

# M<sup>3</sup>P: Learning Universal Representations via Multitask Multilingual Multimodal Pre-training

Minheng Ni<sup>1\*†</sup> Haoyang Huang<sup>2†</sup> Lin Su<sup>3†</sup> Edward Cui<sup>3</sup> Taroon Bharti<sup>3</sup> Lijuan Wang<sup>4</sup>  
Dongdong Zhang<sup>2</sup> Nan Duan<sup>2‡</sup>

<sup>1</sup> Research Center for Social Computing and Information Retrieval  
Harbin Institute of Technology, China

<sup>2</sup> Natural Language Computing, Microsoft Research Asia, China

<sup>3</sup> Bing Multimedia Team, Microsoft, China

<sup>4</sup> Cloud+AI, Microsoft, United States

mhni@ir.hit.edu.cn

{haohua, lins, edwac, tbharti, lijuanw, Dongdong.Zhang, nanduan}@microsoft.com

## Abstract

We present **M<sup>3</sup>P**, a **Multitask Multilingual Multimodal Pre-trained model** that combines multilingual pre-training and multimodal pre-training into a unified framework via multitask pre-training. Our goal is to learn universal representations that can map objects occurred in different modalities or texts expressed in different languages into a common semantic space. In addition, to explicitly encourage fine-grained alignment between images and non-English languages, we also propose **Multimodal Code-switched Training (MCT)** to combine monolingual pre-training and multimodal pre-training via a code-switch strategy. Experiments are performed on the multilingual image retrieval task across two benchmark datasets, including MSCOCO and Multi30K. **M<sup>3</sup>P** can achieve comparable results for English and new state-of-the-art results for non-English languages.

## 1. Introduction

Recently, we witness the rise of a new paradigm of natural language processing (NLP), where general knowledge is learned from raw texts by self-supervised pre-training and then applied to downstream tasks by task-specific fine-tuning. Now, these state-of-the-art monolingual pre-trained language models, such as BERT [7], RoBERTa [23] and GPT-2 [28], have been expanded to *multilingual scenarios*, such as Multilingual BERT [7], XLM/XLM-R [5, 4], Unicoder [13]. Moreover, some pre-training models under *multimodal sce-*

*narios*, such as Unicoder-VL [19], UNITER [3], ERNIE-ViL [36], VILLA [10] and Oscar [21], also come out.

However, it is still challenging to extend these pre-trained models to multilingual-multimodal scenarios. The multilingual pre-trained language models cannot handle vision data (e.g., images or videos) directly, whereas many pre-trained multimodal models are trained on English corpora thus cannot perform very well on non-English languages. Therefore, high quality multilingual multimodal training corpus is essential to combine multilingual pre-training and multimodal pre-training. However, there are only a few multilingual multimodal corpora exist, and they also have low language coverage. Moreover, relying on high-quality machine translation engines to generate such data from English multimodal corpora is both time-consuming and computationally expensive. Learning explicit alignments between vision and non-English languages during pre-training is lacking.

To address these challenges, this paper presents **M<sup>3</sup>P**, a **Multitask Multilingual Multimodal Pre-trained model**, which aims to learn universal representations that can map objects occurred in different modalities or texts expressed in different languages into a common semantic space. In order to alleviate the issue of lacking enough non-English labeled data for multimodal pre-training, we introduce *Multimodal Code-switched Training (MCT)* to enforce the explicit alignments between images and non-English languages. The goal is achieved by (i) learning to represent multilingual data using multilingual corpora (e.g., sentences from Wikipedia covering 100 languages) by multilingual pre-training, (ii) learning multilingual-multimodal representations by randomly replacing some English words with their translations in other languages from multimodal corpora (e.g., image-caption pairs labeled in English), and (iii) generalizing these rep-

\*Work is done during an internship at Microsoft Research Asia.

†These authors contributed equally to this work.

‡Corresponding Author.

representations to deal with multilingual-multimodal tasks by Multitask learning.

In summary, the main contributions of this paper are:

- We present M<sup>3</sup>P, the first known effort on combining multilingual pre-training and multimodal pre-training into a unified framework.
- We propose a novel *Multimodal Code-switched Training* (MCT) method, an effective way to enhance the multilingual transfer ability of M<sup>3</sup>P in the zero-shot and few-shot settings.
- We achieve new state-of-the-art results for the multilingual image-text retrieval task on both Multi30K and MSCOCO for non-English languages, outperforming existing multilingual methods by a large margin. The proposed model can also achieve comparable results for English on these two datasets, compared to the state-of-the-art monolingual multimodal models.
- Last but not least, we conduct extensive experiments and analysis to provide insights on the effectiveness of using *Multimodal Code-switched Training* (MCT) and each pre-training task.

## 2. Related Work

**Multilingual Pre-trained Models** Multilingual BERT (M-BERT) [7] demonstrates that by performing masked language modeling on a multilingual corpus with shared vocabulary and weights for 102 languages, surprisingly good results can be achieved on the cross-lingual natural language inference (XNLI) [6] task in 15 languages. XLM [5] and Unicoder [13] further improve the multilingual BERT by introducing new pre-training tasks based on a bilingual corpus. However, all such models work for NLP tasks and are not well designed for multimodal tasks such as Multilingual Image-text Retrieval or Multimodal Machine Translation.

