Generalizing Multimodal Pre-training into Multilingual via Language Acquisition

Liang Zhang [1] **Anwen Hu** [1] **Qin Jin** [1]

Abstract

English-based Vision-Language Pre-training (VLP) has achieved great success in various downstream tasks. Some efforts have been taken to generalize this success to non-English languages through Multilingual Vision-Language Pre-training (M-VLP). However, due to the large number of languages, M-VLP models often require huge computing resources and cannot be flexibly extended to new languages. In this work, we propose a **M**ulti**L**ingual **A**cquisition (MLA) framework that can easily generalize a monolingual Vision-Language Pre-training model into multilingual. Specifically, we design a lightweight language acquisition encoder based on state-of-the-art monolingual VLP models. We further propose a two-stage training strategy to optimize the language acquisition encoder, namely the Native Language Transfer stage and the Language Exposure stage. With much less multilingual training data and computing resources, our model achieves state-of-the-art performance on multilingual image-text and video-text retrieval benchmarks.

1. Introduction

We are living in a multimodal and multilingual world w. The information we receive in our daily lives may come from different modalities and languages. Therefore, building multimodal and multilingual models to effectively understand such information has attracted much research attention (Gella et al., 2017; Wehrmann et al., 2019; Kim et al., 2020; Burns et al., 2020). Recently, Multilingual Vision-Language Pre-training (M-VLP) achieves convincing performance in various cross-lingual cross-modal tasks such as multilingual image-text retrieval (Ni et al., 2021; Zhou et al., 2021; Fei et al., 2021; Huang et al., 2021; Jain et al., 2021) and multimodal machine translation (Song et al., 2021). As shown in Figure 1(a), M-VLP models handle multiple languages and modalities simultaneously during pre-training. Despite their successes, M-VLP models suffer from two problems. First, pre-training on vision and multilingual data consumes huge computing resources. For example, the state-of-the-art M-VLP model MURAL (Jain et al., 2021) is pre-trained on 128 Cloud TPUv3 for four days. It could support multimodal tasks on 100+ languages. However, considering there are 6,900+ languages worldwide (Zhou et al., 2021), building such a single model to handle all languages will be highly expensive. Second, M-VLP models cannot be flexibly extended to new languages. Additional training is required for M-VLP models to achieve satisfactory performance on a new language. However, this training process will cause performance degeneration of M-VLP models on the original languages due to the limited model capacity. For example, the limited model capacity even results in M-VLP models performing worse than their monolingual counterparts on English (Ni et al., 2021; Zhou et al., 2021).

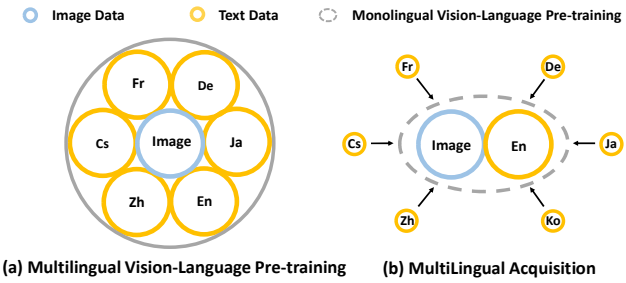

Figure 1: Comparison of data usage between M-VLP and MLA. The size of a circle reflects the amount of training data. M-VLPs learn on vision-language data from multiple languages simultaneously. Instead, MLA generalize monolingual VLP into multilingual on much less training data.

To build multimodal and multilingual models with low-cost and high-flexibility, we refer to our human learning habits when acquiring new languages. We humans normally learn our native language during childhood and practice it through interactions with the multimodal living environments. When learning a new language, we humans initially tend to align it with the native language, as we can easily map words in the native language to real-world objects and concepts. After having a certain language foundation, we could further master it by interacting with the environment directly using the new language. This is known as the language exposure (Castello, 2015). The whole learning process rarely degrades our native language capability.

Inspired by this, we propose a new framework, **M**ulti**L**ingual **A**cquisition (MLA), which constructs multimodal and multilingual models based on monolingual VLPs. The topology of the MLA-based multimodal and multilingual model is illustrated in Figure 1(b). Unlike M-VLPs, which handle data from multiple languages and modalities in a single model, MLA generalizes monolingual VLPs into multilingual using much less training data through a language acquisition encoder. The language acquisition encoder is realized by inserting our proposed lightweight language acquirers into the pre-trained monolingual encoder of the VLP model. During training, original parameters in the pre-trained monolingual encoder are fixed, only multilingual embeddings and language acquirers for each new language are optimized. Following the human learning habits, we propose a two-stage training strategy to train the language acquisition encoder. In the Native Language Transfer (NLT) stage, the model is optimized to establish the correspondence between the new languages with the native language. In the Language Exposure (LE) stage, the model is optimized to build cross-modal alignment between new languages and images. We apply our proposed MLA to the monolingual VLP model CLIP (Radford et al., 2021) and achieve state-of-the-art results on both multilingual image-text and video-text retrieval benchmarks with much less training data and computing resources. Ablation studies demonstrate the effectiveness of our training strategy. Owing to the independence merit of the language acquirers, the MLA-based models can be easily extended to new languages without compromising the performance of their original languages. The main contributions of our work are as follows:

- We propose a lightweight MultiLingual Acquisition (MLA) framework that can easily generalize monolingual VLPs into multilingual.
- We propose a two-stage training strategy to optimize the MLA-based models inspired by the language learning habits of humans. Ablation studies prove the effectiveness of the strategy.
- We apply MLA to the monolingual VLP model CLIP and achieve the new state-of-the-art results on both multilingual image-text and video-text retrieval benchmarks with much less training data and parameters.

2. Related Work

Vision-Language Pre-training: There are increasing interest in building Vision-Language Pre-training (VLP) models. From the perspective of how to interact between vision and language modalities, existing models can be divided into two categories: single-stream and dual-stream models. The single-stream models perform interaction on image and text directly with a cross-modal transformer (Chen et al., 2020; Li et al., 2020b; Kim et al., 2021). In contrast, the dual-stream models encode image and text with two independent encoders and optimize via simple objectives like image-text contrastive learning (Radford et al., 2021; Jia et al., 2021; Yuan et al., 2021). Compared with the single-stream models, the dual-stream models are more efficient to utilize noisy image-text data harvested from the web (Huo et al., 2021), and thus achieve better performance and transferability across downstream tasks. Meanwhile, the dual-stream models are more flexible for extension. Since the dual-stream models process images and text through independent encoders, we can fix the vision encoders and focus on extending the text encoders to support new languages. Therefore, we focus on generalizing dual-stream VLPs into multilingual in this work.

Multilingual Vision-Language Pre-training: To achieve both multilingual and multimodal capability, many works try to learn the relationship between multiple languages and modalities simultaneously through pre-training. M^3P (Ni et al., 2021) introduces the multimodal code-switched training method to enhance multilingual transferability. UC^2 (Zhou et al., 2021) augments the English image-text data to other languages through machine translation and proposes MRTM and VTLM objectives to encourage fine-grained alignment between images and multiple languages. More recently, MURAL (Jain et al., 2021) adopts the dual-stream structure. It is pre-trained with image-text and text-text contrastive objectives on multilingual image-text pairs and translation pairs. M-VLP models significantly outperform previous non-pretraining models (Gella et al., 2017; Wehrmann et al., 2019; Kim et al., 2020; Burns et al., 2020) on multilingual image-text retrieval. Despite their success, these models typically consume huge computing resources and large-scale multilingual training data. Moreover, they fail to take full advantage of the cross-modal knowledge learnt in monolingual VLP, and building cross-modal cross-lingual representations from scratch can be very hard. In contrast, our MLA framework aims to generalize VLP models into multilingual and it builds multimodal and multilingual models with much less data and computing cost.

