# mCLIP: Multilingual CLIP via Cross-lingual Transfer

**Guanhua Chen[1], Lu Hou[2], Yun Chen[3], Wenliang Dai[5], Lifeng Shang[2], Xin Jiang[2], Qun Liu[2], Jia Pan[4], Wenping Wang[6]**
[1]Southern University of Science and Technology; [2]Huawei Noah's Ark Lab
[3]Shanghai University of Finance and Economics; [4]The University of Hong Kong;
[5]The Hong Kong University of Science and Technology; [6]Texas A&M University
chengh3@sustech.edu.cn, yunchen@sufe.edu.cn, wdaiai@connect.ust.hk
{houlu3, shang.lifeng, jiang.xin, qun.liu}@huawei.com
jpan@cs.hku.hk, wenping@tamu.edu

## Abstract

Large-scale vision-language pretrained (VLP) models like CLIP have shown remarkable performance on various downstream cross-modal tasks. However, they are usually biased towards English due to the lack of sufficient non-English image-text pairs. Existing multilingual VLP methods often learn retrieval-inefficient single-stream models by translation-augmented non-English image-text pairs. In this paper, we introduce mCLIP, a retrieval-efficient dual-stream multilingual VLP model, trained by aligning the CLIP model and a Multilingual Text Encoder (MTE) through a novel Triangle Cross-modal Knowledge Distillation (TriKD) method. It is parameter-efficient as only two light projectors on the top of them are updated during distillation. Furthermore, to enhance the token- and sentence-level multilingual representation of the MTE, we propose to train it with machine translation and contrastive learning jointly before the TriKD to provide a better initialization. Empirical results show that mCLIP achieves new state-of-the-art performance for both zero-shot and finetuned multilingual image-text retrieval task.

## 1 Introduction

Recently, large-scale dual-stream vision-language pretrained (VLP) models, such as CLIP (Radford et al., 2021), ALIGN (Jia et al., 2021) and their variants (Yao et al., 2021; Mu et al., 2021; Zhai et al., 2021), have shown remarkable performance on various downstream multimodal tasks. These models use separate encoders for the images and texts, and allow efficient inference in the image-text retrieval task because the image or text features can be computed offline. However, most current VLP models are biased toward English, due to the lack of sufficient high-quality multilingual multimodal datasets for direct large-scale pretraining.

Despite the lack of sufficient non-English image-text pairs, previous methods attempt to create word-level code-switched image-text pairs by looking up bilingual dictionaries (Ni et al., 2021), or sentence-level augmented multilingual image-text pairs by translating the English text to other languages (i.e, translate-train pipeline) (Zhou et al., 2021). Then each image and its paired text are concatenated as a single sequence to train a single-stream Transformer-based model. Despite the good performance of these models (Ni et al., 2021; Zhou et al., 2021), they are less efficient than dual-stream models on large-scale image-text retrieval tasks, as the data from both modalities are intertwined to compute the self-attention and the unimodal features can not be pre-computed. Instead of creating word-level or sentence-level multilingual image-text pairs, MURAL (Jain et al., 2021) extends the ALIGN (Jia et al., 2021) model with multilinguality by an additional text-text contrastive loss among hundreds of languages. However, MURAL is trained from scratch and requires large-scale training data with high computation cost to obtain strong performance on multilingual cross-modal retrieval tasks.

To tackle the aforementioned problems, we propose the *triangle cross-modal knowledge distillation* (TriKD) to learn a dual-stream multilingual VLP model mCLIP, which learns triangle alignment among the pretrained CLIP's image encoder, CLIP's text encoder and a pretrained Multilingual Text Encoder (MTE) through knowledge distillation. Specifically, to avoid catastrophic forgetting of the knowledge already learned in the pretrained CLIP and MTE, they are kept frozen. The triangle alignment is achieved by adjusting a linear projector on top of CLIP and a shallow Transformer-based X-projector on top of the MTE. Since the XLM-R (Conneau et al., 2020) used for initializing the MTE has unsatisfactory performance when directly used for retrieval tasks (Hu et al., 2020), before performing the TriKD, we propose to enhance the MTE via both the machine translation task and a contrastive loss to improve the token-

*Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*
*Volume 1: Long Papers*, pages 13028–13043
July 9-14, 2023 ©2023 Association for Computational Linguistics

and sentence-level multilingual representation.

The proposed mCLIP is both parameter- and computation-efficient as only the projectors are trained, which accounts for only 3% of the total parameters of mCLIP. Empirical results of zero-shot and finetuned multilingual image-text retrieval on MSCOCO (Lin et al., 2014) and Multi30K (Elliott et al., 2016) show that the proposed mCLIP achieves better performance while being much more efficient in inference than single-stream baselines or using less training data than MURAL. The proposed method can also be extended to train a multilingual VLP based on a unimodal image encoder and the MTE, 89.4% performance retained.[1]

## 2 Related Work

**Multilingual VLP Models.** Monolingual vision-language pretrained (VLP) models (Radford et al., 2021; Yao et al., 2021; Jia et al., 2021; Li et al., 2022) trained with large-scale image-text pairs have shown remarkable performance on various downstream tasks like image-text retrieval. Recently, some attempts have extended VLP models to the multilingual scenario. The first line of work applies the translation method to create multilingual image-text pairs and then concatenates the multilingual text with its paired image as a single sequential input to a single-stream Transformer-based encoder. For instance, M3P (Ni et al., 2021) constructs a multilingual code-switched text by randomly replacing English words with translations of other languages, and UC2 (Zhou et al., 2021) directly translates a whole sentence into other languages. However, these single-stream models are inefficient for the image-text retrieval task as unimodal features cannot be pre-computed beforehand. MURAL (Jain et al., 2021) directly trains from scratch with both augmented multilingual image-text pairs and parallel text corpus, which is expensive in both data and computation. Besides retrieval tasks, recent PaLI (Chen et al., 2022b) and ERNIE-UniX2 (Shan et al., 2022) use the encoder-decoder architectures for multilingual multimodal generation tasks. In this paper, we introduce a data- and parameter-efficient knowledge distillation method to train a dual-stream multilingual VLP model by aligning a frozen English VLP and a frozen MTE.

**Knowledge Distillation.** Knowledge distillation (Hinton et al., 2015) is firstly proposed for model compression. The knowledge in the output logits of a large teacher model can be transferred to a smaller student model without significant performance degradation. Besides the logits, the hidden states and attention outputs can also be used for knowledge distillation (Jiao et al., 2020; Hou et al., 2020). Recently, Tian et al. (2020b) propose to distill knowledge with contrastive learning, which maximizes the mutual information between the teacher and student models. For multimodal models, Wang et al. (2021) propose to train a dual-stream VLP model with the knowledge distilled from a single-stream model for faster inference. Furthermore, VLKD (Dai et al., 2022) augments a dual-stream VLP model with a pretrained language model via vision-language knowledge distillation, enabling the multimodal generation ability without hurting the original NLP ability. However, to the best of our knowledge, knowledge distillation has not been studied for training multilingual VLP models, for which efficiency is an important factor due to the data scarcity issue. In this paper, we introduce a novel triangle cross-modal knowledge distillation method to efficiently align a multilingual text encoder to the multimodal space of a pretrained dual-stream VLP model.