**Multimodal Pre-trained Models** Recently, a large number of multimodal pre-trained models, such as ViLBERT [24], Unicoder-VL [19], UNITER [3], VLP [37] and Oscar [21], are developed for vision-language tasks using multi-layer Transformer as the backbone. However, as it is not easy to collect well-aligned visual-linguistic training data in multiple languages, all these models are pre-trained for English only based on monolingual multimodal corpora, such as Conceptual Captions [29], SBU Captions [26], Visual Genome [17] and MSCOCO [2]. Hence, it is not feasible to apply them into multimodal tasks with non-English inputs.

**Code-switched Training** Code-switched training [27] [33] converts the original training corpus to code-switched

corpus, which can help the model explicitly model the relationship among corresponding words in different languages. Existing work uses a rule-based word replacement strategy to replace the original word with translated word randomly by bilingual dictionaries. This approach provides a significant improvement to the low-resource language. However, existing works use Code-switching for text-only tasks and ignore its application on multimodal pre-training model under *multilingual-multimodal scenarios*.

## 3. M<sup>3</sup>P: Multitask Multilingual Multimodal Pre-training

In this section, we describe how we train M<sup>3</sup>P using a multilingual-monomodal corpus (e.g., sentences extracted from Wikipedia) and a monolingual-multimodal corpus (e.g., English image-caption pairs). As outlined in Figure 1, we use the self-attentive transformer architecture of BERT, and design two pre-training objectives with three types of data streams. Multitask training is employed into the pre-training stage to optimize all pre-training objectives simultaneously for better performance. We optimize the accumulated loss of both pre-training objectives with the same weight in each iteration to train them by turns.

### 3.1. Data Stream

We use two basic data streams, Multilingual Monomodal Stream and Monolingual Multimodal Stream, from the multilingual corpus and multimodal corpus, respectively. We also design Multimodal Code-switched Stream to utilize multilingual data and multimodal data at the same time. Details regarding the three data streams are introduced below.

**Multilingual Monomodal Stream** To apply multilingual pre-training, we use Multilingual Monomodal Stream as model input. Given an input text in any language  $w^{[l_i]}$ , we first tokenize it into a sequence of BPE tokens via Sentence Piece [18]. Then we can obtain a text representation sequence by summing up the text embedding and the position embedding of each BPE token. Moreover, a language embedding [5] is added to each token to indicate its language attribute. Specifically, the input data is defined as:

$$\{\mathbf{w}^{[l_i]}\} = \{(\mathbf{w}_1^{[l_i]}, \mathbf{w}_2^{[l_i]}, \dots, \mathbf{w}_M^{[l_i]})\}$$

where  $M$  denotes the length of  $w^{[l_i]}$  and  $l_i$  denotes a language in the language set  $L$ . We denote this stream as  $D^{[X]}$ .

**Monolingual Multimodal Stream** To apply multimodal pre-training, we use Monolingual Multimodal Stream as model input. Given a pair of English text and image  $(w^{[EN]}, v)$ , the text representation sequence of  $w^{[EN]}$  is obtained similarly as we described in Multilingual Monomodal

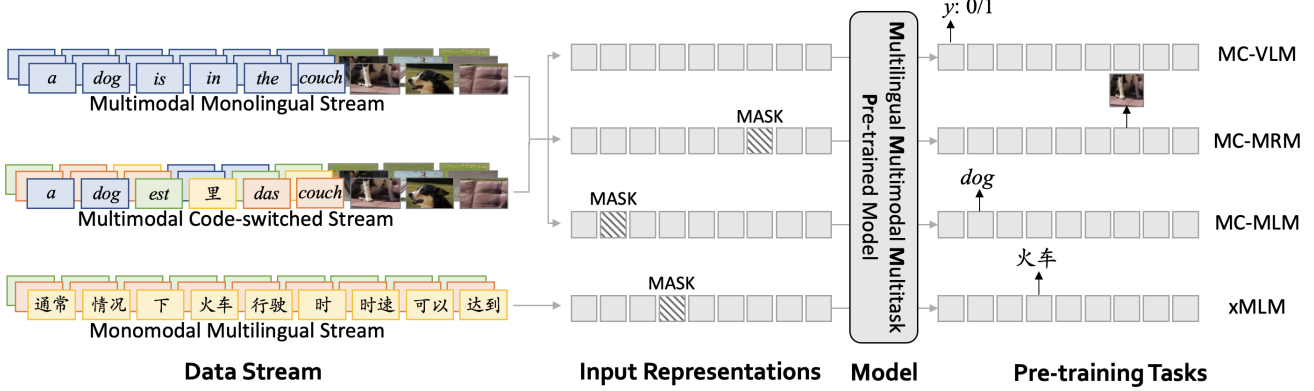


Figure 1: Three data streams and four pre-training tasks used in M<sup>3</sup>P. Blue blocks denote English text, and Yellow, Green and Orange blocks denote non-English text.

Stream section, where English is used as the language embedding. For the image  $v$ , we use Faster-RCNN [12] to detect image regions and use corresponding visual features in each region as a visual feature sequence. We also add a spatial embedding to each visual token, which is a 5-D vector based on its normalized top-left, bottom-right coordinates, and the fraction of the image area covered. We project these two vectors to the same dimension of the text representation using two fully-connected (FC) layers. Therefore, the image representation sequence is obtained by summing up its projected visual feature vector and spatial embedding vector of each region in the image. Furthermore, we add a stream tag [IMG] at the beginning of the image region sequence to separate text tokens and image tokens, and concatenate them to form an input stream:

$$\{\mathbf{w}^{[\text{EN}]}, \mathbf{v}\} = \{(\mathbf{w}_1^{[\text{EN}]}, \mathbf{w}_2^{[\text{EN}]}, \dots, \mathbf{w}_M^{[\text{EN}]}) , (\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_N)\}$$

We denote this stream as  $D^{[\text{EN}]}$ .