Multilingual Extension: Some works explore making pre-trained monolingual language models multilingual. Reimers et al. extend sentence embeddings from monolingual to multilingual by Multilingual Knowledge Distillation (MKD) (Reimers & Gurevych, 2020). Given translation pairs, MKD optimizes the multilingual student model to produce similar sentence embeddings with the monolingual teacher model. Artetxe et al. extend monolingual models by training additional word embeddings (Artetxe et al., 2020). MAD-X (Pfeiffer et al., 2020) extends multilingual pre-training models to support low-resource languages through adapters (Houlsby et al., 2019). By extending state-of-the-art pre-trained language models, these works have achieved impressive results in NLP tasks such as bitext retrieval (Reimers

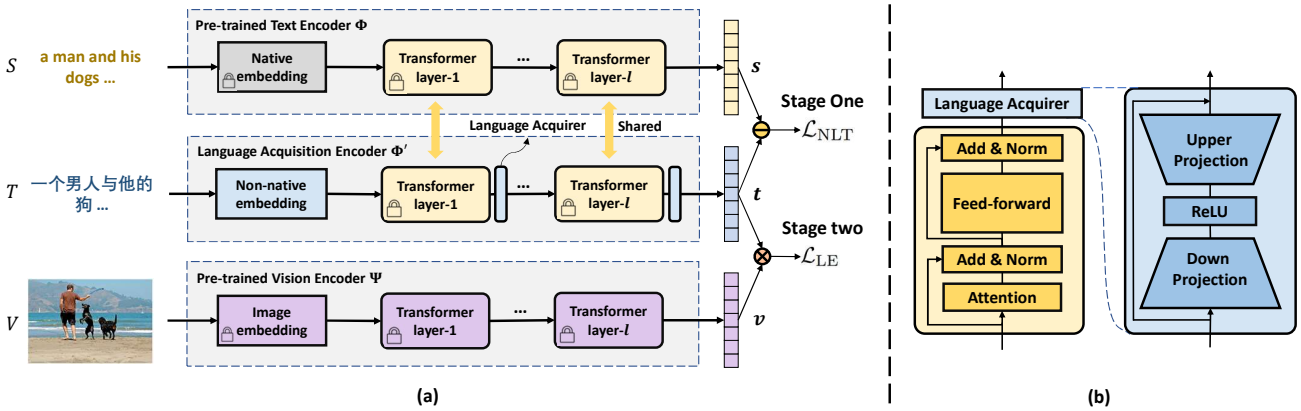

Figure 2: Model illustration: (a) The overview of MLA framework. (b) The structure of a language acquirer

& Gurevych, 2020), cross-lingual QA and NER (Pfeiffer et al., 2020; Reimers & Gurevych, 2020). However, few works focus on making VLP models multilingual. Work in (Pfeiffer et al., 2021) is the first to extend single-stream VLP model OSCAR (Li et al., 2020b). It adopts a similar strategy with MAD-X (Pfeiffer et al., 2020) that trains language adapters with Masked Language Modeling (MLM) for each language. During inference, it replaces the English language adapters with the target language adapters to achieve zero-shot cross-lingual transfer. However, it generalizes poorly on other languages since the MLM-based training strategy can only implicitly establish the correspondence between other languages and English, let alone vision correspondences. In contrast, MLA directly builds the connection of other languages with English and then with vision in the two-stage training strategy. Therefore, MLA achieves comparable results on non-English languages as on English in downstream tasks.

3. Method

The MultiLingual Acquisition (MLA) framework is proposed to empower a dual-stream monolingual VLP model with multilingual capability. We define the *native language* of a VLP as its pre-training language. In this paper, we choose CLIP-ViT-B (Radford et al., 2021) as the VLP model. It is pre-trained with 400M image-text pairs in English (Radford et al., 2021). Note that MLA can also be applied to VLP models with different native languages.

Since the state-of-the-art VLP models can project vision and native language into a shared multimodal space, we design a language acquisition encoder to process non-native languages. We then simulate the learning habits of human beings and propose a two-stage training strategy to optimize the language acquisition encoder. We first introduce the architecture of the MLA framework in Sec.3.1. Then, we describe our training strategy in Sec.3.2.

3.1. Architecture

Figure 2(a) illustrates the overview of the MLA framework, which consists of three modules: the pre-trained text encoder, the pre-trained vision encoder, and the language acquisition encoder.

Pre-trained Text Encoder. Given a sentence S in the native language, the corresponding sentence representation $\boldsymbol{s} = \Phi(S;\theta_\Phi)$ is generated through the pre-trained text encoder Φ. To preserve the cross-model knowledge of VLP, θ_Φ is keep fixed during training. As shown in the top part of Figure 2(a), the pre-trained text encoder contains a native embedding block and l transformer layers (Vaswani et al., 2017). The native embedding block first tokenizes S with byte pair encoding (BPE) (Sennrich et al., 2016). Then, it converts words into embeddings $E_S = [e_{0=\texttt{[SOS]}}, e_1, \ldots, e_{M=\texttt{[EOS]}}]$. `[SOS]` and `[EOS]` are special tokens denoting the boundary of S. The word embeddings are then passed through the transformer layers:

$$H^0 = [e_{0=\texttt{[SOS]}}, e_1, \ldots, e_{M=\texttt{[EOS]}}] + E_{pos} \quad (1)$$

$$H^i = \texttt{TransformerLayer}(H^{i-1};\theta^i_\Phi) \quad (2)$$

where $H^i = [h^i_0, \ldots, h^i_M]$ is the hidden state of the layer i. θ^i_Φ denotes the parameters of the layer i. E_{pos} is the positional encoding. Note that the causal self-attention mask is used in the transformer layers (Radford et al., 2021). The last hidden state of the `[EOS]` token is chosen to generate the sentence representation:

$$\boldsymbol{s} = W_a h^l_M \quad (3)$$

where $\boldsymbol{s}$ is the sentence representation of S, and W_a denotes a linear projection.

Pre-trained Vision Encoder. We extract the representation $\boldsymbol{v} = \Psi(V;\theta_\Psi)$ of an image V with the pre-trained vision encoder Ψ. Similar with the pre-trained text encoder, θ_Ψ is also frozen. The pre-trained vision encoder is implemented as a Vision Transformer (Dosovitskiy et al., 2020).

As shown in the bottom part of Figure 2(a), it consists of a image embedding block and l transformer layers. Given an image V, the image embedding block first divides V into patches $V' = [v'_1, \ldots, v'_N]$ following (Dosovitskiy et al., 2020). Then, they are linearly projected into patch embeddings $E_p = [e_{\texttt{[CLASS]}}, W_p v'_1, \ldots, W_p v'_N]$, where $e_{\texttt{[CLASS]}}$ is a special embedding for the whole image and W_p is the linear projection. The patch embeddings are then fed into transformer layers:

$$Z^0 = [e_{\texttt{[CLASS]}}, W_p v'_1, \ldots, W_p v'_N] + E_{pos} \quad (4)$$

$$Z^i = \texttt{TransformerLayer}(Z_{i-1}; \theta^i_V) \quad (5)$$

where $Z^i = [z^i_0, \ldots, z^i_N]$ is the hidden state of the layer i. The last hidden state of the `[CLASS]` embedding z^l_0 is selected to produce the representation of image V:

$$\boldsymbol{v} = W_b z^l_0 \quad (6)$$

where $\boldsymbol{v}$ is the image representation of V, and W_b denotes a linear projection.