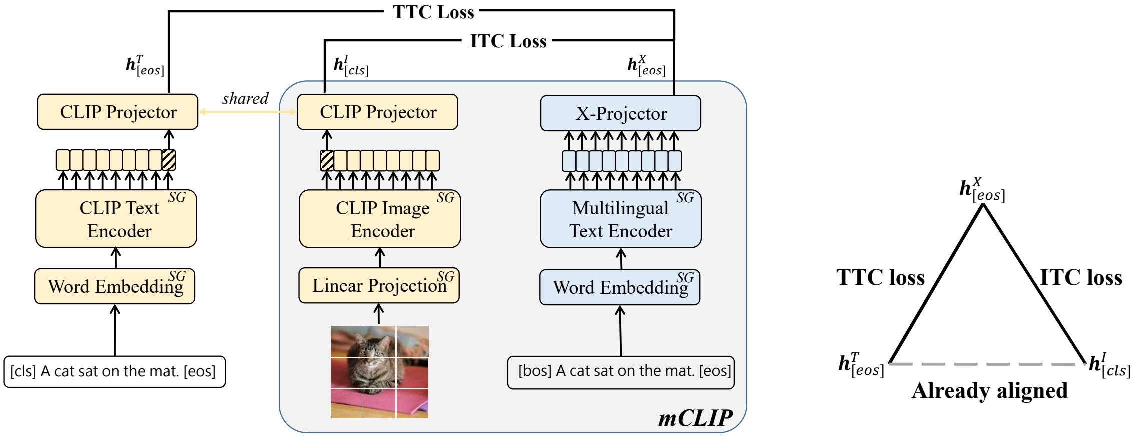
## 3 Method

In this section, we first introduce the architecture of mCLIP in Section 3.1. It extends the monolingual VLP model CLIP to a multilingual one by aligning CLIP and a multilingual text encoder (MTE) to a shared space, through a novel triangle cross-modal knowledge distillation (TriKD) using English image-text pairs (Section 3.2). The performance of mCLIP on non-English image-text retrieval is highly dependent on the quality of the multilingual representation of the MTE. Thus in Section 3.3, we propose to first improve the token- and sentence-level cross-lingual representation of the MTE with the neural machine translation (NMT) task and contrastive learning (CTL).

### 3.1 Model Structure

The architecture of mCLIP is shown in Figure 1a. Like CLIP, mCLIP is a dual-stream model with separate image and text encoders. The vision encoder of mCLIP is the original CLIP ViT image encoder, while the text encoder is a multilingual one initialized from XLM-R (Conneau et al., 2020) with enhanced representations.

[1] Our code is publicly available at `https://github.com/huawei-noah/noah-research/NLP/mclip`.



(a) Architecture of mCLIP.

(b) TriKD.

Figure 1: Architecture of mCLIP, obtained by Triangle cross-modal Knowledge Distillation (TriKD). *SG* is short for the *stop gradient* operation, which indicates a part is kept frozen without any gradient update.

**CLIP.** The image encoder of the pretrained CLIP (Radford et al., 2021) is aligned with the English text encoder, by contrastive learning over 400M English image-text pairs. The Vision Transformer (ViT) is used as a kind of CLIP image encoder, which takes image patches as input and generates the final feature through a Transformer-based model. An additional [cls] token is added before the image patches, and its output at the last Transformer layer represents the image's global feature. The CLIP text encoder has a similar structure to the GPT (Radford et al., 2019) model. The final output of the [eos] token represents the global feature of an English sentence. Note that CLIP's text encoder is only used during the training of mCLIP, but not inference.

**Multilingual Text Encoder.** Instead of using the original CLIP's English text encoder, we use the multilingual encoder XLM-R (Conneau et al., 2020) with enhanced token- and sentence-level cross-lingual representations (Section 3.3.)

Our ultimate goal is to learn the *triangle* alignment among the CLIP's image encoder, CLIP's English text encoder and the multilingual text encoder (MTE) in a shared multilingual multimodal representation space. In Section 3.2, we propose triangle cross-modal knowledge distillation (TriKD) to achieve this goal while maintaining the already learned alignment between the image and English text of CLIP, as well as the multilinguality of the learned MTE. Specifically, as is shown in Figure 1a, to avoid destroying the pretrained alignment between CLIP's image and text encoders, we freeze the parameters of both CLIP's image and text encoders, and use a shared linear projection (i.e., the CLIP-projector) on the top of them. On the other hand, to keep the learned multilinguality of XLM-R, we also freeze its parameters and align it to CLIP's multimodal space by optimizing the learnable X-projector, which consists of two randomly initialized XLM-R Transformer layers (Huang et al., 2021). The input to the X-projector is the outputs of all positions from the MTE. The [eos] output representation after the X-projector is used as the global representation of the text.

## 3.2 Triangle Cross-Modal Knowledge Distillation

Contrastive learning is proved effective in both unimodal (Tian et al., 2020a; Gao et al., 2021) and cross-modal (Radford et al., 2021) representation learning. Here, we also consider using contrastive losses to learn the triangle alignment among CLIP's image encoder, CLIP's English text encoder, and the multilingual text encoder (MTE). Since the image and text encoders of CLIP are already aligned, the TriKD contains only (i) an image-text contrastive (ITC) loss to align the MTE and CLIP image encoder; and (ii) a text-text contrastive (TTC) loss to align the MTE and CLIP's English text encoder (Figure 1).

In contrastive learning, the model parameters are optimized by letting the features of paired samples close and apart otherwise. Specifically, consider a training batch of $N$ samples, where $\mathbf{x}_i, \mathbf{y}_i$ are a pair of features from two views of the $i^{\text{th}}$ sample, e.g., the image and text features of an image-text pair; or the text features of the same text from two different text encoders. We use in-batch negatives, i.e. for $\mathbf{x}_i$, $\mathbf{y}_i$ is its positive, and all the other $\mathbf{y}_j$'s (where $j \neq i$) are its negatives. Denote $\mathbf{x} = \{\mathbf{x}_i\}_{i=1}^N$, $\mathbf{y} = \{\mathbf{y}_i\}_{i=1}^N$, and the temperature parameter is $\tau$. the contrastive loss can be written as

$$\ell(\mathbf{x}, \mathbf{y}) = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(\mathbf{x}_i^\top \mathbf{y}_i / \tau)}{\sum_{j=1}^N \exp(\mathbf{x}_i^\top \mathbf{y}_j / \tau)}. \quad (1)$$

**Image-Text Contrastive Loss.** For the $i^{\text{th}}$ image-text pair in a training batch, denote the $\ell_2$-normalized output of the [cls] token after the CLIP image encoder and CLIP-projector as $\mathbf{h}_i^I$ and the $\ell_2$-normalized output of [eos] after the MTE and X-projector as $\mathbf{h}_i^X$. Denote $\mathbf{h}^I = \{\mathbf{h}_i^I\}_{i=1}^N$ and $\mathbf{h}^X = \{\mathbf{h}_i^X\}_{i=1}^N$, the ITC loss $\mathcal{L}_{\text{ITC}}$ is formulated as the average of image-to-text ($\mathcal{L}_{\text{i2x}}$) loss and text-to-image ($\mathcal{L}_{\text{x2i}}$) loss:

$$\begin{aligned} \mathcal{L}_{\text{ITC}} &= 1/2(\mathcal{L}_{\text{i2x}} + \mathcal{L}_{\text{x2i}}) \\ &= 1/2[\ell(\mathbf{h}^I, \mathbf{h}^X) + \ell(\mathbf{h}^X, \mathbf{h}^I)]. \quad (2) \end{aligned}$$

**Text-Text Contrastive Loss.** For the $i^{\text{th}}$ image-text pair, suppose the $\ell_2$-normalized output of the [eos] token after the CLIP text encoder and CLIP-projector is $\mathbf{h}_i^T$. Denote $\mathbf{h}^T = \{\mathbf{h}_i^T\}_{i=1}^N$, the TTC loss is calculated as the average of contrastive losses in both directions:

$$\begin{aligned} \mathcal{L}_{\text{TTC}} &= 1/2(\mathcal{L}_{\text{t2x}} + \mathcal{L}_{\text{x2t}}) \\ &= 1/2[\ell(\mathbf{h}^T, \mathbf{h}^X) + \ell(\mathbf{h}^X, \mathbf{h}^T)], \quad (3) \end{aligned}$$

where $\mathcal{L}_{\text{t2x}}, \mathcal{L}_{\text{x2t}}$ are the contrastive losses of CLIP text features to XLM-R features and vice versa.