**Multimodal Code-switched Stream** We generate Multimodal Code-switched Stream from Monolingual Multimodal Stream by code-switched method, given English text and image pairs  $(w^{[\text{EN}]}, v)$ , the set of code-switched languages  $\mathbf{C} = \{c_1, c_2, \dots, c_k\}$ , and bilingual dictionaries which can translate a word from English to any language  $c_i$ . Following [27], for each word  $w_i^{[\text{EN}]}$  in English text  $w^{[\text{EN}]}$ , we replace it with a translated word with a probability of  $\beta$ . If a word has multiple translations, we choose a random one. Similar to the generation process of Multilingual Monolingual Stream, we obtain the text representation sequence of the Code-switched text  $w^{[\text{C}]}$  in the same way while keeping the original language embedding.\* Similar with Mono-

\*We have tried to change language embedding in Code-switched Stream, but no significant gain was obtained.

lingual Multimodal Stream, the text and image representation sequences are concatenated as the final input stream:  $\{(\mathbf{w}_1^{[d_1]}, \mathbf{w}_2^{[d_2]}, \dots, \mathbf{w}_M^{[d_M]}), (\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_N)\}$ , where  $d_i$  is a random language in  $\{\text{EN}\} \cup \mathbf{C}$ . We simplify the input sequence as:

$$\{\mathbf{w}^{[\text{C}]}, \mathbf{v}\} = \{(\mathbf{w}_1^{[\text{C}]}, \mathbf{w}_2^{[\text{C}]}, \dots, \mathbf{w}_M^{[\text{C}]}) , (\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_N)\}$$

We denote this stream as  $D^{[\text{C}]}$ .

### 3.2. Pre-training Objectives

To pre-train M<sup>3</sup>P under *multilingual-multimodal scenario*, we designed two types of pre-training objectives. *Multilingual Training* aims to learn grammar or syntax from well-formed multilingual sentences. *Multimodal Code-switched Training* (MCT) aims to learn different languages from the shared vision modal and the alignment between vision and non-English texts.

#### 3.2.1 Multilingual Training

**Multilingual Masked Language Modeling (xMLM)** Similar to Multilingual BERT [7], XLM [5] and Unicoder [13], this task performs masked language modeling based on the multilingual corpus. At each iteration, a batch is composed of sentences sampled from different languages. The sampling probability of a language  $l_i$  is defined as  $\lambda_{l_i} = p_{l_i}^\alpha / \sum_{l_i} p_{l_i}^\alpha$ , where  $p_{l_i}$  is the percentage of  $l_i$  in the entire multilingual corpus, and the smoothing factor  $\alpha$  is set to 0.3. In each batch, we randomly sample 15% of the words and (i) replace them with a special symbol [MASK], (ii) replace them with random tokens, or (iii) keep them unchanged with a probability of 80%, 10% and 10%, respectively. We only use Multilingual Monomodal Stream  $D^{[\text{X}]}$  for we do not need to use Code-switching to extend it to multilingual

corpus. The loss function is defined as:

$$\mathcal{L}_{\text{xMLM}}(\theta) = -\mathbb{E}_{\mathbf{w}^{[l_i]} \sim D^{[x]}} \log q_{\theta}(w_m^{[l_i]} | \mathbf{w}_{\setminus m}^{[l_i]})$$

, where  $w_m^{[l_i]}$  is the masked token and  $\mathbf{w}_{\setminus m}^{[l_i]}$  is its context.

### 3.2.2 Multimodal Code-switched Training

Because of the lack of labeled data for the non-English multimodal scenario, the model can only learn multilingualism and multimodality independently. To help the model learn different language representations under the shared vision modal, we propose three Multimodal Code-switched Training tasks: MC-MLM, MC-MRM and MC-VLM. We mix Multimodal Code-switched Stream  $D^{[C]}$  and Monolingual Multimodal Stream  $D^{[EN]}$  with a proportion ratio of  $\alpha$  and  $1 - \alpha$ , respectively, in train these tasks. To simplify the symbols, we denote the mixed data stream as  $D$  and omit the mask [EN] or [C] as  $[\cdot]$ .

**Multimodal Code-switched Masked Language Modeling (MC-MLM)** Different from the pre-training tasks in ViLBERT [24] and Unicoder-VL [19], this task aims to learn the representation of different languages based on the shared vision modal. Mixed data stream  $D$  is used for training this objective. Specifically, the model predicts each masked token  $w_m^{[\cdot]}$  in the caption  $\mathbf{w}^{[\cdot]}$  based on its surrounding tokens  $\mathbf{w}_{\setminus m}^{[\cdot]}$  and all image regions  $\mathbf{v}$ . We follow the same masking strategy used in xMLM to mask tokens in the input caption. The loss function is defined as:

$$\mathcal{L}_{\text{MC-MLM}}(\theta) = -\mathbb{E}_{(\mathbf{w}^{[\cdot]}, \mathbf{v}) \sim D} \log p_{\theta}(w_m^{[\cdot]} | \mathbf{w}_{\setminus m}^{[\cdot]}, \mathbf{v})$$

, where  $D$  is the mixed data stream.