Language Acquisition Encoder. As shown in the middle part of Figure 2(a), the language acquisition encoder is built upon the pre-trained text encoder. Suppose T is a sentence written in a non-native language L, we get the representation of T through language acquisition encoder $\boldsymbol{t} = \Phi'(T; \theta_\Phi, \theta_{emb}, \theta_L)$, where θ_Φ are fixed parameters of the pre-trained text encoder, θ_{emb} refers to a shared non-native embedding block and θ_L represents specialized language acquirers for language L. Non-native sentence T is first tokenized and processed into word embeddings $E_T = [u_{0=\texttt{[SOS]}}, \ldots, u_{M=\texttt{[EOS]}}]$ through the non-native embedding block. The word embeddings are then encoded through the pre-trained transformer layers and language acquirers:

$$X^0 = [W_e u_{0=\texttt{[SOS]}}, W_e u_1, \ldots, W_e u_{m=\texttt{[EOS]}}] + E_{pos} \quad (7)$$

$$H^i = \texttt{TransformerLayer}(X^{i-1}; \theta^i_\Phi) \quad (8)$$

$$X^i = \texttt{LA}(H^i; \theta^i_L) \quad (9)$$

where $X^i = [x^i_0, \ldots, x^i_m]$ is the hidden state of the layer i. W_e is a linear projection to keep dimension consistency. θ^i_L denotes the parameters of the i-th language acquirer for language L. Note that each non-native language has independent language acquirers, and all of them share the same word embedding block. As shown in Figure 2(b), the language acquirer is implemented as a bottleneck MLP with residual connection (He et al., 2016):

$$\texttt{LA}(X) = W_{upper}\texttt{ReLU}(W_{down}X) + X \quad (10)$$

Similar with the pre-trained text encoder, the last hidden state of the `[EOS]` token is projected into the sentence representation $\boldsymbol{t}$:

$$\boldsymbol{t} = W_a x^l_m \quad (11)$$

Note that Eq.11 shares the same linear projection W_a with Eq.3. The main advantage of the language acquisition encoder is that it can extend the VLP models to support new languages without influencing the existing languages, as it handles different languages with independent language acquirers.

3.2. Training Strategy

To simulate the language learning habits of humans, we optimize the model in two stages: the Native Language Transfer (NLT) stage and the Language Exposure (LE) stage.

Native Language Transfer. When learning a new language, we humans initially tend to align it with the native language. To simulate this learning phase, we align the non-native representations to the native representations during the Native Language Transfer (NLT) stage. Specifically, suppose $\{(S_1, T_1), ..., (S_n, T_n)\}$ are translation pairs, where S_i is in the native language, and T_i is in a non-native language L. The objective in the NLT stage is minimizing the Mean Square Error (MSE) between the native representation $\boldsymbol{s}_i = \Phi(S_i; \theta_\Phi)$ and the non-native representation $\boldsymbol{t}_i = \Phi'(T_i; \theta_\Phi, \theta_L, \theta_{emb})$:

$$\mathcal{L}_{\text{NLT}} = \frac{1}{B}\sum_{i=1}^{B} \|\boldsymbol{s}_i - \boldsymbol{t}_i\|^2 \quad (12)$$

where B is the batch size. Note that θ_Φ is loaded from the VLP model and is kept frozen. θ_L is trained for non-native language L. θ_{emb} is shared among non-native languages.

During the NLT stage, the non-native correspondence with vision can be built pivoting on the native language, since the correspondence between the native language and vision is well established through VLP.

Language Exposure. After the NLT stage, the model has built an implicit connection between non-native languages and vision. However, due to the existence of synonyms, two same words in the native language may correspond to different images. Thus, ambiguity may arise when learning non-native languages solely by relying on the native language. Actually, we can regard the language acquisition encoder after the NLT stage as a person with a certain language foundation. He/She has learned the basic usage of a language through native language teaching. To master it, he/she may practice the non-native language by interacting with the multimodal living environments. Inspired by this learning phase, we directly establish the cross-modal alignment between non-native languages and vision during the Language Exposure (LE) stage. Given image-text pairs $\{(V_1, T_1), ..., (V_n, T_n)\}$ where T_i is

in a non-native language L, the sentence representation $\boldsymbol{t}_i = \Phi^{'}(T_i; \theta_\Phi, \theta_L, \theta_{emb})$ should be closer to the aligned image representation $\boldsymbol{v}_i = \Psi(V_i; \theta_\Psi)$, and away from the misaligned one $\boldsymbol{v}_j = \Psi(V_j; \theta_\Psi), j \neq i$. This can be achieved by performing contrastive learning between non-native languages and images. For a non-native sentence T_i, we treat the corresponding image V_i as a positive sample, and other images in the same batch $V_j, j \neq i$ as negative samples. Vice versa for images. The objective in the LE stage is minimizing the NCE loss defined as follows:

$$\mathcal{L}_{\text{LE}} = \frac{1}{2}(\mathcal{L}_{v2t} + \mathcal{L}_{t2v}) \tag{13}$$

$$\mathcal{L}_{v2t} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(\text{sim}(\boldsymbol{v}_i, \boldsymbol{t}_i)/\tau)}{\sum_{k=1}^{N}\exp(\text{sim}(\boldsymbol{v}_i, \boldsymbol{t}_k)/\tau)} \tag{14}$$

$$\mathcal{L}_{t2v} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(\text{sim}(\boldsymbol{v}_i, \boldsymbol{t}_i)/\tau)}{\sum_{k=1}^{N}\exp(\text{sim}(\boldsymbol{v}_k, \boldsymbol{t}_i)/\tau)} \tag{15}$$

where B is the batch size. $\text{sim}(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x}^\top \boldsymbol{y}}{\|\boldsymbol{x}\|\|\boldsymbol{y}\|}$ is the cosine similarity between two vectors. τ is a temperature hyper-parameter to scale the logits. Note that though the image-to-text loss $\mathcal{L}_{v2t}$ is optimized, the pre-trained vision encoder is kept frozen during training. Similar to NLT, the trainable parameters in LE come from the language acquirers and the non-native embedding block.

4. Experiments

In this section, we first introduce the datasets used in this paper, and then present detailed experiments to evaluate the proposed MLA framework.

4.1. Dataset Description

We train our model with the Conceptual Captions (CC) dataset (Sharma et al., 2018) and two translation enhanced versions of the CC (Zhou et al., 2021; Carlsson, 2021). We use Multi30K (Elliott et al., 2016), MSCOCO (Chen et al., 2015; Li et al., 2019; Yoshikawa et al., 2017) and XTD (Aggarwal & Kale, 2020) for multilingual image-text retrieval evaluation, and MSRVTT (Xu et al., 2016; Huang et al., 2021) for multilingual video-text retrieval evaluation.

Conceptual Captions (CC) (Sharma et al., 2018) contains 3.3 million image-text pairs in English crawled from the Web[1]. We also randomly select 300K image-text pairs denoted as **CC300K** for training our model to show the low-cost merit of MLA. For multilingual sentences, we leverage two translation augmented CC datasets: (1) **CC6L** (Zhou et al., 2021) that translates all English captions of the CC into 5 languages (German(de), French(fr), Czech(cs), Chinese(zh)); and (2) **CC69L** (Carlsson, 2021) that contains 27K captions in each of the 68 languages translated from English.[2] Considering the languages of the downstream datasets, we train the model with CC6L for multilingual image-text retrieval, and with CC69L for multilingual video-text retrieval.

Multi30K (Elliott et al., 2016) is built upon Flickr30K (Young et al., 2014). The English(en) captions are manually translated into German(de), French(fr) and Czech(cs). It contains 31K images paired with 5 captions per image in English and German, and 1 caption in French and Czech. We use the standard train, dev and test splits defined in (Young et al., 2014).

