downstream tasks and research. However, most pre-training studies are limited to English corpora. In order to overcome the language barrier and benefit a wider range of audience, it is important to extend the success of English-centric research to a multilingual scenario. Recent works have demonstrated the effectiveness of cross-lingual language modeling (Conneau and Lample, 2019; Conneau et al., 2020). Based on large-scale multilingual corpora, the models are able to learn universal representations for texts in multiple languages.

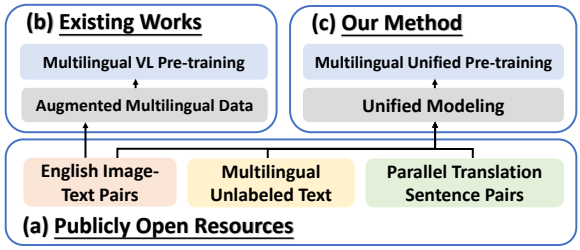

Figure 1: An illustration of the differences between existing multilingual VLP methods and our method.

However, large-scale and tightly associated multilingual image-text pairs are unavailable and costly to acquire. Therefore, it is not straightforward to transfer existing VLP methods to other languages. As shown in Figure 1 (a, b), previous works (Ni et al., 2021; Zhou et al., 2021) address the problem by transferring English data to other languages through different data augmentation strategies (e.g., code switch or translation engines), and then perform VLP on the generated multilingual data. Despite being simple and intuitive, these methods have limitations since the augmentation is either constrained to specific forms that differ from natural data, or it is time-consuming to ensure the equality, making it difficult to scale effectively to more languages and larger datasets. Meanwhile, large volumes of natural language texts are readily accessible in various languages as shown in Figure 1 (a). This raises a question: can we relax the requirement for multilingual image-text pairs and use existing multilingual text resources to transfer English VLP to other languages?

*Corresponding author.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 5939–5958

In this paper, we explore this weakly supervised setting to develop a more scalable multilingual VLP framework. We argue that by unifying cross-lingual text modeling in cross-modal models, universal representations can be learned for multiple languages, and thus the cross-modal modeling ability learned from English image-text pairs can be easily transferred to other languages. The biggest challenge for such unification is that data in different forms (i.e., different languages or modalities) have different intrinsic properties. Therefore, the key issue is how to effectively incorporate multilingual text pre-training into the VLP framework without conflicts among different data streams.

From the model perspective, if we simply feed all data to a vanilla Transformer model (Vaswani et al., 2017), different desired interactions are entangled in self-attention layers and may compete with each other during pre-training. To address this issue, we propose to disentangle different functionalities into different modules. Specifically, we design a novel architecture by incorporating pluggable cross-attention layers into standard Transformer layers. These layers can be activated to perform cross-modal and cross-sentence modeling or skipped for unpaired text modeling. Multilingual text learning can thus transfer the self-attention to fit more languages, indirectly requiring the cross-attention to adapt to the universal representations rather than competing with cross-lingual modeling in self-attention.

In terms of training, VLP and language modeling methods tend to optimize different objectives. For unified pre-training, we introduce two tasks that share unified formulations for different data streams and guide unified cross-lingual and cross-modal learning. Before the cross-modal fusion, we propose unified contrastive learning to simultaneously align parallel sentences in different languages and English image-text pairs, making it easier for the upper encoder to learn interactions shared across languages and modalities. On top of the whole model, we consider three self-supervised tasks: cross-lingual masked language modeling on unpaired texts to achieve universal multilingual representations, visual language modeling on English image-text pairs to learn cross-modal interaction, and translation language modeling on parallel sentences to enhance cross-lingual alignment. Three tasks are further unified as a mask modeling task and cross-lingual learning is naturally unified with cross-modal learning.

Our contributions can be summarized as follows:

- We explore weakly supervised multilingual VLP by unifying cross-lingual modeling from multilingual text corpora and cross-modal modeling from English image-text corpora.
- To effectively unify multilingual text modeling with VLP, we introduce a flexible architecture to consistently encode different data streams and unified pre-training tasks to efficiently learn different capabilities from them.
- We conduct extensive experiments to validate the effectiveness of our approach. Our pre-trained model can encode universal multilingual multimodal representations, enabling effective cross-lingual and cross-modal transfer.

2 Related Works

2.1 Multi-modal Pre-training

Vision-Langauge Pre-training VLP methods aim to learn generalized representations for image-text pairs. To represent images with visual sequences, pioneer works employ frozen object detectors to extract region features from images (Lu et al., 2019; Su et al., 2019), recent works demonstrate the effectiveness of end-to-end VLP with vision Transformers (CNNs) to encode patch (grid) features (Huang et al., 2020; Kim et al., 2021; Li et al., 2021a). As for the architecture, single-stream models first concatenate the textual and visual sequences, then encode the multimodal sequences with self-attention layers where intra-modality and cross-modality interactions are jointly learned (Su et al., 2019; Chen et al., 2020; Li et al., 2020). Two-stream models further disentangle the process with separate self-attention layers and cross-attention layers (Lu et al., 2019; Tan and Bansal, 2019; Li et al., 2021a). Due to the absence of large-scale multilingual image-text pairs, most VLP methods are only able to handle English inputs.

Unified Pre-training Li et al. (2021b) first explore unified pre-training on texts, images, and image-text pairs. To make parameters efficiently shared across modalities, Bao et al. (2021b) propose an architecture with modality-specific experts. Recent works further extend the idea for large-scale pre-training (Wang et al., 2022a,b). Unified pre-training benefits both multimodal and uni-modal

learning. In this work, we claim that unifying multilingual text pre-training helps overcome the language barrier in previous VLP methods.

2.2 Multilingual Pre-training

Multilingual Langauge Modeling Multilingual BERT (Devlin et al., 2019) first validates the effectiveness of masked language modeling on an unlabeled multilingual corpus. Universal cross-lingual representations are learned, allowing effective cross-lingual transfer on downstream tasks. XLM (Conneau and Lample, 2019) and Unicoder (Huang et al., 2019) enhance cross-lingual alignment with additional tasks on parallel translation corpora. XLM-R (Conneau et al., 2020) further scales up the cross-lingual pre-training in terms of the number of languages and the amount of data.

Multilingual VLP MURAL (Jain et al., 2021) extends the contrastive framework in (Radford et al., 2021) by explicitly aligning different languages, but the dual-encoder architecture is not capable of fulfilling reasoning tasks like VQA. The most related works to ours are M^{3}P (Ni et al., 2021) and UC^2 (Zhou et al., 2021). Both methods address the data problem through different augmentation methods. M^{3}P generates code-switched pairs in which English words are replaced with their translation in other languages. UC^2 utilizes translation engines to transform English image captions into other languages, CCLM (Zeng et al., 2022) further extends this idea with existing translation pairs to enhance the cross-lingual alignment but CCLM relies on a larger backbone to show its effectiveness. In contrast to using generated multilingual pairs, we explore weakly supervised multilingual VLP with unified pre-training on existing resources.

3 Method

3.1 Data Stream

Different from prior works, our approach does not rely on multilingual image-text pairs and explores weak supervision in available datasets. We consider N languages $\{l_i\}_{i=1}^{N}$ including English and adopt three parts of publicly open resources. To learn cross-lingual modeling, we utilize a multilingual text corpus $D_m = \cup_{i=1}^{N}\{T_j^{l_i}\}_{j=1}^{N_{l_i}}$, where $T_j^{l_i}$ is the j-th sentence in language l_i and N_{l_i} is the number of sentences in language l_i. Following (Conneau and Lample, 2019), we make use of parallel translation corpora $D_t = \cup_{i=1}^{N}\{(T^{\text{en}}, T^{l_i})_j\}_{j=1}^{N'_{l_i}}$ to learn cross-lingual alignment, where $(T^{\text{en}}, T^{l_i})_j$ is an English-l_i translation sentence pair and N'_{l_i} is the size of the English-l_i dataset. In order to learn cross-modal modeling, we adopt an English image-text corpus $D_v = \{(I, T^{\text{en}})_i\}_{i=1}^{N_m}$, where $(I, T^{\text{en}})_i$ is an English image-sentence pair and N_m is the number of paired samples.