The training loss of the TriKD is the weighted sum of ITC and TTC losses:

$$\mathcal{L}_{\text{TriKD}} = \mathcal{L}_{\text{ITC}} + \lambda \mathcal{L}_{\text{TTC}}.$$

We use $\lambda = 0.1$ following Jain et al. (2021). For training with non-English image-text pairs, only ITC loss is applied as the CLIP text encoder does not support non-English languages.

Since the backbones of image and text encoders are frozen and only the additional projectors (3% of total parameters) are learnable, the training is efficient and allows a large batch size, which is shown to be crucial to the success of contrastive learning (Chen et al., 2020; Radford et al., 2021).

Through TriKD, though mCLIP learns only on English image-text pairs, it already implicitly has the ability to transfer to other languages through the multilinguality embedded in the frozen MTE. The retrieval performance on non-English languages relies on both the English text-image retrieval performance and the cross-lingual transferability of the MTE. However, the original XLM-R is not directly optimized for retrieval and its cross-lingual ability for retrieval is not satisfactory (Hu et al., 2020), so in Section 3.3, we propose a two-stage training method to enhance the MTE before TriKD.

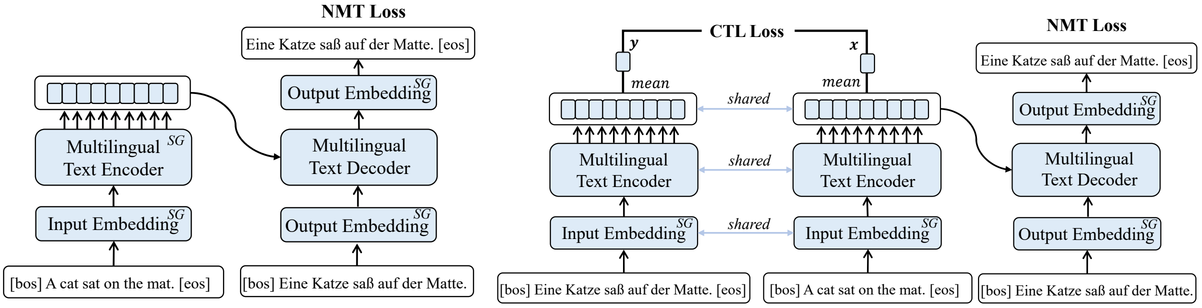
### 3.3 Multilingual Text Encoder

In this section, we propose to enhance the token- and sentence-level alignment among different languages of XLM-R for retrieval tasks, with the neural machine translation (NMT) task and contrastive learning on the textual-only multilingual parallel corpus. Intuitively, an NMT decoder generates semantic-equivalent translation with token-level interactions with the encoder output, encouraging the encoder output to maintain fine-grained token-level information, which is required as the X-projector is trained over token-level inputs during TriKD. On the other hand, the contrastive loss benefits cross-lingual transfer by explicitly aligning the sentence-level representations of parallel sentences.

Note that XLM-R is only an encoder, to train with the NMT loss, we add a decoder with randomly initialized weights (Figure 2a). Inspired by Chen et al. (2021), we adopt a two-stage training schedule to avoid catastrophic forgetting of the strong multilinguality of the pretrained XLM-R encoder. Before joint training with the NMT and contrastive loss, we freeze the encoder and train this decoder with the NMT task on parallel text corpus at the first stage. Note that all embeddings are initialized with XLM-R and fixed all the time. With a slight abuse of notation, here we denote $\mathbf{x}_i$ and $\mathbf{y}_i$ as the $i^{\text{th}}$ source and target sentence in a batch of $N$ paired sentences, and $|\mathbf{y}_i|$ is the length for sentence $\mathbf{y}_i$, the NMT loss can be formulated as:

$$\begin{aligned} &\mathcal{L}_{\text{stage\_1}} = \mathcal{L}_{\text{NMT}} \\ &\quad = -\frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{|\mathbf{y}_i|+1} \log p([\mathbf{y}_i]_t \mid [\mathbf{y}_i]_{0:t-1}, \mathbf{x}_i), \end{aligned}$$

At the second stage, we tune both the XLM-R encoder and the decoder with both NMT and contrastive loss (Figure 2b). Note that we do not tune the embeddings as no further improvements are observed empirically. Specifically, for the $i^{\text{th}}$ sentence pair, denote $\mathbf{h}_i^S, \mathbf{h}_i^O$ as the averaged representation of all the tokens of the source and target


NMT Loss
Eine Katze saß auf der Matte. [eos]
Output Embedding SG
Multilingual Text Encoder SG
Multilingual Text Decoder
Input Embedding SG
Output Embedding SG
[bos] A cat sat on the mat. [eos]
[bos] Eine Katze saß auf der Matte.
CTL Loss
y
x
mean
mean
shared
shared
shared
NMT Loss
Eine Katze saß auf der Matte. [eos]
Output Embedding SG
Multilingual Text Encoder
Multilingual Text Encoder
Multilingual Text Decoder
Input Embedding SG
Input Embedding SG
Output Embedding SG
[bos] Eine Katze saß auf der Matte. [eos]
[bos] A cat sat on the mat. [eos]
[bos] Eine Katze saß auf der Matte.

(a) First stage.

(b) Second stage.

Figure 2: Two-stage training to enhance the multilingual text encoder. At the 1st stage, a decoder is trained with the NMT loss. At the 2nd stage, we simultaneously train with the NMT and the contrastive losses to learn both fine-grained token-level and retrieval-friendly sentence-level representations.

sentences from the last encoder layer, respectively. Denote $\mathbf{h}^S = \{\mathbf{h}_i^S\}_{i=1}^N$, $\mathbf{h}^O = \{\mathbf{h}_i^O\}_{i=1}^N$, the contrastive loss and training loss are:

$$\begin{aligned}\mathcal{L}_{\text{CTL}} &= 1/2[\ell(\mathbf{h}^S, \mathbf{h}^O) + \ell(\mathbf{h}^O, \mathbf{h}^S)], \\ \mathcal{L}_{\text{stage\_2}} &= \mathcal{L}_{\text{NMT}} + \alpha\mathcal{L}_{\text{CTL}},\end{aligned}$$

where $\ell(\cdot,\cdot)$ is the contrastive loss defined in Equation 1, and $\alpha$ is the weight to balance the two loss terms, which is set as $\alpha = 2.0$ in our experiments.

Note that when computing the contrastive loss during the second stage, we select the average representation over all tokens as the sentence-level text feature instead of the [eos] feature, as the former empirically performs better in the cross-modal retrieval task. We speculate this is because the X-projector uses token-level outputs from the MTE instead of the [eos] representation for learning alignment between images and texts. After the two-stage training, the MTE is used to initialize mCLIP using the TriKD method in Section 3.2.