MSCOCO (Chen et al., 2015) contains 123K images with 5 English captions per image. (Yoshikawa et al., 2017) annotates 5 Japanese captions per image, and (Li et al., 2019) extends MSCOCO with Chinese captions for 20K images. We follow the standard train, dev and test splits for English and Japanese as in (Karpathy & Fei-Fei, 2015). For Chinese, we can only perform zero-shot evaluation on the test split defined in (Li et al., 2019), as the full splits have overlaps with English and Japanese splits.

XTD (Aggarwal & Kale, 2020) provides captions in 11 languages (English(en), German(de), French(fr), Chinese(zh), Japanese(ja), Italian(it), Spanish(es), Russian(ru), Polish(pl), Turkish(tr), Korean(ko)) for 1K MSCOCO images. Except for Japanese, all non-English captions are translated from the English caption directly. We use this dataset for zero-shot image-text retrieval evaluation only.

MSRVTT (Xu et al., 2016) is a video caption dataset with 10K videos, where each video is annotated with 20 English captions. Huang et al. translates the English captions into 8 languages (German(de), French(fr), Russian(ru), Spanish(es), Czech(cz), Swahili(sw), Chinese(zh) and Vietnamese(vi)) via machine translation service (Huang et al., 2021). We follow the standard train/dev splits in (Xu et al., 2016), and evaluate on the 1K test split as described in (Yu et al., 2018).

4.2. Implementation Details

We apply MLA on two VLP models: CLIP-ViT-B-32 and CLIP-ViT-B-16 (Radford et al., 2021), denoted as MLA_{CLIP} and $\text{MLA}_{\text{CLIP16}}$ respectively. The hidden dimension of the language acquirers is set to 256, and all language acquirers for each non-native language cost only 3.14 MB parameters. The non-native embedding matrix is initialized with M-BERT (Devlin et al., 2019). It costs 92.2 MB and shared with all non-native languages. We train two separate models for multilingual image-text retrieval and video-text retrieval. For the image model, we train with CC6L (Zhou et al., 2021). For the video model, we use multilingual captions from CC69L (Carlsson, 2021). For both models, we optimize multiple language acquirers iteratively

[1] We can only access ∼2.5 million images due to broken URLs.

[2] We remove captions of unaccessible images, leaving ∼20K captions for each language.

	Method	Training Data	Multi30K				MSCOCO 1K		MSCOCO 5K	
			en	de	fr	cs	en	ja	en	ja
Zero-shot	Unicoder-VL	CC3M (English only)	72.0	-	-	-	63.7	-	-	-
	ALIGN	AT-en (English only)	84.3	-	-	-	80.0	-	60.6	-
	M^3P	CC3M+Wiki	57.9	36.8	27.1	20.4	63.1	33.3	-	-
	UC^2	TrTrain(CC3M)	66.6	62.5	60.4	55.1	70.9	62.3	-	-
	MURAL	TrTrain(CC12M)+EOBT	80.9	76.0	75.7	68.2	78.1	72.5	58.0	49.7
	MURAL†	AT+MBT	82.4	76.2	75.0	64.6	79.2	73.4	59.5	54.4
	MLA_{CLIP}	TrTrain(CC300K)	84.4	78.7	77.7	70.8	79.4	74.9	60.5	54.1
	MLA_{CLIP16}	TrTrain(CC300K)	**86.4**	**80.8**	**80.9**	**72.9**	**80.9**	**76.7**	**62.6**	**57.0**
FT-En	M^3P	CC3M+Wiki	87.4	82.1	67.3	65.0	88.6	56.0	-	-
	UC^2	TrTrain(CC3M)	87.2	83.8	77.6	74.2	88.1	71.7	-	-
	MLA_{CLIP}	TrTrain(CC300K)	92.0	82.6	85.1	76.2	89.3	80.4	75.7	62.1
	MLA_{CLIP16}	TrTrain(CC300K)	**94.5**	**86.4**	**87.3**	**79.5**	**91.3**	**82.6**	**79.4**	**65.5**
FT-All	M^3P‡	CC3M+Wiki	87.7	82.7	73.9	72.2	88.7‡	87.9‡	-	-
	UC^2‡	TrTrain(CC3M)	88.2	84.5	83.9	81.2	88.1‡	87.5‡	-	-
	MURAL	TrTrain(CC12M)+EOBT	91.0	87.3	86.4	82.4	89.4	87.4	73.7	71.9
	MURAL†	AT+MBT	92.2	88.6	87.6	84.2	88.6	88.4	75.4	74.9
	MLA_{CLIP}	TrTrain(CC300K)	92.0	86.8	85.4	82.3	89.3	88.1	75.7	73.2
	MLA_{CLIP16}	TrTrain(CC300K)	**94.5**	**89.7**	**89.2**	**85.9**	**91.3**	**90.4**	**79.4**	**76.5**

Table 1: Multilingual image-text retrieval results on Multi30K and MSCOCO. TrTrain: Translate-train, FT-En: *Fine-tune on English*, FT-All: *Fine-tune on All*. †: Models trained with publicly unavailable datasets. ‡: Models fine-tuned on COCO-CN (Li et al., 2019), which has an overlap train split with the test split of English and Japanese. Best results are in bold and second best are underlined.

Method	Trainable Params	Computing Costs
M^3P	566 M	4×V100×7d
UC^2	478 M	8×V100×4d
MURAL	300 M	128×TPUv3×4d
Ours (MLA_{CLIP})	108 M	1×V100×0.5d

Table 2: Comparison of trainable parameters and computing costs between MLA and M-VLPs.

with a batch size of 128. The NLT stage performs 117,150 steps with a learning rate of 1e-4, and the LE stage performs 11,715 steps with a learning rate of 3e-6. The temperature τ is set to 0.01. For both stages, we use the Adam optimizer (Kingma & Ba, 2015) with a linear warm-up for the first 10% of steps. The whole training process takes about 12 hours to converge on 1 Nvidia V100 GPU.

4.3. Evaluation on Multilingual Image-Text Retrieval

In multilingual image-text retrieval, models are given a sentence in a certain language to find the most semantically relevant image from an image database and vice versa. We compare our model with state-of-the-art multilingual vision-language pre-training methods under three settings:

- ***Zero-shot:*** we directly evaluate the model without fine-tuning on downstream datasets.
- ***Fine-tune on English:*** we first fine-tune the VLP model on downstream English data. We then insert the language acquirers and non-native embedding block into the fine-tuned model and evaluate on other languages directly.
- ***Fine-tune on All:*** after ***Fine-tune on English***, we fine-tune the language acquirers and non-native embedding block and freeze other parts of the model.

Following previous works (Ni et al., 2021; Zhou et al., 2021; Jain et al., 2021), we report Average Recall (AR), which is the average score over Recall@1, Recall@5, and Recall@10 on two retrieval directions (image→text, text→image). The results are shown in Table 1. Also, the comparison of computing costs and parameters can be found in Table 2.