For a sentence T^{l_i} in language l_i, we employ the learned multilingual SentencePiece (Kudo and Richardson, 2018) tokenizer in (Conneau et al., 2020) to transform it into tokens $t^{l_i} = \{t_{\text{cls}}, t_1^{l_i}, ..., t_n^{l_i}, t_{\text{sep}}\}$. All languages share special tokens like CLS and SEP. Following (Dosovitskiy et al., 2020), each 2D image $I \in \mathbb{R}^{H\times W\times C}$ is split into $M = HW/P^2$ fixed-size patches, where C is the number of channels, (H, W) is the image resolution, and (P, P) is the size of each patch. An image is further represented by a visual sequence $\{v_{\text{cls}}, v_1, ..., v_M\}$, each visual token $v_i \in \mathbb{R}^{P^2\times C}$ is a flattened vector of pixel values in the corresponding patch, v_{cls} is a special embedding vector to gather the global information.

3.2 Unified Model Architecture

As our model is required to handle inputs of different forms, we introduce a novel unified model architecture as shown in Figure 2. To disentangle intra-modality and cross-modality modeling, we follow (Tan and Bansal, 2019; Li et al., 2021a) to construct a two-stream model that consists of an image encoder, a text encoder, and a high-level unified encoder. In Figure 2, we use the colors of rectangles to indicate the data flow in our model. Textual tokens and image patches are first fed to the corresponding uni-modal encoders to perform intra-modality interaction. The text and image encoders are standard Transformer (Vaswani et al., 2017) encoders with N_L and N_V layers respectively.

Based on uni-modal representations, a N_C-layer unified high-level encoder is learned. To consistently encode different data streams, we introduce a novel architecture for the high-level unified encoder. Each unified layer comprises a self-attention layer, a feedforward layer, and a pluggable cross-attention layer. We consider different routines for different data streams. Once the cross-attention layers are activated, the encoder serves as a condition-grounded text encoder where the conditional information comes from the paired images or translation source sentences in another language. The cross-attention can be skipped for unconditional

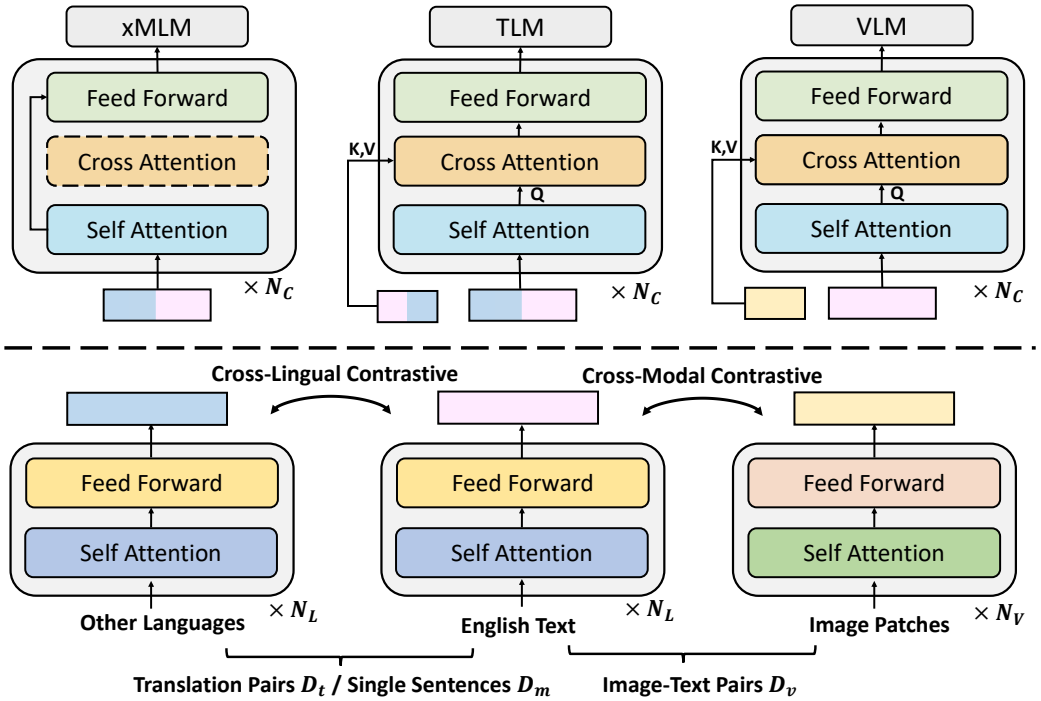

Figure 2: The proposed pre-training framework. The bottom part illustrates the unified contrastive learning on top of uni-modal encoders. The top part displays the unified MLM based on the proposed unified high-level encoder. Modules with the same color share the same parameters. We use the colors of the rectangles to indicate the data flow in the model, rectangles with mixed colors denote data streams that consist of texts in both English and other languages. Residual connections in the model architecture are omitted for brevity.

text modeling on unpaired texts. Different routines are comprehensively illustrated in Figure 5. Compared with the previous methods to entangle cross-modal and intra-model interaction in a single self-attention layer, our unified architecture disentangles different functionalities into different modules. Such a design would allow the knowledge learned from different data streams to be better integrated into the unified model without conflicts.

3.3 Unified Pre-training Tasks

We propose two pre-training tasks: unified contrastive learning and unified masked language modeling. These tasks share unified formulations for different data streams and help the unified model acquire cross-lingual and cross-modal modeling capabilities from them.

3.3.1 Unified Contrastive Learning

As introduced in Section 3.2, the unified high-level encoder relies on attention to perform cross-modal, cross-sentence, and intra-sentence interactions. In order to make the learned modeling capability transferable across languages and modalities, we propose to learn an aligned cross-lingual cross-modal semantic space on top of the uni-modal encoders.

Since no multilingual image-text pairs are accessible, we propose unified contrastive learning (UCL) to simultaneously align cross-lingual texts and English image-text pairs. UCL is based on InfoNCE loss (Oord et al., 2018):

$$\mathcal{L}_{\text{UCL}} = -\mathbb{E}_{(a,b)\sim D_{v,t}}[\log \frac{\exp(s(a,b)/\tau)}{\sum_{\hat{b}\in B}\exp(s(a,\hat{b})/\tau)} + \log \frac{\exp(s(a,b)/\tau)}{\sum_{\hat{a}\in A}\exp(s(\hat{a},b)/\tau)}] \quad (1)$$

where (a, b) is a image-text pair or translation pair sampled from $D_{v,t} = D_v \cup D_t$. A is a batch including the positive sample a and $|A| - 1$ negative samples, the same for B. $s(a, b)$ computes the global similarity between a and b, which is the cosine similarity between the uni-modal CLS representations of a and b. τ is the learnable temperature.

UCL employs English texts as natural anchors to bridge both the language and modality gap.

3.3.2 Unified Masked Language Modeling

To learn token-level contextual representations, the effectiveness of masked language modeling (MLM)

has been validated in various domains (Devlin et al., 2019; Chen et al., 2020; Conneau and Lample, 2019). We consider three variants of MLM: cross-lingual MLM (xMLM) on multilingual sentences in D_m, translation language modeling (TLM) on translation pairs in D_t, and visual language modeling (VLM) on image-text pairs in D_v.

xMLM is a standard MLM task on multilingual texts. As pointed out in previous works (Conneau et al., 2020; Artetxe et al., 2020), it enhances cross-lingual text modeling and endows universal representations for multiple languages.