# 4 Experiments

## 4.1 Setup

**Models and Pretraining Datasets.** We train two models (i.e., mCLIP and mCLIP+) based on the officially released CLIP ViT-B/32 and XLM-R (Conneau et al., 2020) base models. For the vanilla mCLIP, the enhanced MTE in Section 3.3 is trained with the parallel text corpus MT6, which contains 120M parallel sentences between English and six languages and covers 12 language directions (Chen et al., 2022a). Then we perform the TriKD in Section 3.2 with the cross-modal dataset CC3M (Sharma et al., 2018). For mCLIP+, its MTE is trained with OPUS-100 (Zhang et al., 2020) dataset in addition to MT6, covering a total of 175M parallel sentences among 100 languages. The TriKD of mCLIP+ is performed with TrTrain(CC12M), which is obtained by applying the translate-train method and translating the English captions of CC12M (Changpinyo et al., 2021) into Czech, German, Japanese and French with an in-house translator. Note that the TTC loss is removed during TriKD for non-English image-text pairs. More details about MT6 and OPUS-100 are in Appendix A.2. We use the XTD10 (Aggarwal and Kale, 2020) Spanish image-text pairs as the validation set to select the checkpoints, as we care more about the multilingual cross-modal performance.

**Downstream Tasks and Evaluation Metrics.** We test the efficacy of the proposed mCLIP on both multilingual image-to-text and text-to-image retrieval tasks, on the test sets of Multi30K (Elliott et al., 2016) and MSCOCO (Lin et al., 2014). We use the same data splits as Young et al. (2014) and Karpathy and Fei-Fei (2015). More details are in Appendix A.1. For both retrieval tasks, we compute the recall of top-K candidates (recall@K) with K=1, 5, and 10. The mean recall averaged over all these 6 scores is used as the evaluation metric. Following Ni et al. (2021), we evaluate the model's zero-shot and finetuned performance. Under the *zero-shot* setting, the pretrained mCLIP is directly tested on multilingual retrieval tasks. We use three finetuned settings: (i) *English-only Finetune*: finetune the pretrained mCLIP with only English Multi30K or MSCOCO and test on each target language; (ii) *Single-language Finetune:* finetune with training data of target language and test; and (iii) *All-language Finetune:* finetune on training data of all languages and test on each language.

**Compared Methods.** We compare our proposed method against the recent multilingual multimodal models M3P (Ni et al., 2021), UC2 (Zhou et al., 2021) and MURAL (Jain et al., 2021). The results of these models are taken from their original pa-

| Model | Pretraining Data |  | Multi30K |  |  |  | MSCOCO |  |  | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|
|  | Image-text Pairs | Text (#languages) | En | De | Fr | Cs | En | Ja | Zh |  |
| *Zero-shot* |  |  |  |  |  |  |  |  |  |  |
| M3P | CC3M | 101G (100) | 57.9 | 36.8 | 27.1 | 20.4 | 63.1 | 33.3 | 32.3 | 38.7 |
| MURAL | TrTrain(CC12M) | 500M (124) | 80.9 | 76.0 | 75.7 | 68.2 | 58 | 49.7 | – | 68.1⋆ |
| mCLIP | CC3M | 120M (6) | 72.3 | 62.4 | 45.2 | 55.3 | 53.2 | 36.1 | 63.0 | 55.4/54.1⋆ |
| mCLIP+ | TrTrain(CC12M) | 175M (100) | 77.1 | 76.6 | 76.1 | 74.5 | 59.2 | 55.6 | 71.8 | 70.1/69.9⋆ |
| *English-only Finetune* |  |  |  |  |  |  |  |  |  |  |
| M3P | CC3M | 101G (100) | 87.4 | 58.5 | 46.0 | 36.8 | 88.6 | 53.8 | 56.0 | 61.0 |
| UC2 | TrTrain(CC3M) | - | 87.2 | 74.9 | 74.0 | 67.9 | – | – | 82.0 | 77.2† |
| MURAL | TrTrain(CC12M) | 500M (124) | 91.0 | 87.3 | 86.4 | 82.4 | 73.7 | 71.9 | – | 82.1⋆ |
| mCLIP | CC3M | 120M (6) | 97.6 | 83.0 | 61.5 | 77.7 | 69.4 | 50.6 | 76.5 | 73.8/79.3†/73.3⋆ |
| mCLIP+ | TrTrain(CC12M) | 175M (100) | 98.5 | 91.4 | 91.7 | 89.1 | 71.3 | 64.1 | 80.5 | 83.8/90.2†/84.4⋆ |
| *Single-language Finetune* |  |  |  |  |  |  |  |  |  |  |
| M3P | CC3M | 101G (100) | 87.4 | 82.1 | 67.3 | 65.0 | 88.6 | 80.1 | 75.8 | 78.0 |
| UC2 | TrTrain(CC3M) | - | 87.2 | 83.8 | 77.6 | 74.2 | – | – | 84.9 | 81.5† |
| mCLIP | CC3M | 120M (6) | 97.6 | 80.9 | 75.9 | 76.7 | 69.4 | 68.2 | 82.3 | 78.7/82.7† |
| mCLIP+ | TrTrain(CC12M) | 175M (100) | 98.5 | 82.9 | 81.6 | 79.9 | 71.3 | 71.6 | 83.6 | 81.3/85.3† |
| *All-language Finetune* |  |  |  |  |  |  |  |  |  |  |
| M3P | CC3M | 101G (100) | 87.7 | 82.7 | 73.9 | 72.2 | 88.7 | 87.9 | 86.2 | 81.0 |
| UC2 | TrTrain(CC3M) | - | 88.2 | 84.5 | 83.9 | 81.2 | – | – | 87.5 | 85.1† |
| mCLIP | CC3M | 120M (6) | 96.6 | 91.9 | 89.9 | 90.0 | 69.1 | 68.7 | 82.8 | 84.1/90.2† |
| mCLIP+ | TrTrain(CC12M) | 175M (100) | 94.5 | 89.8 | 90.1 | 88.0 | 71.8 | 71.7 | 85.9 | 84.5/89.7† |

Table 1: Mean recall on cross-modal retrieval test sets. † and ⋆ denote the score averaged in the same languages as UC2 and MURAL, respectively.

pers. MURAL-base with the similar model size is compared. Note that UC2 does not report its zero-shot results. Its results on English and Japanese MSCOCO test sets are not directly comparable with the other methods, because they simplified the task by splitting the 5k images and 25k captions into five smaller test sets to calculate the scores. The training details and hyperparameters can be found in Appendix A.4.

### 4.2 Main Results

The zero-shot and finetuned cross-modal retrieval results on Multi30K and MSCOCO are shown in Table 1. As can be seen, finetuning and using more pretraining data improves the performance of our model. In particular, All-language Finetune has the highest mean recall score for both mCLIP and mCLIP+. We speculate this is because image-text pairs with diverse languages allow the projectors to learn the multilingual multimodal alignment *explicitly*, instead of relying on the *implicit* multilinguality embedded in the MTE.