Under the ***Zero-shot*** setting, we observe that MLA_{CLIP} performs significantly better than state-of-the-art M-VLP models on English. This is because MLA_{CLIP} could completely maintain the strong English performance of CLIP. In contrast, M-VLP models typically perform worse than their monolingual counterparts on English (M^3P 57.9 vs. Unicoder-VL (Li et al., 2020a) 72.0, MURAL 80.9 vs. ALIGN (Jia et al., 2021) 84.3). MLA_{CLIP} also outperforms M-VLP models on other languages. For example, MLA_{CLIP} achieves 78.7 average recall score on German, outperforming MURAL by 2.7%. Note that the pre-training dataset of MURAL contains 12 million image-text pairs for each language, while MLA_{CLIP} only uses 300K training image-text pairs. It demonstrates that MLA is a high-data-efficient method to empower monolingual VLP models with multilingual capability. Under the ***Fine-tune on English*** setting, MLA shows strong cross-lingual transfer capability. Under the ***Fine-tune on All*** setting, MLA_{CLIP} performs slightly worse than MURAL which was pre-trained on publicly unavailable dataset AT+MBT (Jain et al., 2021). We consider the reason is that MURAL has more trainable parameters than MLA_{CLIP} (300M vs 108M, as shown in Table 2) for fine-tuning, which makes it easier to fit the downstream datasets with a certain scale such as Multi30K and MSCOCO. MLA_{CLIP16} achieves state-of-the-art re-

	Method	en	de	fr	cs	zh	ru	vi	sw	es	mean
ZS	Ours(MLA_{CLIP} w/o LE)	**30.8**	18.3	18.9	14.5	**18.6**	12.6	7.2	10.2	19.3	16.7
	Ours(MLA_{CLIP})	**30.8**	**20.1**	**22.0**	**15.7**	18.3	**14.4**	**8.2**	**10.7**	**20.2**	**17.8**
FT-En	XLM-R-MMP (Huang et al., 2021)	23.8	19.4	20.7	19.3	18.2	19.1	8.2	8.4	20.4	17.5
	Ours(MLA_{CLIP})	**42.5**	**26.1**	**26.7**	**20.5**	**25.3**	**18.9**	**12.9**	**12.6**	**27.2**	**23.6**
FT-All	XLM-R-MMP (Huang et al., 2021)	23.1	21.1	21.8	20.7	20.0	20.5	10.9	14.4	21.9	19.4
	Ours(MLA_{CLIP})	**42.5**	**33.1**	**34.5**	**30.5**	**31.6**	**28.9**	**16.9**	**24.3**	**33.5**	**30.6**

Table 3: Multilingual video-text retrieval results on MSRVTT. ZS: *Zero-shot*, FT-En: *Fine-tune on English*, FT-All: *Fine-tune on All*.

sults on all languages under three settings. It indicates that if stronger VLP models such as ALIGN-L2 (Jia et al., 2021) or Florence (Yuan et al., 2021) are provided, better performance on multilingual image-text retrieval could be reached through MLA.

4.4. Evaluation on Multilingual Video-Text Retrieval

In multilingual video-text retrieval, the model searches for the most semantically relevant videos given a text query in a certain language. Following (Luo et al., 2021), we first uniformly sample 12 frames from each video, and use the pre-trained vision encoder to extract representations for each frame. We then perform mean pooling over frame representations to get the video representation.

We also evaluate the models under three settings as in Sec.4.3. We report the text→video Recall@1 score in Table 3. Under ***Zero-shot*** setting, MLA_{CLIP}, which is trained on CC69L without using any video data, achieves comparable or even better results than the fine-tuning results of the state-of-the-art M-VLP model XLM-R-MMP (Huang et al., 2021) on several languages (de: 20.1 vs. 21.1; fr: 22.0 vs. 21.8; es: 20.2 vs. 21.9). Under the ***Fine-tune on English*** and ***Fine-tune on All*** settings, MLA_{CLIP} also outperforms XLM-R-MMP significantly. We consider the convincing performance comes from two reasons: 1) CLIP is a strong VLP model that can generalize well on video data. 2) The proposed MLA framework can well transfer the open-domain knowledge learned by CLIP to other languages. These results suggest that MLA could maintain the open-domain capability of the VLP model which generalizes well on different downstream data.

4.5. Ablation Study

A. Training Strategy

We conduct an ablation study in Table 4 to validate the effectiveness of the proposed MLA training strategy. For those settings with NLT and LE at the same stage, we add the loss of the two objectives together during training. By comparing row 1 to row 2&3, we observe that LE at stage one leads to poor performance. This indicates that aligning with the native language is more important for the VLP model to acquire new languages at an early stage. It is consistent with the learning habits of humans. By comparing row 1 and row 4, we see that LE at stage two could bring improvements on the new languages. Additionally, comparing row 4 and row 5 suggests that optimizing the model with NLT and LE together at stage two does not bring improvements.

Row	Stage one		Stage two		Multi30K			MSCOCO 1K	
	NLT	LE	NLT	LE	de	fr	cs	ja	zh
1	✓				76.3	74.2	67.2	72.1	75.7
2		✓			68.2	67.7	58.6	65.9	71.7
3	✓	✓			71.1	69.7	59.8	67.6	73.9
4	✓			✓	**78.7**	**77.7**	**70.8**	**74.9**	**78.5**
5	✓		✓	✓	78.4	77.3	69.9	74.2	78.1

Table 4: Ablation study on training strategy

B. Language Acquirers and Embedding Initialization

In order to validate the effectiveness of the proposed Language Acquirers, we remove the language acquirers and the M-BERT embedding initialization from the model respectively and evaluate on zero-shot multilingual image-text retrieval. As shown in Table 5, the performance on all languages drops significantly without language acquirers. Meanwhile, initializing the embedding with M-BERT (Devlin et al., 2019) only brings incremental improvements. It indicates that the language acquirers contribute most to the performance, and MLA does not depend much on the initialization of non-native embedding.

Methods	Multi30K			MSCOCO 1K	
	de	fr	cs	ja	zh
MLA_{CLIP}	**78.7**	**77.7**	**70.8**	**74.9**	**78.5**
MLA_{CLIP} w/o LA	76.1	74.9	65.7	70.3	76.5
MLA_{CLIP} w/o EI	77.9	76.2	69.4	74.6	78.1

Table 5: Ablation study on language acquirers and embedding initialization. LA: Language Acquirers, EI: M-BERT Embedding Initialization

C. Low-resource Languages

Image-text pairs may be rare for low-resource languages. To explore the performance of MLA under this situation, we further simulate a **low-resource scenario** using XTD dataset. We finetune MLA_{CLIP} and UC^2 (pre-trained on CC6L) with small amount of data from XTD in an unseen language. We randomly sample 600 pairs for finetuning, and the remained 400 samples are evenly divided for validation and testing. Korean is chosen to perform simulation as its script and language family are not covered by CC6L.

Row	Method	Seen languages					Unseen languages					
		en	de	fr	zh	ja	it	es	ru	pl	tr	ko
1	UC^2 w/o unseen language training	71.8	67.5	68.4	61.9	51.5	-	-	-	-	-	
2	UC^2 w/ unseen language training	63.6	57.8	57.6	57.6	48.4	56.4	56.2	51.3	56.4	51.62	51.3
3	UC^2 w/ all language training	65.2	59.3	59.7	60.1	50.5	57.7	56.5	50.9	55.3	53.2	50.2
4	MLA_{CLIP} w/o unseen language training	75.9	**72.6**	**72.9**	73.7	**67.2**	-	-	-	-	-	-
5	MLA_{CLIP} w/ unseen language training	**76.0**	**72.6**	**72.9**	**73.8**	**67.2**	**64.7**	**62.8**	**58.1**	**63.0**	**56.5**	**57.3**

Table 6: Language extention experiments on XTD dataset.

Experimental results in Table 7 show that MLA can achieve competitive results with **very small amount of text-text pairs only** (row 2), and adding image-text pairs brings further improvement (row 3). It demonstrates that MLA is still an attractive method for low-resource languages even without any image-text pairs.

	Methods	Data	Training samples 100 / 200 / 600
1	UC^2	Img-Txt	47.0 / 60.1 / 78.3
2	MLA_{CLIP}	Txt-Txt	51.7 / 62.8 / 78.7
3	MLA_{CLIP}	Both	**56.7 / 66.9 / 80.1**

Table 7: Low resource performance on image-Korean retrieval.