TLM and VLM can be unified as conditional MLM with complement information available. VLM trains the model to learn visually grounded representations by cross-modal interaction. TLM helps the model learn token-level alignment across languages through cross-sentence interaction. At the same time, TLM and VLM are consistent with each other for two reasons: our unified encoder is agnostic to the modality of the conditional information and the representations of conditions in different modalities are aligned through UCL.

Generally, three tasks share the same mask-then-predict paradigm. The target token sequence is masked with 0.15 probability following (Devlin et al., 2019). The model is optimized to recover the original tokens based on the contextual outputs of the unified high-level encoder. The unified MLM loss is formulated as:

$$\begin{aligned}\mathcal{L}_{\text{MLM}} = &- \mathbb{E}_{t^{l_i}\sim D_m} \log P_{\text{MLM}}(t_m^{l_i}|t_{\setminus m}^{l_i}) \\ &- \mathbb{E}_{(t^{l_i},t^{l_j})\sim D_t} \log P_{\text{MLM}}(t_m^{l_i}|t_{\setminus m}^{l_i}, t^{l_j}) \\ &- \mathbb{E}_{(t^{\text{en}},I)\sim D_v} \log P_{\text{MLM}}(t_m^{\text{en}}|t_{\setminus m}^{\text{en}}, I) \quad (2)\end{aligned}$$

where $t_m^{l_i}$ and $t_{\setminus m}^{l_i}$ denote the masked and masked tokens respectively, P_{MLM} is the predicted distribution over the vocabulary for masked tokens.

In addition, we adopt a commonly-used task, image-text matching (ITM), for global cross-modal learning. ITM is a binary classification task based on the image-grounded text encoder:

$$\begin{aligned}\mathcal{L}_{\text{ITM}} = -\mathbb{E}_{(T^{\text{en}},I)\sim D_v}[&\log(P_{\text{ITM}}(T^{\text{en}}, I)) \\ &+ \log(1 - P_{\text{ITM}}(\hat{T}^{\text{en}}, \hat{I}))] \quad (3)\end{aligned}$$

where P_{ITM} is the predicted matching probability. $(\hat{T}^{\text{en}}, \hat{I})$ is a negative pair, we follow (Li et al., 2021a) to utilize the similarities $s(a, b)$ in Equation 1 to perform in-batch hard negative sampling.

4 Experiments

4.1 Pre-training Details

Pre-training Corpora We consider 21 languages including English to cover the target languages in downstream tasks. We construct D_v by including 4M image-text pairs from Conceptual Captions (Sharma et al., 2018), MSCOCO (Lin et al., 2014), and Visual Genome (Krishna et al., 2017). D_t is composed of 19M parallel translation pairs between English and other 20 languages collected from WikiMatrix (Schwenk et al., 2021). As for D_m, we adopt CC-100[1] which is an open-source recreation of the dataset for training XLM-R (Conneau et al., 2020), we sample a subset of 0.8B sentences following the language distribution used in XLM-R. More details are in Appendix B.1.1.

Implementation Details For each transformer layer, we consider the base size in (Devlin et al., 2019) and we set $N_L = N_C = 6$ and N_V=12. The image encoder is initialized from (Li et al., 2021a), while the textual encoder and the high-level encoder are initialized from the first 6 and last 6 layers of XLM-R (Conneau et al., 2020) respectively. As XLM-R does not contain cross-attention layers, we initialize those layers in our high-level encoder with the parameters of self-attention layers.

Our model is pre-trained to minimize $\mathcal{L}_{\text{UCL}} + \mathcal{L}_{\text{MLM}} + \mathcal{L}_{\text{ITM}}$ for 240K steps with AdamW (Loshchilov and Hutter, 2018) optimizer. Each training batch comprises 512 image-text pairs, 2048 translation pairs, and 2048 multilingual sentences. The learning rate is warmed-up from 0 to 1e-4 in the first 24K steps and then linearly decays to 0. Based on the ZERO-2 optimization and half-precision training under the framework of DeepSpeed (Rasley et al., 2020), the pre-training takes around 6 days on 8 RTX 3090 GPUs. More pre-training hyper-parameters are provided in Appendix B.1.3. Notice that our method can be easily scaled up in terms of the number of languages, the model size, and the scale of the dataset used. We adopt the current setup for a fair comparison with existing models.

4.2 Downstream Tasks

To comprehensively evaluate the learned universal multilingual multimodal representations, We conduct experiments on downstream vision-language (V-L) and text tasks under different settings.

[1] https://data.statmt.org/cc-100/

Model	VNLI	VQA	Reasoning	Retrieval			
	XVNLI	xGQA	MaRVL	xFlickr&CO		WIT	
				IR	TR	IR	TR
mUNITER	53.7 (76.4)	10.0 (54.7)	53.7 (71.9)	8.1 (44.5)	8.9 (40.9)	9.2 (19.9)	10.48 (22.3)
xUNITER	58.5 (75.8)	21.7 (54.8)	54.6 (71.6)	14.0 (38.5)	13.5 (32.1)	8.7 (16.7)	9.8 (18.5)
UC^2	62.1 (76.4)	29.4 (55.2)	57.3 (70.6)	20.3 (37.4)	17.9 (34.6)	7.8 (17.9)	9.1 (19.7)
M^{3}P	58.3 (76.9)	28.2 (53.8)	56.0 (68.2)	12.9 (31.4)	11.9 (24.6)	8.1 (15.5)	10.0 (15.3)
Ours	**69.5** (79.7)	**42.1** (57.4)	**62.1** (75.3)	**59.8** (86.6)	**58.7** (91.7)	**36.3** (56.0)	**36.6** (56.2)

Table 1: Zero-shot performance of multilingual VLP models trained on English and evaluated on target languages in IGLUE (Bugliarello et al., 2022). The results are averaged over all target languages. IR and TR are short for image retrieval and text retrieval respectively. Numbers in brackets are evaluation results on the English test sets.

Cross-lingual transfer on V-L tasks To validate that the learned cross-modal modeling capability can be transferred across languages, we evaluate our method on the IGLUE (Bugliarello et al., 2022) benchmark. IGLUE incorporates different kinds of tasks including visually-grounded NLI (VNLI) in XVNLI, visual question answering (VQA) in xGQA (Pfeiffer et al., 2022), V-L reasoning in MaRVL (Liu et al., 2021), and image-text retrieval in xFlickr&CO and WIT (Srinivasan et al., 2021). For all tasks, we consider the zero-shot language transfer setting where the model is trained in English and directly evaluated in other languages. Accuracies are reported for XVNLI, xGQA, and MaRVL. As for xFlickr&CO and WIT, while for text-to-image retrieval and image-to-text retrieval, recall is adopted as the evaluation metric.

Multilingual fine-tuning on V-L tasks Following (Ni et al., 2021; Zhou et al., 2021), we adapt our method to multilingual image-text retrieval task on the multilingual extensions of MSCOCO (Lin et al., 2014) and Flickr30K (Young et al., 2014). The training and test sets are valid in all languages. Mean recall (mR) is used as the evaluation metric, which is the average of recall at $K = \{1, 5, 10\}$ of both image and text retrieval. In addition to retrieval, we fine-tune our model on 2 multilingual VQA datasets: Japanese-VQA (Shimizu et al., 2018) and FM-IQA Chinese (Gao et al., 2015).