**Multimodal Code-switched Masked Region Modeling (MC-MRM)** This task aims to learn vision representations with multilingual text as the context in mixed data stream  $D$ . The model reconstructs each masked image region  $v_n$  based on the remaining regions  $\mathbf{v}_{\setminus n}$  and all caption tokens  $\mathbf{w}^{[\cdot]}$ . We randomly mask image regions with a probability of 15%. The input representation of each masked image region is set to zeros or kept as the original values with a probability of 90% and 10%, respectively. We apply an FC layer to convert the Transformer output of each masked region  $v_k$  into a vector  $h_{\theta}(v_k)$  of the same dimension with the visual feature  $f(v_k)$ . We use cross-entropy loss  $\text{CE}(g_{\theta}(v_k), C(v_k))$  to predict the object category of each masked region  $v_k$ . We also apply another FC layer to convert the Transformer output of each masked region  $v_k$  to predict the scores of  $K$  object classes, which further go through a softmax function to be transformed into a normalized distribution  $g_{\theta}(v_k)$ . We take the predicted object category with the highest confidence

| Dataset                                  | Images | Texts | Languages |
|--|--------|-------|-----------|
| <i>Pre-training Corpus</i>               |        |       |           |
| Wikipedia                                | -      | 101G  | 100       |
| Conceptual Captions [29]                 | 3.3M   | 3.3M  | 1         |
| <i>Fine-tuning and Evaluation Corpus</i> |        |       |           |
| Multi30K [35]                            | 32K    | 384K  | 5         |
| MSCOCO [2] [34] [20]                     | 120K   | 1.5M  | 3         |

Table 1: Statistics of datasets.

score outputted by Faster-RCNN as the ground-truth label of  $v_k$ , and convert it into a one-hot vector  $C(v_k) \in \mathbb{R}^K$ . Due to the top-1 category predicted by Faster-RCNN is not always correct, we leave minimizing the KL divergence between two distributions for our future work. The loss function can be defined as:

$$\mathcal{L}_{\text{MC-MRM}}(\theta) = -\mathbb{E}_{(\mathbf{w}^{[\cdot]}, \mathbf{v}) \sim D} \sum_k [\text{MSE}(h_{\theta}(v_k), f(v_k)) + \text{CE}(g_{\theta}(v_k), C(v_k))]$$

where  $k$  enumerates the index of each masked image region and  $\text{MSE}(h_{\theta}(v_k), f(v_k))$  denotes the mean-square-error loss that regresses the Transformer output of each masked region  $v_k$  to its visual feature  $f(v_k)$ .

**Multimodal Code-switched Visual-Linguistic Matching (MC-VLM)** This task aims to learn alignment between multilingual texts and images with mixed data stream  $D$ . An FC layer  $s_{\theta}(\mathbf{w}^{[\cdot]}, \mathbf{v})$  is applied on the Transformer output of [CLS] to predict whether the input image  $\mathbf{v}$  and the input English or Code-switched text  $\mathbf{w}^{[\cdot]}$  are semantically matched. Negative image-caption pairs are created by replacing the image or text in a matched sample with a randomly-selected image or text from other samples. We use Binary Cross-Entropy as the loss function:

$$\mathcal{L}_{\text{MC-VLM}}(\theta) = -E_{(\mathbf{w}^{[\cdot]}, \mathbf{v}) \sim D} [\text{BCE}(s_{\theta}(\mathbf{w}^{[\cdot]}, \mathbf{v}), y)]$$

where  $y \in \{0, 1\}$  indicates whether the input image-text pair is matched and BCE indicates binary-cross-entropy loss.

## 4. Experiments

In this section, we describe detailed experimental settings during pre-training, fine-tuning and evaluating M<sup>3</sup>P model.

### 4.1. Dataset Description

As shown in Table 1, we construct our pre-training dataset based on multimodal corpus, Conceptual Captions [29], and multilingual corpus, Wikipedia. We evaluate M<sup>3</sup>P on multilingual image-text retrieval task on two datasets: Multi30K



[9, 8] and MSCOCO [2, 25, 20]. Panlex<sup>†</sup> is used as the bilingual dictionary during *Multimodal Code-switched Training*.

#### 4.1.1 Pre-training Corpus

**Conceptual Captions** We use Conceptual Captions [29] as the multimodal corpus. It contains 3.3 million English image-caption pairs harvested from the Web and does not contain any non-English text.

**Wikipedia** We use sentences extracted from the Wikipedia dump as the multilingual corpus. It includes 101G sentences covering 100 languages without any vision information.

#### 4.1.2 Fine-tuning and Evaluation Corpus

**Multi30K** This dataset extended Flickr30K [35] from English (en) to German (de), French (fr) and Czech (cs). It contains 31,783 images and provides five captions per image in English and German and one caption per image in French and Czech. The train, dev, and test splits are defined in [35].

**MSCOCO** This dataset contains 123,287 images and provides five captions per image in English (en), but fewer in Chinese (zh) and Japanese (ja). STAIR Captions [34] extended MSCOCO [2] with 820K Japanese captions for COCO images. [20] extended MSCOCO [2] with Chinese captions for 20K images. We use the same train, dev, and test splits for English and Japanese as defined in [14]. As for Chinese, we use the COCO-CN split [20].

#### 4.1.3 Code-switched Dictionary

The word-level bilingual dictionaries used by Code-switched training are from Panlex, the world’s largest open-source lexical translation database. We extract top 50 scale English to other language bilingual dictionaries.