D. Amount of Training Data

Multilingual image-text pairs may be rare in practice. To explore the performance of MLA under low-resource conditions, we conduct experiments to control the numbers of image-text pairs used for each language. We train the models with CC6L and evaluate on MSCOCO 1K and Multi30K under the zero-shot setting. The corresponding mean AR over non-English languages (de, fr, cs, ja, zh) are drawn in Figure 3. We observe that MLA performs significantly better than MKD (Reimers & Gurevych, 2020) in all cases. Note that when the amount of training data is small, the advantage of MLA is more obvious, which could outperform MKD even without the LE training stage. Additionally, when training with only 30K image-text pairs per language, MLA outperforms UC^2, which is pre-trained with 3M pairs per language. MLA is thus a data-efficient method to build multilingual and multimodal models.

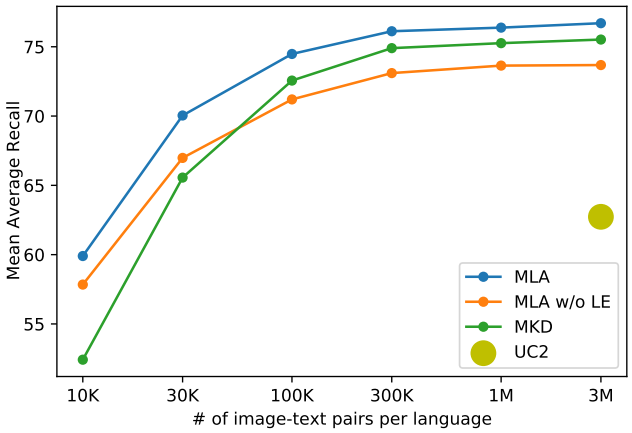

Figure 3: Mean AR vs. number of image-text pairs per language.

E. Language Extensibility

Multilingual models often encounter the need to support new languages that do not occur in the training stage. We conduct language extension experiments to compare MLA_{CLIP} with M-VLP model UC^2 (Zhou et al., 2021) on the XTD dataset (Aggarwal & Kale, 2020). XTD supports 11 languages, and 5 of them (en, de, fr, cs, zh, ja) are seen in the pre-training stage of UC^2, while other 6 languages (it, es, ru, pl, tr, ko) are unseen. To make a fair comparison, we first train MLA_{CLIP} with the same data as UC^2 and then train both of them on unseen languages with CC69L. The zero-shot image-text retrieval results on XTD are shown in Table 6. We observe a significant performance degeneration on the seen languages for UC^2 when training solely with unseen languages (row 1 vs. row 2). Even keep training with the seen languages, the performance is still significantly reduced due to the limited model capacity (row 1 vs. row 3). In contrast, as MLA decoupled multiple languages through acquirers, the performance of the seen languages is rarely affected (row 4 vs. row 5) . This suggests that MLA framework can build multimodal multilingual models that are suitable for supporting increasing numbers of languages.

5. Conclusion

In this paper, we propose the MultiLingual Acquisition (MLA) framework that can generalize monolingual Vision-Language Pre-training models into multilingual with low-cost and high-flexibility. MLA injects language acquirers and a non-native embedding block into VLPs to support new languages. Inspired by the language learning habits of humans, we propose a two-stage training strategy to optimize the language acquirers and non-native embedding block. MLA applied on CLIP achieves state-of-the-art performances on multilingual image-text and video-text retrieval benchmarks with much less computing costs and training data. Extensive ablation studies demonstrate that MLA is a flexible, effective, and efficient method to build multimodal and multilingual models.

References

Aggarwal, P. and Kale, A. Towards zero-shot cross-lingual image retrieval. *CoRR*, abs/2012.05107, 2020.

Artetxe, M., Ruder, S., and Yogatama, D. On the cross-lingual transferability of monolingual representations. In *58th Annual Meeting of the Association for Computational Linguistics*, pp. 4623–4637. Association for Computational Linguistics, 2020.

Burns, A., Kim, D., Wijaya, D., Saenko, K., and Plummer, B. A. Learning to scale multilingual representations for vision-language tasks. In *European Conference on Computer Vision*, pp. 197–213. Springer, 2020.

Carlsson, F. Multilingual clip. https://github.com/FreddeFrallan/Multilingual-CLIP, 2021.

Castello, D. First language acquisition and classroom language learning: Similarities and differences. *ELAL College of Arts & Law*, pp. 1–18, 2015.

Chen, X., Fang, H., Lin, T., Vedantam, R., Gupta, S., Dollár, P., and Zitnick, C. L. Microsoft COCO captions: Data collection and evaluation server. *CoRR*, abs/1504.00325, 2015.

Chen, Y.-C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., and Liu, J. Uniter: Universal image-text representation learning. In *European conference on computer vision*. Springer, 2020.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, 2019.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

Elliott, D., Frank, S., Sima'an, K., and Specia, L. Multi30k: Multilingual english-german image descriptions. In *VL@ ACL*, 2016.

Fei, H., Yu, T., and Li, P. Cross-lingual cross-modal pretraining for multimodal retrieval. In *2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3644–3650, 2021.

Gella, S., Sennrich, R., Keller, F., and Lapata, M. Image pivoting for learning multilingual multimodal representations. In *EMNLP 2017: Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2017.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *IEEE conference on computer vision and pattern recognition*, 2016.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *ICCV*, 2021a.

Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., and Song, D. Natural adversarial examples. In *CVPR*, 2021b.

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., and Gelly, S. Parameter-efficient transfer learning for NLP. In *ICML*, 2019.

Huang, P.-Y., Patrick, M., Hu, J., Neubig, G., Metze, F., and Hauptmann, A. G. Multilingual multimodal pre-training for zero-shot cross-lingual transfer of vision-language models. In *NAACL*, 2021.

Huo, Y., Zhang, M., Liu, G., Lu, H., Gao, Y., et al. Wenlan: Bridging vision and language by large-scale multi-modal pre-training, 2021.

Jain, A., Guo, M., Srinivasan, K., Chen, T., Kudugunta, S., Jia, C., Yang, Y., and Baldridge, J. MURAL: Multimodal, multitask representations across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 3449–3463, 2021.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 4904–4916. PMLR, 2021.

Karpathy, A. and Fei-Fei, L. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

Kim, D., Saito, K., Saenko, K., Sclaroff, S., and Plummer, B. Mule: Multimodal universal language embedding. In *AAAI Conference on Artificial Intelligence*, volume 34, pp. 11254–11261, 2020.

Kim, W., Son, B., and Kim, I. Vilt: Vision-and-language transformer without convolution or region supervision. In *38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*. PMLR, 2021.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *ICLR (Poster)*, 2015.

Krizhevsky, A. Learning multiple layers of features from tiny images. *Master's thesis, University of Tront*, 2009.

Li, G., Duan, N., Fang, Y., Gong, M., and Jiang, D. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 11336–11344, 2020a.

Li, X., Xu, C., Wang, X., Lan, W., Jia, Z., Yang, G., and Xu, J. Coco-cn for cross-lingual image tagging, captioning, and retrieval. *IEEE Transactions on Multimedia*, 21(9): 2347–2360, 2019.

Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*. Springer, 2020b.

Luo, H., Ji, L., Zhong, M., Chen, Y., Lei, W., Duan, N., and Li, T. Clip4clip: An empirical study of clip for end to end video clip retrieval. *arXiv preprint arXiv:2104.08860*, 2021.

Ni, M., Huang, H., Su, L., Cui, E., Bharti, T., Wang, L., Zhang, D., and Duan, N. M3p: Learning universal representations via multitask multilingual multimodal pre-training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3977–3986, 2021.

Pfeiffer, J., Vulić, I., Gurevych, I., and Ruder, S. Mad-x: An adapter-based framework for multi-task cross-lingual transfer. In *2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7654–7673, 2020.

Pfeiffer, J., Geigle, G., Kamath, A., Steitz, J.-M. O., Roth, S., Vulić, I., and Gurevych, I. xgqa: Cross-lingual visual question answering. *arXiv e-prints*, 2021.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In *38th International Conference on Machine Learning*, volume 139, pp. 8748–8763, 2021.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*. PMLR, 2019.