Cross-modal transfer from L to V-L As our method is a unified model, we evaluate the cross-lingual transfer ability for text modeling on xNLI (Conneau et al., 2018). Assuming that the learned modeling capability can be even transferred across modalities, we consider a zero-shot modality-transfer task from NLI to VNLI: models are trained with sentence pairs in the English SNLI dataset (Bowman et al., 2015) and directly evaluated on image-text pairs in XVNLI. More details of different tasks, datasets, and the corresponding fine-tuning settings are summarized in Appendix B.2.

4.3 Compared Models

Baseline The baseline method adopted in the experiment is (Liu et al., 2021), which employs the UNITER architecture and pre-trains with MLM on both cross-lingual texts and English image-text pairs. It can be regarded as the baseline method of ours without the unified architecture and unified pre-training tasks. Two variants of pre-trained models named as mUNITER and xUNITER, are generated by initializing from mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020).

Multulingual VLP We also compare with two existing multilingual VLP models named as M^{3}P (Ni et al., 2021) and UC^2 (Zhou et al., 2021), respectively. Both models are initialized from XLM-R and rely on data augmentation. M^{3}P extends xUNITER with code-switched image-text pairs where English words are randomly replaced with their translation in other languages. UC^2 utilizes translation engines to transform English image captions into other 5 languages. Based on the generated multilingual pairs, M^{3}P performs standard VLP with commonly used tasks while UC^2 introduces 2 more tasks to enhance cross-lingual and cross-modal modeling.

4.4 Main Results

Cross-Lingual Transfer As shown in Table 1, our model shows a superior cross-lingual zero-shot ability on various V-L tasks. For XVNLI, xGQA, and MaRVL, compared with other methods, our method achieves significant performance improvements in other languages, bridging the perfor-

Method	XNLI						SNLI → XVNLI					
	en	ar	es	fr	ru	mean	en	ar	es	fr	ru	mean
XLM-R	**85**.8	**73.8**	**80.7**	**79.7**	**78.1**	**78.1**	-	-	-	-	-	-
UC^2	83.4	65.9	74.5	74.0	72.4	71.7	54.2	37.4	45.0	48.4	41.5	43.1
Ours	82.7	73.0	77.8	78.5	75.4	76.2	**71.5**	**53.9**	**57.8**	**60.1**	**58.2**	**57.5**

Table 2: Cross-lingual and cross-modal zero-shot transfer performance. Models are fine-tuned on English NLI datasets and evaluated on NLI and VNLI datasets in other languages.

Model	Flickr30K				MSCOCO		
	EN	DE	FR	CS	EN	ZH	JA
English-only Fine-tune							
UC^2	87.2	74.9	74	67.9	88.1	82	71.7
M^{3}P	87.4	58.5	46.0	36.8	88.6	53.8	56.0
Ours	**94.9**	**84.4**	**86.1**	**77.2**	**89.6**	**83.3**	**73.1**
Single-Language Fine-tune							
UC^2	87.2	83.8	77.6	74.2	88.1	84.9	87.3
M^{3}P	87.4	82.1	67.3	65.0	88.6	75.8	80.1
Ours	**94.9**	**92.5**	**92.4**	**91.0**	**89.6**	**92.5**	**90.4**
All-Language Fine-tune							
UC^2	88.2	84.5	83.9	81.2	88.1	89.8	87.5
M^{3}P	87.7	82.7	73.9	72.2	88.7	86.2	87.9
Ours	**95.3**	**93.6**	**93.8**	**92.4**	**90.4**	**92.6**	**90.0**

Table 3: Multilingual image-text retrieval performance on Flickr30K and MSCOCO across multiple languages.

mance gap between English and target languages. For retrieval task, the superior performance in English is also effectively transferred to other languages.These results validate the effectiveness of our pre-training framework and under our framework, the cross-modal modeling capability learned from English image-text corpora can be transferred across languages, since our model learns universal multilingual multi-modal representations.

In addition, our model is able to perform cross-lingual transfer on text-only tasks. The results are listed in the left part of Table 2. As a unified model, our pre-trained model is better than UC^2 but slightly worse than XLM-R, we attribute this to that we only sample a small part of the corpus used in XLM-R. Nevertheless, it shows that cross-lingual modeling capabilities for text and image-text pairs are consistently integrated in our model.

Cross-Modal Transfer Unlike previous methods, our unified framework endows cross-model transfer capability and achieves better cross-modal zero-shot transfer performances as shown in the right part of Table 2. The learned interaction between sentence pairs can be directly applied to perform image-text interaction. It further validates our claim that our framework consistently unifies text and cross-modal modeling.

Method	Japanese VQA	FM-IQA
UC^2∗	29.57	30.09
Ours	**32.21**	**34.31**

Table 4: Fine-tuning accuracies on multilingual VQA tasks, UC^2∗ denotes our re-implementation of UC^2.

Cross-Lingual Fine-tuning As shown in the fine-tuning results on retrieval in Table 3 and VQA in Table 4, our model improves the performances of previous methods under different settings. Meanwhile, we notice that the retrieval performance of M^{3}P and UC^2 varies across languages while ours achieves a balanced performance. It indicates that our pre-trained model is a better initialization for downstream V-L tasks in multiple languages.

4.5 Ablation Study

To validate the effects of different components, we conduct ablation studies. All variants compared in this section are only pre-trained for 120K steps to save resources. Results are provided in Table 5.

Effects of TLM It is shown that TLM mainly helps the learning of shared interaction for all languages. Removing TLM significantly degrades the performance on tasks requiring inferring the image-text relationship, namely VNLI and retrieval, it conforms to results in Section 4.4 and further validates the consistency between TLM and VLM.

Effects of xMLM We apply xMLM to help learn universal representations for various languages, the results show that xMLM contributes to all tasks but the effect is not significant. We think the effect of xMLM is weakened by the strong XLM-R initialization and provide further analysis in Ap-

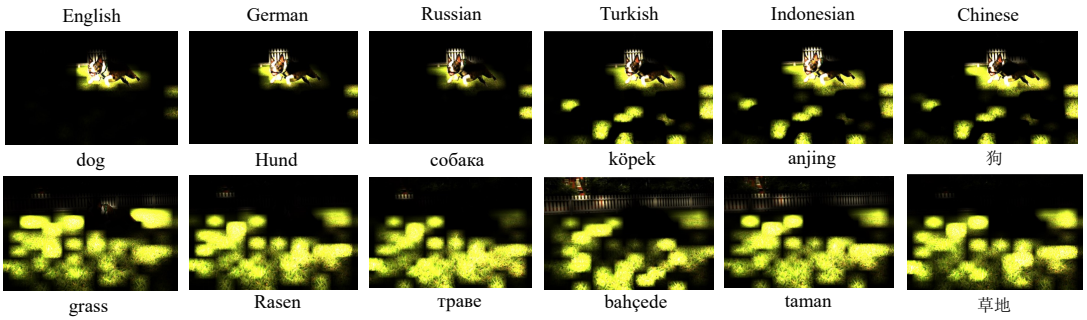

Figure 3: Visualization of the learned cross-attention between image regions and words in different languages. The model is not fine-tuned and the sample comes from the test set of xFlickr&CO that has not been seen by the model.

Method	XVNLI	xGQA	xFlickr&CO IR	xFlickr&CO TR	Multi30K All-Lang
Ours	**67.9**	**42.1**	**58.6**	**57.7**	**91.9**
w/o TLM	64.3	41.9	54.4	53.4	91.6
w/o xMLM	67.3	41.6	58.2	**57.7**	91.6
w/o XLC	67.1	41.0	51.7	50.8	91.5
w/o uni-arch	65.5	40.5	49.1	49.6	88.2

Table 5: Results of ablation studies. For ablated sub-modules, XLC is short for cross-lingual contrastive, and uni-arch denotes the unified architecture.