**Comparison with Baselines.** Compared with M3P, our proposed mCLIP achieves 16.7 and 12.8 more mean recall scores in zero-shot and English-only finetuned settings, respectively, despite that M3P uses more fine-grained code-switched image-text pairs and more languages. Moreover, M3P is a single-stream model and can be less efficient for retrieval tasks. Compared with UC2 pretrained on 5x larger translation-augmented TrTrain(CC3M), mCLIP trained with only English CC3M achieves 2.1 higher mean recall scores on English-only Finetune. Again, UC2 is a single-stream model like M3P and also suffers from inefficient inference. Compared with MURAL, the mean recall of mCLIP+ is 1.8 (resp. 2.3) points higher under the zero-shot (resp. English-only Finetune) setting with about 1/3 parallel texts. This may be because the MTE of mCLIP+ learns strong multilinguality in Section 3.3 and the X-projector only needs to focus on the multimodal alignment rather than the multilingual alignment. In contrast, MURAL has to learn to align the multilingual texts from scratch with parallel texts. Besides achieving better performance with less training data, mCLIP is also parameter-efficient, i.e., the learnable projectors only account for 3% of the total parameters during the triangle distillation.

### 4.3 Ablation Study

In this section, we conduct ablation studies using mCLIP pretrained on CC3M and report results on zero-shot multilingual image-text retrieval tasks.

**Components of Training Objectives.** Table 2 shows the effect of different training objectives in training the enhanced multilingual text encoder (MTE) and during the triangle cross-modal knowledge distillation (TriKD). mCLIP$-\mathcal{L}_{\text{NMT}}$ represents finetuning the XLM-R with only contrastive loss on parallel texts, while mCLIP$-\mathcal{L}_{\text{CTL}}$ is to finetune XLM-R with the two-stage training scheme only on the NMT task. mCLIP$-\mathcal{L}_{\theta}$ represents the mCLIP trained with original XLM-R from Conneau et al. (2020). As can be seen, in the TriKD, both image-text and text-text contrastive loss contribute positively to the performance and the image-text contrastive loss $\mathcal{L}_{\text{ITC}}$ is more crucial to the retrieval performance. In the learning of enhanced MTE, both the contrastive and NMT losses improve the performance in non-English languages. However, NMT loss improves English image-text retrieval while contrastive loss degrades it. This may be because the NMT loss allows the MTE to learn more fine-grained token-level textual representations, and facilitate the learning of the X-projector which relies on these token-level inputs.

| Model | Multi30K | | | | MSCOCO | | | Avg. |
|---|---|---|---|---|---|---|---|---|
| | En | De | Fr | Cs | En | Ja | Zh | |
| mCLIP | 72.3 | 62.4 | 45.2 | 55.3 | 53.2 | 36.1 | 63.0 | 55.4 |
| *Enhanced Multilingual Text Encoder* | | | | | | | | |
| $-\mathcal{L}_{\text{CTL}}$ | 70.6 | 58.6 | 51.6 | 51.7 | 56.2 | 31.4 | 58.1 | 54.0 |
| $-\mathcal{L}_{\text{NMT}}$ | 62.6 | 57.8 | 50.5 | 56.0 | 39.0 | 32.2 | 60.3 | 51.2 |
| $-\mathcal{L}_{\theta}$ | 70.7 | 50.6 | 48.9 | 36.7 | 51.4 | 21.7 | 49.7 | 47.1 |
| *Triangle Cross-modal Knowledge Distillation* | | | | | | | | |
| $-\mathcal{L}_{\theta}-\mathcal{L}_{\text{TTC}}$ | 66.5 | 48.9 | 46.8 | 36.3 | 48.2 | 21.2 | 49.3 | 45.3 |
| $-\mathcal{L}_{\theta}-\mathcal{L}_{\text{ITC}}$ | 30.2 | 25.4 | 24.8 | 19.8 | 14.6 | 8.7 | 25.4 | 21.3 |

Table 2: Ablation on training objectives used in training the enhanced multilingual text encoder and triangle cross-modal knowledge distillation.

**Design Choices of Locked Parameters.** We compare different design choices of locked parameters of mCLIP in Table 3. As can be seen, the performances of all three languages drop when either CLIP or XLM-R is finetuned. Finetuning CLIP degrades the performance because the image encoder gradually forgets its learned knowledge from large-scale pretraining on 400M image-text pairs and tends to overfit to the small CC3M dataset used for triangle distillation. When the XLM-R is finetuned, the ability of multilingual transfer degrades and the text encoder biases toward English. When both CLIP and XLM-R are locked, the knowledge embedded in these two models is maintained, contributing to the success of the cross-lingual cross-modal transfer. Yet another advantage of locking both backbones is the improved training efficiency, which allows the much larger batch size, as 97% parameters are frozen during training.

| CLIP | XLM-R | En | Ja | Zh | Avg. |
|---|---|---|---|---|---|
| locked | locked | 53.2 | 36.1 | 63.0 | 50.8 |
| trainable | locked | 28.4 | 18.4 | 44.4 | 30.4 |
| locked | trainable | 52.5 | 33.3 | 61.7 | 49.2 |
| trainable | trainable | 41.1 | 25.1 | 57.3 | 41.2 |

Table 3: Ablation on different choices of locked parameters during TriKD. The zero-shot mean recall scores on the MSCOCO dataset are reported.

### 4.4 Discussion

**Results on More Languages.** Table 4 compares our model with baselines on more diverse languages of the cross-modal retrieval task in the IGLUE (Bugliarello et al., 2022) benchmark[2]. Following Bugliarello et al. (2022), we report the mean recall@1 score under the zero-shot setting. The results of M3P and UC2 are taken from Bugliarello et al. (2022). We do not compare with MURAL as it is not open-sourced and its original paper does not report results on IGLUE. As can be seen, mCLIP+ has the best performance among all languages, achieving 17.2 and 17.8 higher averaged Recall@1 score than M3P and UC2.

**Different Image Encoder Backbones.** Besides using the CLIP-ViT as the image encoder of mCLIP, we also try to use the Swin Transformer (Swin-B[3]) (Liu et al., 2021), a novel unimodal model trained with only the image classification dataset. We use the same setup as Section 4.1 except that we remove the TTC loss. Empirical results on zero-shot retrieval on Multi30K and MSCOCO in Table 6 show that using Swin Transformer has 89.4% mean recall scores of that using CLIP-ViT. This indicates that our proposed method can also be extended to align a unimodal image encoder and a multilingual text encoder into a multilingual multimodal model.

[2] https://github.com/e-bug/iglue

[3] swin_base_patch4_window7_in22k of timm toolkit

| Model | Ar | Bg | Da | El | Et | Id | Ja | Ko | Tr | Vi | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| M3P | 8.6 | 9.3 | 10.6 | 10.8 | 6.8 | 9.8 | 7.7 | 6.6 | 8.5 | 11.7 | 9.1 |
| UC2 | 7.5 | 8.3 | 9.9 | 10.2 | 5.4 | 10.7 | 10.3 | 5.0 | 8.2 | 9.2 | 8.5 |
| mCLIP | 10.1 | 20.1 | 17.6 | 14.9 | 9.3 | 17.3 | 18.2 | 8.4 | 13.1 | 19.4 | 14.8 |
| mCLIP+ | 22.5 | 26.3 | 31.0 | 24.3 | 20.7 | 32.9 | 23.6 | 19.3 | 28.1 | 34.5 | 26.3 |

Table 4: Recall@1 results of cross-modal retrieval task in IGLUE benchmark.