### 4.2. Training Details

**Pre-training Details** Similar to previous vision-language pre-trained models, the M<sup>3</sup>P model uses the same model architecture as BERT [7]. We initialize M<sup>3</sup>P with XLM-R [4] and continue pre-training on our data. We use the same vocabulary as XLM-R [4], which includes 250K BPE tokens and covers 100 languages. We set the dropout rate to 0.1 and the max input length to 128. We use Adam Optimizer [16] with a linear warm-up [30] and set the learning rate to  $1 \times 10^{-4}$ . The total batch size is 1,024 after gradient accumulation. The pre-training stage takes about seven days to converge on 8 V100 GPUs. We use *Multimodal Code-switched Training* with all top 50 languages from Panlex.

<sup>†</sup><https://panlex.org>

**Fine-tuning Details** The batch size is set to 512, and we sample three negative cases for each positive case in VLM. We experiment with different numbers of negative samples in {1, 3, 5}, and find three yields the best results. We use Adam Optimizer with  $\beta_1 = 0.9$ ,  $\beta_2 = 0.98$  and  $5 \times 10^{-5}$  as the hyper-parameters of learning rate.

### 4.3. Baselines

We compare our work with several related work [15, 31, 11, 32, 1], which are trained on downstream task datasets (MSCOCO and Multi30K) directly without pre-training. In addition, to make the comparison as fair as possible, we take Unicoder-VL as another baseline, as it employs the same pre-training data during image-language pre-training.

Among the baselines, SMALR [1] uses machine translation to augment Multi30K and MSCOCO. But considering that applying machine translation to translate English to all other supported languages lacks generalization and requires a large amount of translators, we leave this as an option for future work. Moreover, note that MULE is using different dev/test splits of MSCOCO compared with other models.

It is also worth noticing that word-level dictionaries are only used in M<sup>3</sup>P, as the *Multimodal Code-switched Training* is firstly used in multilingual multimodal pre-training.

### 4.4. Evaluation Settings

Multilingual image-text retrieval is the task of finding the most relevant images given input texts in different languages, or vice versa. We use mean Recall (mR) as our metric, which is an averaged score of Recall@1, Recall@5, and Recall@10 on image-to-text retrieval and text-to-image retrieval tasks.

We compare M<sup>3</sup>P with baseline methods on multilingual image-text retrieval in four different settings:

- (i) *w/o fine-tune*: apply M<sup>3</sup>P to all test sets directly to obtain the evaluation results without fine-tuning.
- (ii) *w/ fine-tune on en*: fine-tune M<sup>3</sup>P on English and then apply the fine-tuned model to all test sets.
- (iii) *w/ fine-tune on each*: fine-tune M<sup>3</sup>P on each language and apply each model to the test set of this language.
- (iv) *w/ fine-tune on all*: fine-tune M<sup>3</sup>P for all languages using the merged labeled data and then apply the fine-tuned model to all test sets.

## 5. Results and Analysis

In this section, we show the evaluation results of M<sup>3</sup>P compared with existing work and conduct ablation studies in order to better understand the effect of the model.

### 5.1. Overall Results

From Table 2, we have several observations: (1) Our M<sup>3</sup>P model obtains the state-of-the-art results in all non-English languages, which shows its exciting multilingual multimodal

| Model  | Multi30K    |             |             |             | MSCOCO      |             |             |
|--|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
|  | en          | de          | fr          | cs          | en          | ja          | zh          |
| <i>Monolingual supervised results</i>                    |             |             |             |             |             |             |             |
| EmbN [31]  | 72.0        | 60.3        | 54.8        | 46.3        | 76.8        | 73.2        | 73.5        |
| PAR. EmbN [11]   | 69.0        | 62.6        | 60.6        | 54.1        | 78.3        | 76.0        | 74.8        |
| S-LIWE [32]  | 76.3        | 72.1        | 63.4        | 59.4        | 80.9        | 73.6        | 70.0        |
| MULE [15]  | 70.3        | 64.1        | 62.3        | 57.7        | 79.0        | 75.9        | 75.6        |
| SMALR [1]  | 74.5        | 69.8        | 65.9        | 64.8        | 81.5        | 77.5        | 76.7        |
| <i>Monolingual results with multimodal pre-training</i>  |             |             |             |             |             |             |             |
| Unicoder-VL (w/o fine-tune) [19]                         | 72.0        | -           | -           | -           | 63.7        | -           | -           |
| Unicoder-VL (w/ fine-tune on en) [19]                    | <b>88.1</b> | -           | -           | -           | <b>89.2</b> | -           | -           |
| <i>Multilingual results with multimodal pre-training</i> |             |             |             |             |             |             |             |
| M <sup>3</sup> P (w/o fine-tune)                         | 57.9        | 36.8        | 27.1        | 20.4        | 63.1        | 33.3        | 32.3        |
| M <sup>3</sup> P (w/ fine-tune on en)                    | 87.4        | 58.5        | 46.0        | 36.8        | 88.6        | 53.8        | 56.0        |
| M <sup>3</sup> P (w/ fine-tune on each)                  | 87.4        | 82.1        | 67.3        | 65.0        | 88.6        | 80.1        | 75.8        |
| M <sup>3</sup> P (w/ fine-tune on all)                   | 87.7        | <b>82.7</b> | <b>73.9</b> | <b>72.2</b> | 88.7        | <b>87.9</b> | <b>86.2</b> |