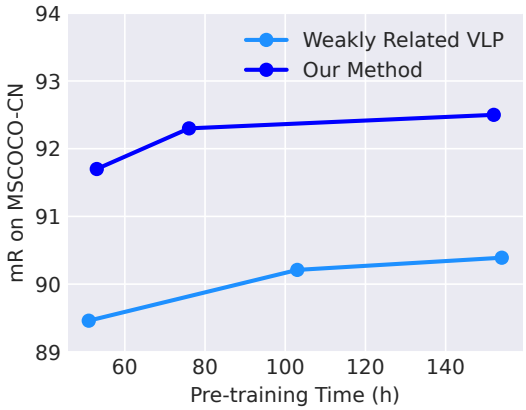

Figure 4: Retrieval performance under different pre-training costs which are controlled by the amounts of pre-training data.

pendix C.1. Another benefit of xMLM is to achieve a balance between languages. For low-resource bn, xMLM helps improve the accuracy from 27.8 to 33.8 in xGQA. This result is consistent with XLM-R (Conneau et al., 2020), it is mainly due to the balanced language distribution in D_m.

Effects of XLC As we argued, cross-lingual contrastive learning (XLC) explicitly guides the alignment among languages, which endows the high-level attention layers with the ability to be transferred across languages. Therefore, when removing XLC, the cross-lingual transfer performance degrades significantly. The results verify the effectiveness of XLC.

Effects of the Unified Architecture We ablate the introduced flexible architecture by removing the pluggable cross-attention layers in the high-level encoder. The entangled model does not perform well on all tasks, supporting our claim that different desirable interactions may compete in self-attention layers and thus hinder the unification.

4.6 Discussion

Results on Weakly Associated Data Another solution for scalable multilingual VLP is to relax the tight association requirement between image-text pairs. As noisy pairs can be crawled from the web (Radford et al., 2021), noisy multilingual pairs can be collected efficiently. We explore this idea with a noisy image-text corpus in Chinese (Gu et al., 2022). Following (Schuhmann et al., 2021), we employ a pre-trained model to filter out those pairs with a similarity lower than 0.25. Then the weakly related data is utilized for regular VLP with objectives used in (Li et al., 2021a).

As the result shown in Figure 4, more pre-training time and data yield better performance of both methods. At the same time, our method is more effective at the same cost. Considering that Wang et al. (2021) utilizes billions of noisy pairs to achieve satisfactory results, we speculate that the noisy data needs to be further scaled up for reliable multilingual VLP. Generally, the result indicates the efficiency of our method.

Attention Visualization We further visualize the learned cross-attention in Figure 3. The attended regions are similar for salient words with similar meanings in different languages, illustrating that

the cross-modal interaction learned in English can be applied to other languages. As German and Russian are in the same language family as English, the learned attention is more effectively transferred.

5 Conclusion

In this paper, we explored weakly supervised multilingual VLP without multilingual image-text pairs. We proposed a flexible architecture and unified tasks to effectively unify cross-lingual modeling on multilingual texts and cross-modal modeling on English image-text pairs. Experimental results validate the effectiveness of our approach to learn universal multilingual multimodal representations.

Limitations

Despite promising, the current work still has limitations. First, the current model mainly focuses on understanding problems. The generation ability of our model has not yet been investigated. It is unclear whether our weakly supervised framework also fits generative models and transfers strong generation capability across languages. Secondly, the current work explores multilingual corpora and overlooks the domain gaps in existing image resources. As argued in (Liu et al., 2021), the visual appearances of objects are diverse across cultures. Bias naturally exists in the distribution of images in existing V-L corpora. To develop a truly generalized multilingual multimodal model, the gap between visual distributions in different cultures should be considered.

Ethics Statement

Although multilingual multimodal representation learning is a promising topic, it has not been studied systematically due to the lack of multilingual data. Our work provides a solution to extend the success of English-centric works to more languages without the need for multilingual image-text pairs. Our pre-trained model can serve as a tool for v-L research or application in other languages and cultures. We hope that our work will motivate multimodal research to develop more effective methods for learning V-L representations in other cultures, benefitting more people in the world.

Acknowledgement

This work is partially supported by Ministry of Science and Technology of China (No.2020AAA0106701). We would also like to thank Xiaoqiang Lin for help with data preparation, the anonymous reviewers for their constructive feedback.

References

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.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. 2021a. Beit: Bert pre-training of image transformers. In *International Conference on Learning Representations*.

Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, and Furu Wei. 2021b. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *arXiv preprint arXiv:2111.02358*.

Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. 2018. 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.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Emanuele Bugliarello, Fangyu Liu, Jonas Pfeiffer, Siva Reddy, Desmond Elliott, Edoardo Maria Ponti, and Ivan Vulić. 2022. Iglue: A benchmark for transfer learning across modalities, tasks, and languages. *arXiv preprint arXiv:2201.11732*.

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 *European conference on computer vision*, pages 104–120. Springer.

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

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk,

and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485.

Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. 2020. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703.

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

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. 2017. 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. Association for Computational Linguistics.

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

Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. 2015. Are you talking to a machine? dataset and methods for multilingual image question. *Advances in neural information processing systems*, 28.

Jiaxi Gu, Xiaojun Meng, Guansong Lu, Lu Hou, Minzhe Niu, Hang Xu, Xiaodan Liang, Wei Zhang, Xin Jiang, and Chunjing Xu. 2022. Wukong: 100 million large-scale chinese cross-modal pre-training dataset and a foundation framework. *arXiv preprint arXiv:2202.06767*.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009.

Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. 2019. Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks. 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 2485–2494.

Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. 2020. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*.

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

Aashi Jain, Mandy Guo, Krishna Srinivasan, Ting Chen, Sneha Kudugunta, Chao Jia, Yinfei Yang, and Jason Baldridge. 2021. Mural: Multimodal, multitask representations across languages. In *Findings of the Association for computational Linguistics: EMNLP 2021*, pages 3449–3463.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021a. Align before fuse: Vision and language representation learning with momentum distillation.

Advances in Neural Information Processing Systems, 34.

Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. 2021b. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2592–2607.

Xirong Li, Chaoxi Xu, Xiaoxu Wang, Weiyu Lan, Zhengxiong Jia, Gang Yang, and Jieping Xu. 2019. Coco-cn for cross-lingual image tagging, captioning, and retrieval. *IEEE Transactions on Multimedia*, 21(9):2347–2360.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 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*, pages 10467–10485.

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 arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

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

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3977–3986.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Jonas Pfeiffer, Gregor Geigle, Aishwarya Kamath, Jan-Martin Steitz, Stefan Roth, Ivan Vulić, and Iryna Gurevych. 2022. xgqa: Cross-lingual visual question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2497–2511.

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

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361.

Bin Shan, Yaqian Han, Weichong Yin, Shuohuan Wang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2022. Ernie-unix2: A unified cross-lingual cross-modal framework for understanding and generation. *arXiv preprint arXiv:2211.04861*.

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.

Nobuyuki Shimizu, Na Rong, and Takashi Miyazaki. 2018. Visual question answering dataset for bilingual image understanding: A study of cross-lingual transfer using attention maps. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1918–1928.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2443–2449.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*.

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 Annual Meeting of the 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)*, pages 5100–5111.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022a. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. 2022b. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2021. Simvlm: Simple visual language model pretraining with weak supervision. In *International Conference on Learning Representations*.

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. 2019. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*.

Yuya Yoshikawa, Yutaro Shigeto, and Akikazu Takeuchi. 2017. Stair captions: Constructing a large-scale japanese image caption dataset. *arXiv preprint arXiv:1705.00823*.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67–78.

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

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4155–4165.