Reimers, N. and Gurevych, I. Making monolingual sentence embeddings multilingual using knowledge distillation. In *2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4512–4525, 2020.

Sennrich, R., Haddow, B., and Birch, A. Neural machine translation of rare words with subword units. In *54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1715–1725, 2016.

Sharma, P., Ding, N., Goodman, S., and Soricut, R. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2556–2565, 2018.

Song, Y., Chen, S., Jin, Q., Luo, W., Xie, J., and Huang, F. Product-oriented machine translation with cross-modal cross-lingual pre-training. In *29th ACM International Conference on Multimedia*, 2021.

Van der Maaten, L. and Hinton, G. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Wehrmann, J., Souza, D. M., Lopes, M. A., and Barros, R. C. Language-agnostic visual-semantic embeddings. In *IEEE/CVF International Conference on Computer Vision*, pp. 5804–5813, 2019.

Xu, J., Mei, T., Yao, T., and Rui, Y. Msr-vtt: A large video description dataset for bridging video and language. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Yoshikawa, Y., Shigeto, Y., and Takeuchi, A. Stair captions: Constructing a large-scale japanese image caption dataset. In *55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2017.

Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. 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.

Yu, Y., Kim, J., and Kim, G. A joint sequence fusion model for video question answering and retrieval. In *European Conference on Computer Vision (ECCV)*, 2018.

Yuan, L., Chen, D., Chen, Y.-L., Codella, N., Dai, X., Gao, J., Hu, H., Huang, X., Li, B., Li, C., et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021.

Zhou, M., Zhou, L., Wang, S., Cheng, Y., Li, L., Yu, Z., and Liu, J. Uc2: Universal cross-lingual cross-modal vision-and-language pre-training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4155–4165, June 2021.

A. Qualitative Analysis

A.1. Case study

In Figure 4, we visualize the top-1 retrieved images for given text queries in 11 languages on XTD dataset (Aggarwal & Kale, 2020). Compared with the multilingual vision-language pre-training model UC^2 (Zhou et al., 2021), MLA can better capture entities, attributes, and actions to retrieve the correct image. Specifically, given simple queries that contain few entities such as Query #1 or Query #2, the images retrieved by MLA show high consistency across languages, since the representations of non-English queries are aligned to English in the NLT stage. For the more complex queries such as Query #3 or Query #4, MLA also shows better fidelity to all entities in most cases.

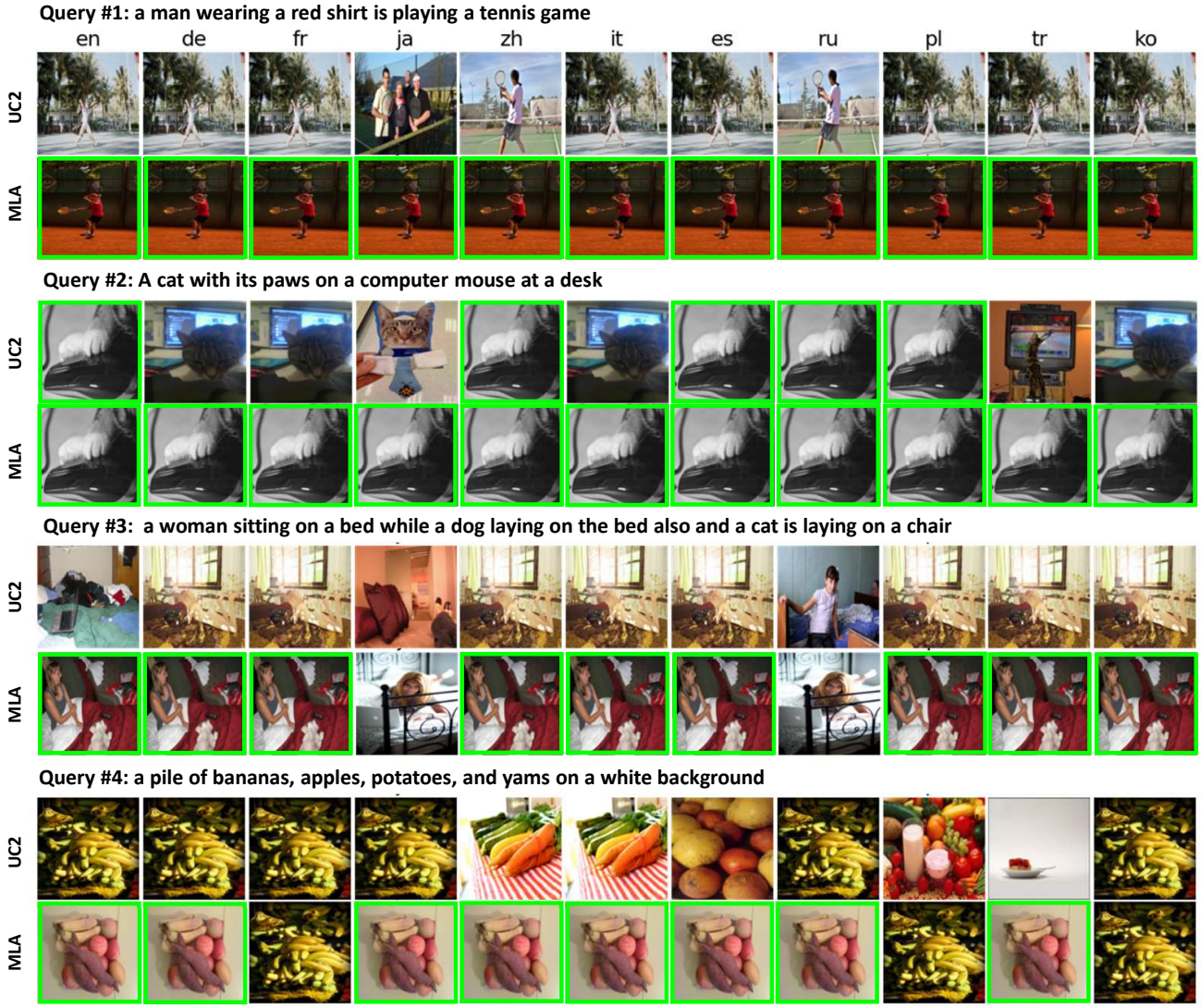

Figure 4: Top-1 retrieved images for given text queries in 11 languages on XTD dataset. Only English queries are shown in this figure. The correct images are bordered green.

A.2. Representation visualization

To visualize the multimodal and multilingual representation space, we translate the English class labels of CIFAR10 (Krizhevsky, 2009) into 5 languages including German (de), French (fr), Czech (cs), Chinese (zh), and Japanese (ja). The images and labels in 6 languages are encoded into representations through MLA_{CLIP}. Figure 5 shows the t-SNE (Van der Maaten & Hinton, 2008) visualization of these representations. We can see that the representations from different languages and modalities are clustered according to the semantics. It suggests that MLA_{CLIP} indeed can project images and multilingual sentences into a shared multimodal and multilingual space.