Table 2: Multilingual image-text retrieval results on Multi30K and MSCOCO. The metric is the mean Recall (mR). Each **bold number** indicates the best mR score in that column. We report the mR results of Unicoder-VL on the English test set, as it is pre-trained based on the same image-caption corpus (i.e., Conceptual Captions) with M<sup>3</sup>P.

transfer capability. (2) Similar to the observations reported in Unicoder [13, 22], the two *fully-supervised* settings (iii) *w/ fine-tune on each* and (iv) *w/ fine-tune on all* can lead to the best results. This means the same sentence in different languages may capture complementary information to help improve performance. (3) Comparing to Unicoder-VL that is pre-trained using English image-caption corpus (i.e. Conceptual Captions) only, M<sup>3</sup>P performs worse on the English test set. The possible reason could be that, M<sup>3</sup>P needs to balance its multilingual capability over 100+ languages, rather than on English only. (4) In both setting (i) *w/o fine-tune* and setting (ii) *w/ fine-tune on en*, integrating *Multimodal Code-switched Training* (MCT) into M<sup>3</sup>P can bring significant gains on non-English datasets, which demonstrates good multilingual transfer ability of *Multimodal Code-switched Training* in the zero-shot setting. It is expected to see such gains become smaller in setting (iii) *w/ fine-tune on each* and setting (iv) *w/ fine-tune on all*, as M<sup>3</sup>P can learn alignments between images and languages from labeled data directly.

## 5.2. Ablation Studies

Although we achieve good results under different settings, we want to deep dive into more aspects of M<sup>3</sup>P: (1) whether *Multimodal Code-switched Training* (MCT) can provide a positive effect under all settings; (2) whether the number of languages used in MCT affects the performance; (3) whether different pre-training tasks affect the performance.

### 5.2.1 The Impact of MCT

To verify whether the *Multimodal Code-switched Training* (MCT) strategy can provide a positive effect in different settings, we compare the performance of M<sup>3</sup>P without MCT and M<sup>3</sup>P with MCT under all fine-tuning settings.

| Setting                     | Multi30K    |             |             |             |
|-----------------------------|-------------|-------------|-------------|-------------|
|                             | en          | de          | fr          | cs          |
| <i>w/o fine-tune</i>        |             |             |             |             |
| M <sup>3</sup> P w/o MCT    | 54.9        | 28.9        | 25.2        | 13.5        |
| w/ MCT                      | <b>57.9</b> | <b>36.8</b> | <b>27.1</b> | <b>20.4</b> |
| <i>w/ fine-tune on en</i>   |             |             |             |             |
| M <sup>3</sup> P w/o MCT    | 86.0        | 48.6        | 37.1        | 34.6        |
| w/ MCT                      | <b>87.4</b> | <b>58.5</b> | <b>46.0</b> | <b>36.8</b> |
| <i>w/ fine-tune on each</i> |             |             |             |             |
| M <sup>3</sup> P w/o MCT    | 86.0        | 80.2        | 67.1        | <b>66.2</b> |
| w/ MCT                      | <b>87.4</b> | <b>82.1</b> | <b>67.3</b> | 65.0        |
| <i>w/ fine-tune on all</i>  |             |             |             |             |
| M <sup>3</sup> P w/o MCT    | 86.7        | 82.0        | 73.5        | 70.2        |
| w/ MCT                      | <b>87.7</b> | <b>82.7</b> | <b>73.9</b> | <b>72.2</b> |

Table 3: The impact of MCT for multilingual image-text retrieval. The metric is the mean Recall (mR). Each **bold number** indicates the best mR score.

For each setting in Table 3, we observe: (1) MCT im-

proves the performance on almost all languages, which shows its exciting robustness and expansibility, and (2) in both setting (i) and setting (ii), integrating MCT into M<sup>3</sup>P can bring significant gains on non-English datasets, which demonstrates the good multilingual transferability of MCT. It is expected to see such gains become smaller in settings (iii) and (iv), as M<sup>3</sup>P can learn alignments between images and languages from labeled data directly.

### 5.2.2 The Impact of Number of Languages in MCT

| Setting                    | Multi30K |      |      |      |
|----------------------------|----------|------|------|------|
|                            | en       | de   | fr   | cs   |
| M <sup>3</sup> P w/o MCT   | 54.9     | 28.9 | 25.2 | 13.5 |
| w/ 3 languages MCT         | 56.4     | 37.1 | 28.7 | 23.0 |
| w/ 5 languages MCT         | 58.2     | 36.7 | 26.9 | 23.6 |
| w/ 50 languages MCT (Full) | 57.9     | 36.8 | 27.1 | 20.4 |

Table 4: Impact of number of languages in *Multimodal Code-switched Training* (MCT). The metric is the mean Recall (mR). "Full" represents the model pre-trained with all Code-switching languages.

To verify whether the number of languages influences the performance of *Multimodal Code-switched Training* (MCT), we conduct an experiment by pre-training M<sup>3</sup>P by MCT with different numbers of languages and evaluate the model directly without fine-tuning. We pre-train M<sup>3</sup>P with the following settings: pre-train M<sup>3</sup>P without MCT, pre-train M<sup>3</sup>P with MCT on 3 languages (de, fr, cs), 5 languages (de, fr, cs, ja, zh), and all 50 languages.

In Table 4, we can find that, for languages like de and fr, there is no significant difference under different settings. On the contrary, for languages like en and cs, M<sup>3</sup>P achieves the best performance when MCT is activated with 5 languages. This implies that activating MCT on more languages can lead to more noise due to a higher probability of inaccurate translation. This noise may improve the robustness of the model but make the model harder to be well-trained.