A Additional Discussion on Related Works

Noticing that M^{3}P also utilizes multilingual texts during pre-training, we comprehensively distinguish between our approach with M^{3}P from the following perspectives: (1) A code-switch-based method is proposed in M^{3}P to further generate multilingual image-text pairs for training, so it is not a weakly-supervised method; (2) M^{3}P simply feeds mixed data streams of multilingual texts and image-text pairs to a vanilla Transformer for joint pre-training. Referring to the results in IGLUE, M^{3}P does not enable effective cross-lingual transfer on downstream tasks, which means that M^{3}P struggles to learn universal representations across languages. Therefore, we propose an appropriate framework that unifies cross-lingual and cross-modal modeling, which is our main contribution.

ERNIE-Unix2 (Shan et al., 2022) is a concurrent work to ours. ERNIE-Unix2 aims to unify understanding and generation in multilingual VLP. To achieve this, ERNIE-Unix2 extends the idea of UC^2 (Zhou et al., 2021) to generate and collect more multilingual pairs, the process introduces an additional cost to scale up. Results of ERNIE-Unix2 are not included and compared in our main experiments since much more data is used, ERNIE-Unix2 consumes 89M multilingual image-text pairs during pre-training. Notice that our method demonstrates commendable performance in the context of an unfair setting.

B Additional Implementation Details

B.1 Pre-training Details

B.1.1 Language Distribution

We list the distribution of all languages $\{l_i\}_{i=1}^N \cup \{\text{en}\}$ considered in our model in Table 6. We use all data of target languages in WikiMatrix (Schwenk et al., 2021), we further transform traditional Chinese sentences into simplified Chinese which is more commonly used in China. For CC-100, we sub-sample 0.8B sentences following the language distribution used in XLM-R (Conneau et al., 2020):

$$q_i = \frac{p_i^\alpha}{\sum_{j=1}^N p_j^\alpha} \quad \text{with} \quad p_i = \frac{n_i}{\sum_{k=1}^N n_k}$$

where n_i is the number of sentences in l_i in the full dataset. α is set to 0.3 for a balanced distribution.

Language			Distribution	
Name	Code	Family	WikiMatrix	CC-100
English	en	Indo-E	1.000	0.085
Arabic	ar	Afro-A	0.051	0.037
Bengalu	bn	Indo-E	0.014	0.029
Bulgarian	bg	Indo-E	0.019	0.045
Czech	cs	Indo-E	0.027	0.037
Danish	da	Indo-E	0.022	0.050
Estonian	et	Uralic	0.013	0.027
German	de	Indo-E	0.077	0.053
Greek	el	Indo-E	0.032	0.043
French	fr	Indo-E	0.139	0.053
Indonesian	id	Austron	0.051	0.070
Japanese	ja	Japonic	0.044	0.054
Korean	ko	Koreanic	0.015	0.052
Chinese	zh	Sino-T	0.041	0.043
Potuguese	pt	Indo-E	0.122	0.050
Russian	ru	Indo-E	0.084	0.066
Spanish	es	Indo-E	0.165	0.051
Swahili	sw	Niger-C	0.003	0.019
Tamil	ta	Dravidian	0.003	0.031
Turkish	tr	Turkic	0.024	0.037
Vietnamese	vi	Austro-A	0.053	0.069

Table 6: Language distribution in the pre-training corpus. Language codes are based on ISO 639-1. A, C, E, and T are short for Asiatic, Congo, European, and Tibetan respectively. Austron denotes Austronesian.

B.1.2 Implementation of TLM

As illustrated in Figure 3.2 and Figure 5, our TLM task is slightly different from the original TLM task introduced in (Conneau and Lample, 2019), we activate the cross-attention layers to perform cross-sentence modeling. If the original TLM is applied, the cross-attention layers will only accommodate English inputs through VLM, our design allows cross-attention layers to adapt to non-English languages, and its effectiveness is demonstrated in Section 4.5.

B.1.3 Hyper-Parameters

For the model size, we follow the base-setting in BERT (Devlin et al., 2019): the hidden size is 768, the intermediate size is 3072, and the number of attention heads is 12. Our model consists of around 377M parameters in which the word embeddings of the large vocabulary take 200M parameters. During pre-training, the image resolution is 256×256 and the patch size is 16×16, and RandAugment (Cubuk et al., 2020) is applied to images. To avoid overfitting, dropout is applied with 0.1 probability, and 0.2 weight decay is used in the optimizer. The maximal lengths of sentences in D_m, D_v, and D_t are respectively 64, 35, and 50.

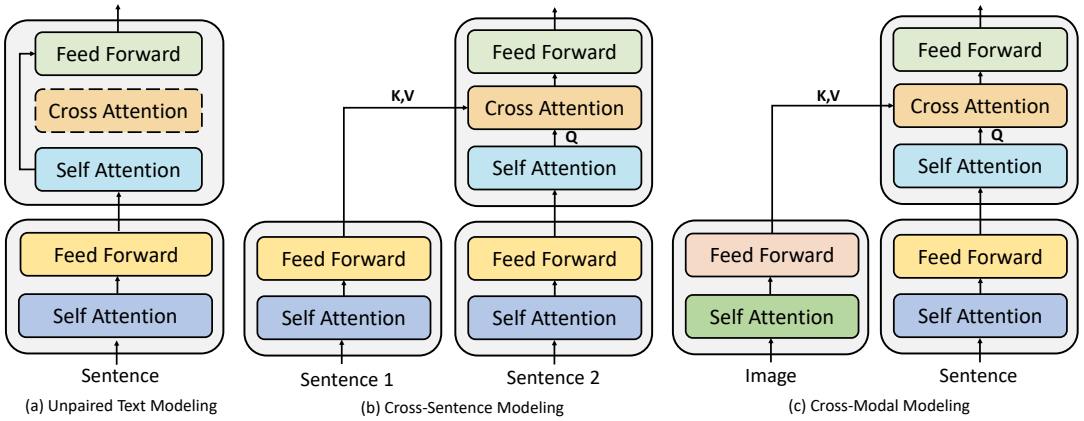

Figure 5: Illustration of the inference procedure for different data streams, the colors correspond to Figure 2.

B.2 Details of Fine-tuning

Due to the limitation of the text length, supplementary details of the fine-tuning experiments on different datasets are provided in this section.

B.2.1 Datasets

XVNLI is introduced in (Bugliarello et al., 2022), which is the multilingual extension of SNLI-VE (Xie et al., 2019). This task requires the model to infer the relationship between image-text pairs, the candidate answers include 'entailment', 'contradiction', and 'neutral'.

xGQA is introduced in (Pfeiffer et al., 2022), they extends the evaluation data of GQA (Hudson and Manning, 2019) dataset with manually translated questions in other 7 languages. The balanced English training set is used for training.

MaRVL is the **M**ulticultural **R**easoning over **V**ision and **L**anguage dataset introduced in (Liu et al., 2021). It can be regarded as a multicultural extension of the English NLVR2 dataset (Suhr et al., 2019). Each description is accompanied by 2 images, the model is asked to distinguish if the description is true for these 2 images. Different from xGQA and XVNLI, MaRVL address the problem of the gap between cultures by employing native speakers to collect images and descriptions which are representative in different cultures. The English training data comes from NLVR2.

xFlickr&CO is also created by IGLUE, they create a new multilingual evaluation set on 1000 images from Flickr30K (Young et al., 2014) and 1000 images from MSCOCO (Lin et al., 2014). They ask annotators to directly describe the images rather than translate the English captions. The English training set is constructed by sampling from the training sets of Flickr30K and MSCOCO.