| Methods | Alignment | | Uniformity | | |
|---|---|---|---|---|---|
| | (Image, English text) | (Image, Non-English text) | Image | English | Non-English |
| mCLIP(vanilla XLM-R) | 1.27 | 1.41 | -2.53 | -3.23 | -3.03 |
| $+\mathcal{L}_{\text{NMT}}$ | 1.20 (0.07) | **1.34 (0.07)** | -2.82 (0.29) | **-3.34 (0.11)** | -3.17 (0.14) |
| $+\mathcal{L}_{\text{CTL}}$ | 1.29 (-0.02) | 1.35 (0.06) | -2.62 (0.09) | -3.23 (0.00) | **-3.21 (0.18)** |
| $+\mathcal{L}_{\text{NMT}} + \mathcal{L}_{\text{CTL}}$ | **1.19 (0.08)** | 1.36 (0.05) | **-2.87 (0.34)** | **-3.34 (0.11)** | -3.17 (0.14) |

Table 5: Alignment and uniformity scores on XTD10 image-text retrieval test sets. Numbers in the bracket show the absolute improvement over the mCLIP with vanilla XLM-R as its MTE.

| Image Encoder | Multi30K | | | | MSCOCO | | | Avg. |
|---|---|---|---|---|---|---|---|---|
| | En | De | Fr | Cs | En | Ja | Zh | |
| CLIP-ViT | 72.3 | 62.4 | 45.2 | 55.3 | 53.2 | 36.1 | 63.0 | 55.4 |
| Swin-ViT | 68.3 | 54.2 | 41.4 | 50.5 | 50.0 | 32.5 | 61.3 | 49.5 |

Table 6: Mean recall scores of using different visual backbones for mCLIP.

## 5 Analysis of the Representations

The training objectives of contrastive learning encourage the positive samples to stay closer (i.e., alignment) while the negative samples to scatter on the hypersphere (i.e., uniformity) (Wang and Isola, 2020). Similarly, a desired multilingual and multimodal model should also learn good alignment between images and multilingual texts, as well as uniform representations within each modality.

We analyze the quality of the learned representations with the uniformity and alignment scores introduced in Wang and Isola (2020). The alignment score $\ell_{\text{align}} = \mathbb{E}(||\mathbf{h}_i^I - \mathbf{h}_i^X||^2)$ measures the distance between the $\ell_2$-normalized features of the image-text pairs (i.e, $\mathbf{h}_i^I, \mathbf{h}_i^X$ for the $i$-th image-text pair), while the uniformity score measures how uniformly the representations are distributed: $\ell_{\text{uniform}} = -\mathbb{E}(\exp(-2||\mathbf{h}_i^* - \mathbf{h}_j^*||^2))$, where $\mathbf{h}_i^*, \mathbf{h}_j^*$ with $* \in \{I, X\}$ are $\ell_2$-normalized features of different samples from the same modality. Smaller alignment and uniformity scores indicate higher alignment and uniformity, and thus better learned representations. We use mCLIP trained on English CC3M and analyze the learned representations of XTD10 (Aggarwal and Kale, 2020) test set with the two metrics. The uniformity score is calculated for each of the three modalities: images, English text, and non-English text. We report results for non-English languages averaged over It, Es, Ru, Pl, Ko, Zh, and Tr.

From Table 5, both CTL and NMT losses improve the alignment and uniformity scores for non-English languages, as well as the uniformity of the images. However, for English, the NMT loss improves both scores while the CTL loss cannot improve or even degrade them. This is consistent with the finding in Table 2 where CTL loss leads to worse English retrieval performance. This again affirms that the NMT loss learns more fine-grained token-level representation which benefits the X-projector for aligning the English image-text pairs, thus rendering better alignment and uniformity scores for English. The NMT loss also reduces the burden of mCLIP projectors to learn multilingual alignment, which contributes to better uniformity of image features. To summarize, mCLIP relies on NMT for English image-text retrieval and CTL for further improvement on non-English retrieval (mainly on the uniformity of the image).

## 6 Conclusion

In this paper, we introduce mCLIP, a novel multilingual vision-language pretrained model which aligns CLIP and an enhanced multilingual text encoder through triangle cross-modal knowledge distillation. This distillation method is both parameter-efficient with only 3% of the total parameters of mCLIP trained, and data-efficient with only English image-text pairs required. The performance of mCLIP can be further improved with more parallel text corpus from more languages and multilingual image-text pairs from the translate-train pipeline. Empirical results show that the proposed mCLIP+ achieves state-of-the-art performance in multilingual image-text retrieval tasks.

## 7 Limitations

This work only explores the multilingual VLP model for the image-text retrieval task. We leave the exploration of other multilingual vision-and-language downstream tasks such as visual question answering as future work. At the same time, our proposed method relies on a well-pretrained vision Transformer and a multilingual text encoder. Its performance is heavily influenced by the performance of the visual and textual backbones. This hinders the mCLIP from further improvements with the given backbones.

## 8 Ethical Considerations

We present a data- and training-efficient approach to build a multilingual VLP model mCLIP, by aligning the pretrained monolingual VLP model CLIP and a multilingual text encoder XLM-R to the same multimodal multilingual space. Despite the strong multimodal and multilingual abilities inherited from both models, the proposed mCLIP also inherits the societal impacts including some negative ones of the original CLIP and XLM-R, e.g., societal biases (Radford et al., 2021) and misuse of language models (Tamkin et al., 2021). The implicit biases are expected to be removed by debiasing either the dataset or the model (Meade et al., 2022; Zhou et al., 2022). Besides, our proposed method makes it simpler to retrieve malicious or offensive content (Welbl et al., 2021) from image-text pairs of different languages. Future explorations are needed to mitigate the misuse of VLP models.

## Acknowledgements

This project was supported by National Natural Science Foundation of China (No. 62106138) and Shanghai Sailing Program (No. 21YF1412100). Jia Pan and Wenping Wang are partially supported by Centre for Transformative Garment Production. We thank the anonymous reviewers for their insightful feedbacks on this work.

## References

Pranav Aggarwal and Ajinkya Kale. 2020. Towards zero-shot cross-lingual image retrieval. Preprint arXiv:2012.05107.

Emanuele Bugliarello, Fangyu Liu, Jonas Pfeiffer, Siva Reddy, Desmond Elliott, Edoardo Maria Ponti, and Ivan Vulić. 2022. IGLUE: A benchmark for transfer learning across modalities, tasks, and languages. In *Proceedings of ICML*, volume 162, pages 2370–2392.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of CVPR*.

Guanhua Chen, Shuming Ma, Yun Chen, Li Dong, Dongdong Zhang, Jia Pan, Wenping Wang, and Furu Wei. 2021. Zero-shot cross-lingual transfer of neural machine translation with multilingual pretrained encoders. In *Proceedings of EMNLP*, pages 15–26.

Guanhua Chen, Shuming Ma, Yun Chen, Dongdong Zhang, Jia Pan, Wenping Wang, and Furu Wei. 2022a. Towards making the most of cross-lingual transfer for zero-shot neural machine translation. In *Proceedings of ACL*, pages 142–157, Dublin, Ireland.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of ICML*.

Xi Chen, Xiao Wang, Soravit Changpinyo, A. J. Piergiovanni, Piotr Padlewski, Daniel M. Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish V. Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. 2022b. PaLI: A jointly-scaled multilingual language-image model. Preprint arXiv:2209.06794.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL*, pages 8440–8451.

Wenliang Dai, Lu Hou, Lifeng Shang, Xin Jiang, Qun Liu, and Pascale Fung. 2022. Enabling multimodal generation on clip via vision-language knowledge distillation. In *Findings of ACL*.

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.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of EMNLP*, pages 6894–6910.

Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. In *NeurIPS Deep Learning and Representation Learning Workshop*.

Lu Hou, Zhiqi Huang, Lifeng Shang, Xin Jiang, Xiao Chen, and Qun Liu. 2020. Dynabert: Dynamic bert with adaptive width and depth. In *Proceedings of NeurIPS*, pages 9782–9793.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multitask benchmark for evaluating cross-lingual generalisation. In *Proceedings of ICML*, volume 119, pages 4411–4421.

Po-Yao Huang, Mandela Patrick, Junjie Hu, Graham Neubig, Florian Metze, and Alexander Hauptmann. 2021. Multilingual multimodal pre-training for zero-shot cross-lingual transfer of vision-language models. In *Proceedings of NAACL*, pages 2443–2459.

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 EMNLP*, pages 3449–3463.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of ICML*, pages 4904–4916.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. TinyBERT: Distilling BERT for natural language understanding. In *Findings of EMNLP*, pages 4163–4174.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of CVPR*, pages 3128–3137.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of ICLR*, pages 100–108.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of ACL*, pages 66–75.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of ICML*, volume 162, pages 12888–12900.

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.

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 *Proceedings of ECCV*, pages 740–755.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of ICCV*.

Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. 2022. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. In *Proceedings of ACL*, pages 1878–1898.

Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. 2021. Slip: Self-supervision meets language-image pre-training. Preprint arXiv:2112.12750.

Minheng Ni, Haoyang Huang, Lin Su, Edward Cui, Taroon Bharti, Lijuan Wang, Jianfeng Gao, Dongdong Zhang, and Nan Duan. 2021. M3p: Learning universal representations via multitask multilingual multimodal pre-training. In *Procceedings of CVPR*, pages 3977–3986.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of ICML*, pages 8748–8763.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. Technical report.

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. 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 ACL*, pages 2556–2565.

Alex Tamkin, Miles Brundage, Jack Clark, and Deep Ganguli. 2021. Understanding the capabilities, limitations, and societal impact of large language models. Preprint arxiv:2102.02503.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2020a. Contrastive multiview coding. In *Proceedings of ECCV*, pages 776–794.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2020b. Contrastive representation distillation. In *Proceedings of ICLR*.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of ICML*, pages 9929–9939.

Zekun Wang, Wenhui Wang, Haichao Zhu, Ming Liu, Bing Qin, and Furu Wei. 2021. Distilled dual-encoder model for vision-language understanding. Preprint arxiv:2112.08723.

Johannes Welbl, Amelia Glaese, Jonathan Uesato, Sumanth Dathathri, John Mellor, Lisa Anne Hendricks, Kirsty Anderson, Pushmeet Kohli, Ben Coppin, and Po-Sen Huang. 2021. Challenges in detoxifying language models. In *Findings of EMNLP*, pages 2447–2469.

Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. 2021. Filip: Fine-grained interactive language-image pre-training. In *Proceedings of ICLR*.

Yuya Yoshikawa, Yutaro Shigeto, and Akikazu Takeuchi. 2017. STAIR captions: Constructing a large-scale Japanese image caption dataset. In *Proceedings of ACL*, pages 417–421.

Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. 2020. Large batch optimization for deep learning: Training bert in 76 minutes. In *Proceedings of ICLR*.

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

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. 2021. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of CVPR*, pages 18102–18112.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of ACL*, pages 1628–1639.

Kun Zhou, Beichen Zhang, Xin Zhao, and Ji-Rong Wen. 2022. Debiased contrastive learning of unsupervised sentence representations. In *Proceedings of ACL*, pages 6120–6130.

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 CVPR*, pages 4155–4165.

# A More Experimental Setup

## A.1 Cross-Modal Dataset for Image-text Retrieval

**Multi30K.** The Flickr30K dataset (Young et al., 2014) contains 31k images in total, each of which has five English captions. The Multi30K dataset (Elliott et al., 2016) extends the Flickr30K dataset to three other languages. Each image has five German captions, one Czech caption and one French caption. Following previous works (Ni et al., 2021; Jain et al., 2021), we use the same train/validation/test splits as Karpathy and Fei-Fei (2015) for each language.

**MSCOCO.** The MSCOCO dataset contains 123k images, each of which has five English captions. Yoshikawa et al. (2017) manually create the Japanese descriptions for MSCOCO images. Li et al. (2019) extend MSCOCO with Chinese captions for 20K images. Following previous works (Ni et al., 2021; Jain et al., 2021), we use the same dataset splits as Karpathy and Fei-Fei (2015) for English and Japanese, and the test set of each language has 5k images and 25k captions. For Chinese, we use the same dataset split as Li et al. (2019), whose test set has 1000 image-text pairs.

## A.2 Machine Translation Dataset

Training the enhanced multilingual text encoder (MTE) in Section 3.3 requires parallel sentences. Thus we create a dataset called MT6, which contains 120 million parallel sentences between English and six languages: Czech, German, Japanese, Russian, Spanish, and Chinese. The MT6 dataset is from WMT translation task[4], CzEng 1.6[5], JParaCrawl v1.0[6] and CCAligned corpus[7]. For Es-En and Ru-En, MT6 uses the first 20M sentence pairs of the CCAligned corpus. The validation sets are from the development and test sets of the WMT translation task. More details are shown in Table 7. To compare with MURAL, we combine MT6 dataset and OPUS-100[8] (Zhang et al., 2020) to train the enhanced MTE. All texts are tokenized by the sentencepiece (Kudo, 2018) tokenizer as used in the original XLM-R model (Conneau et al., 2020). The source sentence length is limited to 512, which is the maximum source sentence length supported by XLM-R.

| Split | Language | Source | # Sentences |
|---|---|---|---|
| Train | Cs-En | CzEng 1.6 | 8.1M |
| | De-En | WMT19 | 41.0M |
| | Es-En | CCAligned | 20.0M |
| | Ja-En | JparaCrawl v1.0 | 8.6M |
| | Ru-En | CCAligned | 20.0M |
| | Zh-En | WMT18 | 22.6M |
| Valid | Cs-En | Newstest 16 | 2,999 |
| | De-En | Newstest 16 | 2,999 |
| | Es-En | Newstest 10 | 2,489 |
| | Ja-En | Newsdev 20 | 1,998 |
| | Ru-En | Newstest 16 | 2,998 |
| | Zh-En | Newstest 17 | 2,001 |

Table 7: Training and validation sets of the MT6 dataset. "# Sentences" denotes the number of parallel sentences.

| ISO | Language | ISO | Language |
|---|---|---|---|
| Ar | Arabic | Id | Indonesian |
| Bg | Bulgarian | It | Italian |
| Cs | Czech | Ja | Japanese |
| Da | Danish | Ko | Korean |
| De | German | Pl | Polish |
| El | Greek | Ru | Russian |
| En | English | Tr | Turkish |
| Es | Spanish | Vi | Vietnamese |
| Et | Estonian | Zh | Chinese |
| Fr | French | | |

Table 8: Languages used in this paper.

## A.3 Language ISO code

The languages used in this paper are shown in Table 8.

## A.4 Training Details

We first train the enhanced multilingual text encoder from XLM-R following Section 3.3. Adam (Kingma and Ba, 2015) is used as the optimizer. Each batch has 32,768 tokens. At the first training stage, the learning rate is warmed up to 0.0005 within 4,000 steps, and then decays to 0. At the second stage, the learning rate decays from 0.0001 to 0 without warmup. The model is trained for one epoch at the first stage and 0.5 epoch at the second stage. The training data of different language pairs are sampled following that of XLM-R: $q_i = p_i^\beta / \sum_j p_j^\beta$, where $\beta = 0.2$ and $p_j$ is the percentage of each language in the training dataset.