### 5.2.3 The Impact of Proposed Tasks

We want to find whether each component during pre-training positively affects the performance and try to explain how they gain the performance by conducting several ablation experiments. Since *Multimodal Code-switched Training* (MCT) influences each task's target, we conduct the ablation experiments on M<sup>3</sup>P without MCT and fine-tune each model on the dataset of each language to compare the performance.

As shown in Table 5, we can observe that: (1) MC-VLM provides the most considerable improvement (+10.6 on en) to the model among all four sub-tasks during the pre-training

| Setting          | Multi30K    |             |             |             |
|------------------|-------------|-------------|-------------|-------------|
|                  | en          | de          | fr          | cs          |
| M <sup>3</sup> P | <b>86.0</b> | <b>80.2</b> | <b>67.1</b> | <b>66.2</b> |
| w/o xMLM         | 79.6        | 70.8        | 56.4        | 54.3        |
| w/o MC-MLM       | 84.3        | 76.2        | 64.1        | 62.2        |
| w/o MC-MRM       | 85.5        | 77.9        | 65.0        | 63.9        |
| w/o MC-VLM       | 75.4        | 68.3        | 52.7        | 50.9        |

Table 5: Ablation study on multilingual image-text retrieval. The metric is the mean Recall (mR). Each **bold number** indicates the best mR score in that column.

stage. We suggest this is because the MC-VLM sub-task successfully models the relationship between image and text. (2) xMLM shows a great impact on non-English results compared with English results, which shows that xMLM will improve the capability of multilinguality. (3) MC-MLM and MC-MRM also show good support to the results in all languages, which we suggest these two tasks will help the model learn the knowledge of multimodality. (4) When combining all tasks, we obtain the highest gain in all languages.

### 5.3. Expanding MCT to Fine-tuning

| Setting                                | Multi30K    |             |             |             |
|--|-------------|-------------|-------------|-------------|
|  | en          | de          | fr          | cs          |
| <i>Pre-trained without MCT</i>         |             |             |             |             |
| M <sup>3</sup> P (w/ Normal Fine-tune) | <b>86.0</b> | 48.6        | 37.1        | 34.6        |
| M <sup>3</sup> P (w/ MCT Fine-tune)    | 85.4        | <b>67.8</b> | <b>59.2</b> | <b>54.0</b> |
| <i>Pre-trained with MCT</i>            |             |             |             |             |
| M <sup>3</sup> P (w/ Normal Fine-tune) | <b>87.4</b> | 58.5        | 46.0        | 36.8        |
| M <sup>3</sup> P (w/ MCT Fine-tune)    | 86.4        | <b>71.8</b> | <b>62.3</b> | <b>59.6</b> |

Table 6: The results of expanding MCT to fine-tuning for multilingual image-text retrieval. The metric is the mean Recall (mR). Each **bold number** indicates the best mR score under the setting. *Normal Fine-tune* represents fine-tuning with English data directly and *MCT Fine-tune* represents fine-tuning with Code-switched English data.

Similar to MC-VLM, we use Code-switched data to fine-tune M<sup>3</sup>P on Multi30K. The results in Table 6 show: (1) *Multimodal Code-switched Training* (MCT) can bring a large margin for non-English language probably because of the lack of labeled image-non English caption pairs during the pre-training stage or fine-tuning stage. (2) Employing MCT into the fine-tuning stage for the model, whatever pre-trained by, will achieve a large increase in non-English performance. (3) MCT in fine-tuning is more effective than MCT in pre-training, which may be explained by that the model can learn multilinguality in a more specific task. (4) The best results

can be achieved when MCT replaces English in both the pre-training and fine-tuning stages.

#### 5.4. Qualitative Studies on MCT

To further explore how *Multimodal Code-switched Training* (MCT) affects the model, we randomly select some text-image pairs generated from it. We want to figure out why *Multimodal Code-switched Training* is very effective on non-English languages and whether it has any limitations.

|  |   |
|--|---|
|   | A dog is <u>sitting</u> on a <u>couch</u> .                           |
|  | A <u>chienne</u> is <u>坐着</u> on a <u>ソファ</u> .                       |
|  | A dog is sitting on a sofa.   |
| (a)  |   |
|  | Red <u>truck</u> driving <u>on</u> the <u>road</u> at <u>sunset</u> . |
|  | Red <u>交易</u> driving <u>sur</u> the <u>Straße</u> at <u>夕焼け</u> .    |
|  | Red <u>transaction</u> driving on the road at sunset.                 |
| (b)  |   |

Figure 2: Qualitative study for *Multimodal Code-switched Training* (MCT). The first row in each table is the original text, and the second row in each table is the Code-switched text. We add the meaning of the Code-switched text in English in the third row of each table.

As Figure 2 (a) shows, the meaning of the code-switched text generated by *Multimodal Code-switched Training* (MCT) is almost the same as that of the original text. Although there are some small differences between the original text (first row) and the generated text translated back to English (third row), it has no influence on the training quality, which demonstrates the reason why MCT brings gains. The key idea of using MCT in M<sup>3</sup>P is to let the model see more Code-switched text and image pairs and learn the joint multilingual multimodal representations from such pairs directly. We guess this helps the model learn richer information of each token from the multilingual context.

We did not consider the grammar or syntax correctness of the Code-switched sentences generated by replacing words in the English sentences with their word translations in other languages. The pre-trained models can learn such knowledge from well-formed multilingual sentences and En-

glish caption sentences. Since we don't have image-caption pairs or high-quality machine translation engines to generate such data for most languages, generating Code-switched sentences is the most effective way to let M<sup>3</sup>P directly see more alignments between non-English languages and images.