WIT is short for the **W**ikipedia-based **I**mage **T**ext dataset (Srinivasan et al., 2021). They collect image-text pairs from Wikipedia in 108 languages. Compared to Flickr30K and MSCOCO, the relationship between the image-text pairs in WIT is relatively weaker and covers a diverse set of concepts. They create an English training set of 500K captions and evaluation sets in 10 languages where each language has at least 500 image-text pairs.

The datasets mentioned above are integrated into the IGLUE benchmark (Bugliarello et al., 2022), please refer to the original paper for more statistics.

Multi30K and MSCOCO Multi30K is based on the English Flickr30K dataset (Young et al., 2014). Several works (Elliott et al., 2016, 2017; Barrault et al., 2018) translate English captions into other languages. An image is paired with 5 captions in English and German, and 1 caption in French and Czech. The dataset is split into 29000/1000/1000 images for the train/val/test sets.

The original MSCOCO dataset is made of 123K images where 5 captions are used to describe an image. STAIR dataset (Yoshikawa et al., 2017) collects 820K Japanese captions for 165K images in COCO, for these 2 datasets, we use the standard Karpath split (Karpathy and Fei-Fei, 2015). COCO-CN (Li et al., 2019) is the Chinese counterpart, we use the human-written part of 20K images with around 1 caption per image and follow their split.

Japanese VQA and FM-IQA Both datasets are created based on the VQA task, which requires the model to answer a question conditioned on the visual content. Japanese VQA (Shimizu et al., 2018)

Task	XVNLI	xGQA	xFLickr&CO	WIT	Multi30K	MACOCO	FM-IQA	Ja-VQA
Peak learning rate	2e-5	3e-5	2e-5	2e-5	2e-5	4e-5	3e-5	3e-5
Epochs	10	10	10	10	15	15	15	15
Batch size	512	512	96	96	96	96	256	256
Max text length	40	40	50	50	50	50	40	40
Re-rank candidates	NA	NA	16	16	32	128	NA	NA
Frozen modules	uni-modal	uni-modal	None	None	None	None	None	None

Table 7: Fine-tuning hyper-parameters of experiments in different datasets.

Task	MaRVL	XNLI	SNLI→ XVNLI
Peak learning rate	4e-5	4e-5	4e-5
Epochs	10	10	10
Batch size	256	1024	1024
Max text length	40	50	50
Re-rank candidates	NA	NA	NA
Frozen modules	uni-modal	uni-modal	uni-modal

Table 8: Additional fine-tuning hyper-parameters.

Method	XVNLI	xGQA	xFlickr&CO IR	xFlickr&CO TR
Ours (XLM-R init)	67.9	42.1	58.6	57.7
w/o xMLM	67.3	41.6	58.2	57.8
Ours (ALBEF init)	65.1	37.3	57.6	56.0
w/o xMLM	63.8	34.3	56.7	55.6

Table 9: Results of ablation studies on xMLM for models with different initializations.

use images from Visual Genome and FM-IQA (Gao et al., 2015) provide Chinese questions for COCO images. Both datasets use natural sentences to answer the question and do not provide simplified answers like GQA (Hudson and Manning, 2019).

XNLI is a multilingual extension (Conneau et al., 2018) of the natural language inference (NLI) task. Sentence pairs are used as input, our model is required to infer the relationship between the premise and hypothesis. XNLI covers 15 languages while we only consider languages included in XVNLI.

B.2.2 General Setup

Inference The inference procedure is illustrated in Figure 5. Data are first encoded by uni-modal encoders, the cross-attention layers in the high-level encoder are skipped for unpaired text modeling. For paired inputs, cross-attention is activated for cross-modal or cross-sentence modeling.

Hyper-parameters The setup of several hyper-parameters is shared by all tasks. The image resolution is 384×384 and the patch size is 16×16, and the new visual position embedding is initialized with 2D interpolation following (Dosovitskiy et al., 2020), RandAugment (Cubuk et al., 2020) is also applied. All tasks are optimized by an AdamW optimizer with 0.2 weight decay. No warming-up is considered and the learning rates always linearly decay to zero. During fine-tuning, we may freeze the uni-modal encoders of our model to ensure the aligned multilingual multimodal semantic space is not influenced by English training data. We list the frozen parts for different tasks in Table 7.

Evaluation Metrics As for the metrics reported in this paper, we report the single-run results for two reasons: the pre-training procedure is costly, and as the pre-trained model provides a good initialization, we find that there is little variation in the fine-tuning results of different runs.

B.2.3 Task-Specific Setup

Retrieval For the retrieval task, we employ the pre-ranking and re-ranking mechanism as in (Li et al., 2021a). Pre-ranking similarities are computed by uni-modal encoders and re-ranking similarities come from the ITM head of the high-level encoder. We list the numbers of candidates for re-ranking in Table 7.

VQA We consider VQA as a classification task, we create the answer set with the N_a labels with the highest frequency in the training set. In xGQA, $N_a =$. We add dataset-specific MLPs on top of the high-level encoder.

MaRVL As each sample consists of 2 images, we first use the full image-grounded encoder to encode 2 image-text pairs, then the global representations of 2 pairs are concatenated and fed to an MLP to predict the score for true description.

NLI and VNLI Both tasks are 3-way classification. For NLI, our model encodes the sentence pairs in the same way as translation pairs, the encoded premise serves as the condition. For VNLI,

Model	VNLI	VQA	Reasoning	Retrieval			
	XVNLI	xGQA	MaRVL	xFlickr&CO IR	xFlickr&CO TR	WIT IR	WIT TR
zero-shot							
UC^2	62.1	29.4	57.3	20.3	17.9	7.8	9.1
M^{3}P	58.3	28.2	56.0	12.9	11.9	8.1	10.0
Ours	**69.5**	**42.1**	**62.1**	**59.8**	**58.7**	**36.3**	**36.6**
translate-test							
UC^2	73.7	50.2	63.1	36.0	30.4	12.7	14.1
M^{3}P	73.4	48.8	62.5	27.7	21.3	11.5	13.6
Ours	**75.5**	**52.5**	**71.1**	**79.1**	**77.6**	**46.6**	**46.8**

Table 10: Zero-shot cross-lingual transfer results in IGLUE. The models are trained in English and evaluated in target languages, the results are averaged over all target languages.

the encoded image is conditional information. The 2 tasks share the same architecture which enables us to test the cross-modal transfer capability.

C Additional Results and Analysis

In this section, we list more results of the main and supplementary experiments. Some complementary analysis is also provided.

C.1 Effects of Initialization

As we propose to perform weakly-supervised multilingual VLP by jointly learning cross-lingual text modeling and cross-modal modeling, the initialization model can provide strong capability in one of the 2 aspects. In the main paper, we use XLM-R (Conneau et al., 2020) for the cross-lingual modeling capability. In this section, we explore utilizing the text encoder of ALBEF (Li et al., 2021a) for the cross-modal modeling capability. The results are listed in Table 9.

It is obvious that the XLM-R initialization is better than the ALBEF initialization for cross-lingual V-L modeling. We think there are several factors that lead to the result. Firstly, XLM-R is a better-trained model that requires much more pre-training cost than ALBEF, this is a common phenomenon of the comparison between VLP and text-only pre-training. Secondly, in our framework, we just perform a relatively small-scale multilingual text pre-training in terms of the scale of data and maximal sequence lengths. The xMLM task in our method can not help the model to be comparable with XLM-R for universal multilingual text modeling.

At the same time, we find that xMLM is much more important for the ALBEF-initialized model to achieve universal cross-lingual representations. As the XLM-R initialization naturally implies strong a cross-lingual modeling capability. The effect of xMLM in Table 5 may be weakened.