[4] https://www.statmt.org/wmt19/translation-task.html

[5] https://ufal.mff.cuni.cz/czeng/czeng16

[6] http://www.kecl.ntt.co.jp/icl/lirg/jparacrawl/

[7] http://www.statmt.org/cc-aligned/

[8] https://opus.nlpl.eu/opus-100.php

| Training Stage | Pretraining | | | Finetuning | | |
|---|---|---|---|---|---|---|
| | MTE-Stage 1 | MTE-Stage 2 | TriKD | English | Non-English | All-language |
| Optimizer | AdamW | AdamW | LAMB | LAMB | LAMB | LAMB |
| Peak Learning Rate | 5e-4 | 1e-4 | 1e-2 | 1e-2 | 1e-3 | 1e-2 |
| Batch Size | 32,768$^{\dagger}$ | 32,768$^{\dagger}$ | 16,384 | 1,024 | 512 | 1,024 |
| Warmup Steps | 4,000 | 0 | 500 | 500 | 500 | 500 |
| Epochs | 1 | 0.5 | 15 | 30 | 30 | 10 |

Table 9: Hyperparameters used in the experiments. "$^{\dagger}$" indicates each batch has 32,768 tokens.

Then we perform the TriKD in Section 3.2. We use the LAMB optimizer (You et al., 2020). The learning rate is linearly warmed up to 0.01 within the first 500 steps and then decayed to 0. The batch size is 16,384. The temperature for the ITC loss is initialized as 0.07 and then learned by gradient descent, while the temperature of TTC loss is fixed as 0.07 (Jain et al., 2021). The models are pretrained for 15 epochs when the smaller dataset CC3M is used, while for 3 epochs when CC12M is used.

When finetuning mCLIP on the downstream image-text retrieval dataset, we use the contrastive loss to finetune the projectors while keeping other parameters frozen. When one training image has multiple captions, all its paired captions are treated as positives.

For all experiments in all pretraining and finetuning stages, we use the inverse square root learning rate scheduler and conduct experiments on 8 NVIDIA V100 GPUs. We use the same dropout method as XLM-R (Conneau et al., 2020). The dropout ratios are set as 0.3. The detailed hyperparameters of different stages are listed in Table 9.

## B More Experimental Results

### B.1 Comparison with Translate-test Method

The translate-test (Conneau et al., 2020; Ni et al., 2021) is another possible method for the multilingual cross-modal retrieval task. It first translates non-English texts into English and then completes the cross-modal retrieval task with an English vision-language pretrained model like CLIP. In this part, we compare mCLIP+ and the translate-test baseline (CLIP+TrTest) on the non-English languages of MSCOCO and Multi30K test sets under the zero-shot setting. For CLIP+TrTest, the captions are translated with the open-sourced m2m-100[9], a recent strong NMT model that is trained with 7.5 billion parallel sentences. The translations are generated with beam size 5 using the 1.2B model checkpoint.

We compare the latency on both the text-to-image and image-to-text retrieval tasks. For the text-to-image retrieval task, we precompute the image features and report the inference time for one text query, which contains (1) the time to extract the feature of one text query, (2) the time of similarity calculation and ranking, and (3) for CLIP+TrTest, the time to translate the non-English text query into English. Similarly, for the image-to-text retrieval task, we precompute the text features and report the inference time for one image query, which contains (1) the time to extract the feature of one image query, and (2) the time of similarity calculation and ranking. Note that for CLIP+TrTest, the translation cost is not included in this latency as the text features are precomputed. All latency values are averaged on all test sets over ten runs using one NVIDIA V100 32G GPU.

From Table 10, mCLIP+ achieves 7.9% better mean recall score than CLIP+TrTest, with fewer model parameters and lower latency. Directly designing a multilingual cross-modal retrieval model like ours is more practical than the translate-test method: (1) The translate-test method has to deploy an additional NMT system, which introduces the storage and computation overhead. Although CLIP+TrTest can be improved with better translations, it usually comes with the cost of larger NMT models and longer latencies. (2) For the translate-test method, the translation process of every text query has to go through the encoder and decoder of the NMT model. The decoding process is usually in an autoregressive manner, which brings non-negligible computing overhead and latency. However, the text query of mCLIP+ only goes through a multilingual encoder, which is computed parallelly. (3) The translate-test method is not suitable for the image-to-text retrieval task. To

[9] https://github.com/facebookresearch/fairseq/tree/main/examples/m2m_100

| Model | Mean Recall | | | | | | Params. (M) | Latency (ms) | |
|---|---|---|---|---|---|---|---|---|---|
| | Multi30K | | | MSCOCO | | Avg. | | | |
| | De | Fr | Cs | Ja | Zh | | | i-to-t | t-to-i |
| mCLIP+ | 76.6 | 76.1 | 74.5 | 55.6 | 71.8 | 70.9 | 444 | 16.3 | 17.2 |
| CLIP+TrTest | 72.8 | 73.4 | 70.0 | 45.3 | 67.1 | 65.7 | 1351 | 16.2 | 438.7 |

Table 10: Comparisons of mCLIP+ and CLIP (Translate-test) on zero-shot cross-modal retrieval tasks. 'Params.' is the total number of model parameters required for inference. 'Latency' is the average inference time for one query.

apply the translate-test method, every non-English image description has to be translated into English and stored in databases, which is infeasible for real billion-level cross-modal retrieval applications.

### B.2 Comparison with English CLIP.

For zero-shot English retrieval on MSCOCO, the mean recall of the original CLIP (ViT-B/32) model is 60.4 (Radford et al., 2021). On the other hand, though mCLIP gains zero-shot cross-lingual transferability on the non-English image-text retrieval, from Table 1, its mean recall score on English is only 53.2, accounting for only 87.2% of the original CLIP's performance. This comparison reveals that one limitation of mCLIP is that the proposed method may slightly degrade the performance on the English cross-modal retrieval task. This performance degradation can be alleviated by using more pretraining data, i.e., mCLIP+ trained with more parallel corpus and multilingual image-text pairs retains 93.8% English retrieval ability of the original CLIP. In addition, since the image backbone of mCLIP initialized from CLIP is frozen during training, one can store an additional CLIP text encoder for English image-text retrieval task when the storage is allowed.

## ACL 2023 Responsible NLP Checklist

### A For every submission:

☑ A1. Did you describe the limitations of your work?
*Section 7*

☑ A2. Did you discuss any potential risks of your work?
*Section 8*

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

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

### B ☑ Did you use or create scientific artifacts?

*Section 4 and Section A of Appendix*

☑ B1. Did you cite the creators of artifacts you used?
*Section 4 and Section A of Appendix*

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
*They are widely used open-sourced data and tools.*

☒ 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)?
*The use is consistent.*

☒ 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 follow the previous works.*

☒ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
*They are covered in the original papers.*

☒ 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.
*They are covered in the original papers.*

### C ☑ Did you run computational experiments?

*Section 4 and Section B of Appendix*

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

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
*Section 4 and Section A of Appendix*

☒ 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?
*The computation cost of one experiment run is high. We follow previous work to report experimental results.*

□ 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.)?
*Not applicable. Left blank.*

**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.?
*No response.*

□ 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)?
*No response.*

□ 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?
*No response.*

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
*No response.*

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
*No response.*