Hence, because of the high accuracy of translation from MCT, multilingual results will significantly increase when no non-English multimodal data is available. However, when the model can access high-quality multilingual multimodal data, the noise from MCT may limit its performance. In Figure 2 (b), we show a negative case in Code-switched text. MCT faultily changes the meaning of the original text. We leave this as future work to solve this problem.

#### 6. Conclusion

We have presented in this paper a new pre-trained model M<sup>3</sup>P which combines Multilingual Pre-training and Multimodal Pre-training into a unified framework via Multitask Pre-training for multilingual multimodal scenarios. We proposed Multimodal Code-switched Training to further alleviate the issue of lacking enough labeled data for non-English multimodal tasks and avoid the tendency to model the relationship between vision and English text.

#### References

- [1] Andrea Burns, Donghyun Kim, Derry Wijaya, Kate Saenko, and Bryan A Plummer. Learning to scale multilingual representations for vision-language tasks. *arXiv preprint arXiv:2004.04312*, 2020. 5, 6
- [2] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 2, 4, 5
- [3] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations. *arXiv preprint arXiv:1909.11740*, 2019. 1, 2
- [4] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019. 1, 5
- [5] Alexis Conneau and Guillaume Lample. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, pages 7057–7067, 2019. 1, 2, 3
- [6] Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. Xnli: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*, 2018. 2
- [7] 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*, pages 4171–4186, 2019. 1, 2, 3, 5



- [8] 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. *arXiv preprint arXiv:1710.07177*, 2017. 5
- [9] Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. Multi30k: Multilingual english-german image descriptions. *arXiv preprint arXiv:1605.00459*, 2016. 5
- [10] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. *arXiv preprint arXiv:2006.06195*, 2020. 1
- [11] Spandana Gella, Rico Sennrich, Frank Keller, and Mirella Lapata. Image pivoting for learning multilingual multimodal representations. In: *Empirical Methods in Natural Language Processing (EMNLP) (2017)*, 2017. 5, 6
- [12] Ross Girshick, Ilija Radosavovic, Georgia Gkioxari, Piotr Dollár, and Kaiming He. Detectron. 2018. 3
- [13] Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks. *arXiv preprint arXiv:1909.00964*, 2019. 1, 2, 3, 6
- [14] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015. 5
- [15] Donghyun Kim, Kuniaki Saito, Kate Saenko, Stan Sclaroff, and Bryan A Plummer. Mule: Multimodal universal language embedding. In: *AAAI Conference on Artificial Intelligence (2020)*, 2020. 5, 6
- [16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *international conference on learning representations*, 2015. 5
- [17] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017. 2
- [18] Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *EMNLP*, 2018. 2
- [19] Gen Li, Nan Duan, Yuejian Fang, Daxin Jiang, and Ming Zhou. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. *AAAI*, 2020. 1, 2, 4, 6
- [20] Xirong Li, Chaoxi Xu, Xiaoxu Wang, Weiyu Lan, Zhengxiong Jia, Gang Yang, and Jieping Xu. Coco-cn for cross-lingual image tagging, captioning and retrieval. In *IEEE Transactions on Multimedia*, 2019. 4, 5
- [21] 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. *arXiv preprint arXiv:2004.06165*, 2020. 1, 2
- [22] Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, et al. Xglue: A new benchmark dataset for cross-lingual pre-training, understanding and generation. *arXiv preprint arXiv:2004.01401*, 2020. 6
- [23] 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. *arXiv preprint arXiv:1907.11692*, 2019. 1
- [24] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23, 2019. 2, 4
- [25] Takashi Miyazaki and Nobuyuki Shimizu. Cross-lingual image caption generation. In *ACL*, 2016. 5
- [26] Vicente Ordonez, Girish Kulkarni, and Tamara L Berg. Im2text: Describing images using 1 million captioned photographs. In *Advances in neural information processing systems*, pages 1143–1151, 2011. 2
- [27] Libo Qin, Minheng Ni, Yue Zhang, and Wanxiang Che. Cosda-ml: Multi-lingual code-switching data augmentation for zero-shot cross-lingual nlp. *arXiv preprint arXiv:2006.06402*, 2020. 2, 3
- [28] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 1
- [29] 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 the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 2, 4, 5
- [30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 5
- [31] Liwei Wang, Yin Li, Jing Huang, and Svetlana Lazebnik. Learning two-branch neural networks for image-text matching tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):394–407, 2018. 5, 6
- [32] Jônatas Wehrmann, Douglas M Souza, Mauricio A Lopes, and Rodrigo C Barros. Language-agnostic visual-semantic embeddings. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5804–5813, 2019. 5, 6
- [33] Zhen Yang, Bojie Hu, Ambyera Han, Shen Huang, and Qi Ju. Csp: Code-switching pre-training for neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2624–2636, 2020. 2
- [34] Yuya Yoshikawa, Yutaro Shigeto, and Akikazu Takeuchi. Stair captions: Constructing a large-scale japanese image caption dataset. *arXiv preprint arXiv:1705.00823*, 2017. 4, 5
- [35] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 4, 5
- [36] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-vil: Knowledge enhanced vision-

language representations through scene graph. *arXiv preprint arXiv:2006.16934*, 2020. [1](#)

- [37] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. *AAAI*, 2020. [2](#)