Model	Language ar	es	fr	ru	mean
Compared models					
UC^2	56.2	57.5	69.7	64.9	62.1
M^{3}P	55.3	58.9	56.4	62.5	58.3
Ours	**66.3**	**69.5**	**71.7**	**70.4**	**69.5**
Ablation study					
Ours	**62.9**	69.7	**70.8**	**68.1**	**67.9**
w/o TLM	60.4	65.7	66.1	65.1	64.3
w/o xMLM	61.2	**70.7**	**70.8**	66.3	67.3
w/o XLC	61.0	69.6	70.0	68.0	67.1
w/o uni-arch	59.5	68.5	68.7	65.3	65.5

Table 11: Language-specific results of cross-lingual zero-shot transfer experiments in XVNLI.

C.2 Translation-Test Baselines

Following previous works on cross-lingual transfer (Conneau et al., 2018; Conneau and Lample, 2019; Conneau et al., 2020; Bugliarello et al., 2022), there are strong baseline models to utilize translation engines to perform translate-test: the test sets in other languages are translated to English and evaluated. Generally, these baselines are really competitive due to the strong translation engines. We provide the results in Table 10.

We can see that the translate-test baseline mod-

Model	Language							mean
	bn	de	id	ko	pt	ru	zh	
Compared models								
UC^2	20.0	42.9	28.7	21.4	30.4	31.0	31.2	29.4
M^{3}P	18.6	33.4	32.5	25.1	31.4	27.5	28.7	28.2
Ours	**31.9**	**48.7**	**45.3**	**39.1**	**47.0**	**39.0**	**43.4**	**42.1**
Ablation study								
Ours	**33.8**	47.1	**45.2**	38.6	**47.3**	40.5	42.5	**42.1**
w/o TLM	33.6	46.4	43.3	39.1	45.8	**42.8**	42.4	41.9
w/o xMLM	27.8	**47.2**	45.1	**39.4**	47.1	41.6	**43.2**	41.6
w/o XLC	33.6	46.9	44.3	37.6	45.1	36.5	43.0	41.0
w/o uni-arch	31.5	45.4	42.1	37.7	43.6	38.0	41.5	40.0

Table 12: Language-specific results of cross-lingual zero-shot transfer experiments in xGQA.

Model	Language					mean
	id	sw	ta	tr	zh	
Compared models						
UC^2	56.7	52.6	**60.5**	56.7	59.9	57.3
M^{3}P	56.5	55.7	56.0	56.8	55.0	56.0
Ours	**65.3**	**58.7**	60.3	**65.3**	**60.6**	**62.1**

Table 13: Language-specific results of cross-lingual zero-shot transfer experiments in MaRVL.

els always perform better. At the same time, our method narrows the gap between the zero-shot and translate-test performance, which means that our method learns better universal multilingual multimodal representations. This result conforms with the main results in Section 4.4.

C.3 Language-Specific Results of IGLUE

In this section, we provide the experimental results of IGLUE in all languages separately. Results of XVNLI, xGQA, MaRVL, xFlickr&CO, and WIT are respectively listed in Table 11, 12, 13, 14, 15. The results of the ablated variants are also included for XVNLI, xGQA, and xFlickr&CO.

C.3.1 Supplementary Analysis

Generally, our method outperforms M^{3}P and UC^2 across languages and tasks in different datasets except ta in MaRVL. At the same time, we notice that the UC^2 is skewed towards the languages (de, fr, cs, zh, ja) in which the translated image captions are generated. M^{3}P does not perform well in low-resource languages like bn. However, our model achieves a more balanced performance among various languages.

xMLM is the main factor of balanced performance. The balanced language distribution of D_m helps the learning of low-resource languages bn (in Table 12) and minority languages in WikiMatrix like ar and ru (in Table 11).

Model	Language														mean	
	de		es		id		ja		ru		tr		zh			
	IR	TR	IR	TR	IR	TR	IR	TR	IR	TR	IR	TR	IR	TR	IR	TR
Compared models																
UC^2	28.6	23.9	16.0	15.3	14.6	13.6	24.3	22.4	20.0	16.8	7.2	7.0	31.6	26.3	20.3	17.9
M^{3}P	13.4	11.9	13.4	12.2	13.2	12.1	10.3	9.7	16.0	14.5	7.8	8.4	16.5	14.8	12.9	11.9
Ours	**58.2**	**57.2**	**69.6**	**68.7**	**62.7**	**60.6**	**49.8**	**48.2**	**63.2**	**62.6**	**50.8**	**50.8**	**64.2**	**63.2**	**59.8**	**58.7**
Ablation Study																
Ours	**58.2**	**56.8**	**67.6**	**67.5**	**61.3**	**60.1**	48.5	46.1	62.5	60.5	**48.1**	**50.5**	**64.1**	**62.5**	**58.6**	**57.7**
w/o xMLM	56.6	55.7	66.4	66	59.4	59.4	**51.8**	**52.3**	**62.8**	**62.6**	47.4	47.0	63.1	62.2	58.2	**57.7**
w/o TLM	55.6	54.1	62.2	63.0	57.1	55.7	44.5	40.0	55.6	56.2	45.9	45.1	59.8	59.9	54.4	53.4
w/o XLC	53.1	51.9	61.7	60.5	53.8	52	42	41.5	55.4	54.0	39.4	40.0	56.8	55.8	51.7	50.8
w/o uni-arch	49.0	48.8	58.4	57.4	50.5	51.6	42.3	43.1	48.1	50.6	41.5	40.0	53.9	55.7	49.1	49.6

Table 14: Language-specific results of cross-lingual zero-shot transfer experiments in xFlickr&CO.

Model	Language										mean
	ar	bg	da	el	et	id	ja	ko	tr	vi	
Image Retrieval											
UC^2	6.6	8.8	9.4	8.8	4.7	9.9	9.8	4.3	7.5	8.5	7.8
M^{3}P	8.9	8.8	9.4	9.7	5.4	8.7	7.0	6.1	6.5	10.8	8.1
Ours	**37.3**	**30.8**	**41.8**	**37.7**	**26.5**	**47.1**	**31.9**	**25.6**	**36.1**	**48.1**	**36.3**
Text Retrieval											
UC^2	8.3	7.7	10.4	11.6	6.0	11.5	10.8	5.7	8.8	9.9	9.1
M^{3}P	8.3	9.8	11.8	12.0	8.2	10.9	8.4	7.1	10.6	12.7	10.0
Ours	**37.8**	**31.4**	**40.7**	**37.3**	**26.9**	**44.0**	**33.3**	**26.0**	**40.8**	**47.3**	**36.6**

Table 15: Language-specific results of cross-lingual zero-shot transfer experiments in WIT. Results are recalls for image retrieval and text retrieval

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
In the "Limitations" section.

☒ A2. Did you discuss any potential risks of your work?
We only adopt publicly open resources including data and packages. Those resources are commonly used in corresponding domains. Our work does not introduce additional risk.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
In the abstract and section 1.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Yes, we propose a framework in Section 3.

☑ B1. Did you cite the creators of artifacts you used?
We cite the artifacts used in Sections 4.1, 4.2, and Appendix A.1, A.2.

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
We use multiple existing open-source artifacts that are based on different licenses, making it difficult to summarize. We cite the resources of utilized artifacts where the license details can be found.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
In Sections 1, 3, 4.1 and Appendix A.2.2.

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
We only adopt publicly open datasets. Those resources are commonly used in corresponding domains. And the information security issues have been discussed in the papers where the datasets are introduced.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
In Appendix A.1.1.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
In Section 4.1 and Appendix A.1.1, A.2.1.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

C ☑ Did you run computational experiments?

In Section 4.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
In Section 4.1 and Appendix A.1.2.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
In Section 4.1 and Appendix A.1.2, A.2.2, and A.2.3.

☐ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
No response.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
In Section 4.1 and Appendix A.1.2, A.2.2, and A.2.3.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.