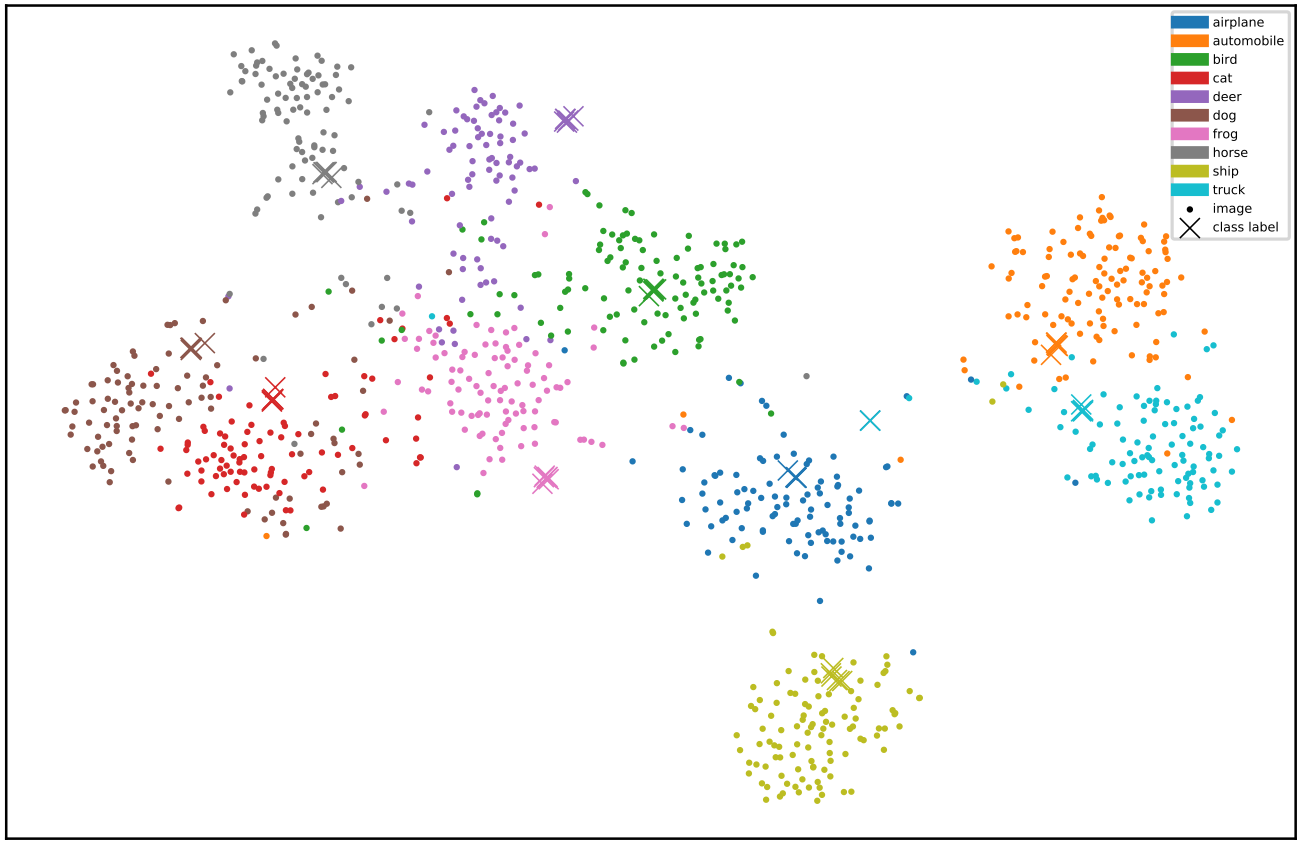

Figure 5: Representation visualization with t-SNE. The categories are color coded. '•' denotes a image representation, and '×' denotes a class label representation in a certain language.

B. Additional Ablation studies

We conduct additional ablation studies to verify the effectiveness of MLA. All experiments in this section are conducted on zero-shot image-text retrieval.

B.1. Structure of language acquirer

In our proposed MLA, we implement the language acquirer as a bottleneck MLP. In Table.8, we compare the different structure of the language acquirer, the bottleneck MLP and a linear projection layer with the same amount of parameters. MLP works slightly better than the linear projection. Thus, we choose MLP to conduct our major experiments.

B.2. Objectives in the two-stage training

In the default setting, we use the MSE objective during the NLT stage and the NCE objective during the LE stage. The MSE objective requires paired representations to be completely consistent, while the NCE objective only requires positive pairs to

Table 8: Ablation study on structure of language acquirer.

Methods	Component	Multi30K			MSCOCO 1K	
		de	fr	cs	ja	zh
MLA_{CLIP}	Linear	78.2	77.6	69.3	74.6	78.0
MLA_{CLIP}	MLP	**78.7**	**77.7**	**70.8**	**74.9**	**78.5**

be closer than negative ones. We conduct experiments to use different objectives in the two stages. As shown in Table 9, we observe that the MSE objective is more suitable for the NLT (row 1 vs. row 2, row 7 vs. row 8) stage, and the NCE objective performs better for the LE stage (row 3 vs. row 4, row 5 vs. row 6). We consider the reason is that in the NLT stage, we leverage translation pairs to build alignment between languages. Since the two sentences of a translation pair are highly semantically related, their representations can be very similar. Thus, optimizing a strong objective like MSE during the NLT stage is feasible. However, during the LE stage, the optimization is conducted with image-text pairs. Although the image and text are semantically related, one sentence can hardly describe all the information in the image. Therefore, a weak objective like NCE is suitable for the LE stage.

Table 9: Ablation study on objectives in the two training stages. mse: MSE objective, nce: NCE objective

Row	Stage one		Stage two		Multi30K			MSCOCO 1K	
	NLT	LE	NLT	LE	de	fr	cs	ja	zh
1	mse				76.3	74.2	67.2	72.1	75.7
2	nce				63.0	58.5	49.6	57.6	64.8
3		mse			47.2	47.0	37.4	46.3	54.9
4		nce			68.2	67.7	58.6	65.9	71.7
5	mse			mse	55.0	51.3	43.8	50.9	57.9
6	mse			nce	**78.7**	**77.7**	**70.8**	**74.9**	**78.5**
7	mse		mse	nce	78.4	77.3	69.9	74.2	78.1
8	mse		nce	nce	78.1	77.2	69.5	73.9	78.2

B.3. Multilingual Acquisition vs. Cross-modal Acquisition

MLA adopts the "Multimodal→Multilingual" strategy that empowers VLP models with multilingual capability. However, there is another option of "Multilingual→Multimodal" that empowers multilingual pre-training models with multimodal capability. To make a comparison between these two strategies, we implement the Cross-Modal Acquisition (CMA) that inserts cross-modal acquirers in each layer of the multilingual pre-training model M-BERT (Devlin et al., 2019). We keep the pre-trained M-BERT fixed and train the cross-modal acquirers with the same two-stage strategy as MLA. From Table 10, we find that CMA performs worse than MLA in all languages. It suggests that generalizing multilingual models to multimodal is harder than generalizing multimodal models to multilingual through lightweight acquirers.

Table 10: Multilingual Acquisition vs. Cross-modal Acquisition

Methods	Multi30K				MSCOCO 1K		
	en	de	fr	cs	en	ja	zh
CMA_{CLIP}	80.2	73.9	72.8	67.0	76.3	69.8	75.1
MLA_{CLIP}	**84.4**	**78.7**	**77.7**	**70.8**	**79.4**	**74.9**	**78.5**

C. Open-domain Image Classification

In order to test the open-domain capability of models, we conduct zero-shot open-domain image classification experiments on CIFAR100 (Krizhevsky, 2009), ImageNet-V2 (Recht et al., 2019), ImageNet-R (Hendrycks et al., 2021a) and ImageNet-A (Hendrycks et al., 2021b) datasets. As shown in Table 11, MKD (Reimers & Gurevych, 2020) performs badly on open-domain image classification. We consider the reason is that MKD abandons the original text encoder which contains

open-domain multimodal knowledge from large-scale pre-training. In contrast, MLA keeps the original text encoder fixed and thus could maintain the open-domain capability of the pre-training model.

Table 11: Top-1 Accuracy of zero-shot open-domain image classification.

Methods	CIFAR100	ImageNet-V2	ImageNet-R	ImageNet-A
MKD_{CLIP}	32.8	54.7	37.7	23.5
MLA_{CLIP}	**64.2**	**63.4**	**69.0**	**31.4**