AltCLIP: Altering the Language Encoder in CLIP for Extended Language Capabilities

Zhongzhi Chen *1,2†, **Guang Liu** *1, **Bo-Wen Zhang** 1
Qinghong Yang 2‡, **Ledell Wu** 1‡
1 Beijing Academy of Artificial Intelligence, 2 Beihang University
{liuguang, bwzhang, wuyu}@baai.ac.cn
{jongjyh, yangqh}@buaa.edu.cn

Abstract

CLIP (Contrastive Language–Image Pretraining) is an English multimodal representation model learned from a massive amount of English text-image pairs and has achieved great success in various downstream tasks, including image classification, text-to-image retrieval, and image generation. When extending CLIP to other languages, the major problem is the lack of good-quality text-image pairs. In this work, we present AltCLIP, a simple and low-resource method to build a strong multilingual multimodal representation model. Instead of training a model from scratch on multilingual text-image pairs, we take the original CLIP model trained on English text-image pairs and alter its text encoder with a pre-trained multilingual text encoder (XLM-R). We then align text and image representations by a two-stage training schema consisting of teacher learning and contrastive learning. Our method utilizes the existence of rich parallel text data and pre-trained multilingual language models. We present extensive experimental evaluations to demonstrate the effectiveness of our proposed method. Our model sets new state-of-the-art zero-shot performances on a wide range of tasks in multilingual multimodal benchmarks, including ImageNet-CN/IT/JA/KO serials, Flicker30k-CN, COCO-CN, Multi30k, and XTD. Further, our model outperforms the original CLIP model on zero-shot cross-modal retrieval, Image Classification in the Wild (ICinW) tasks, and CLIP Benchmark. We open-source our code, pre-trained model weights, and evaluation toolkit of multilingual multimodal tasks, to facilitate research on multilingual multimodal representation learning.

1 Introduction

Learning a good representation in a joint space for vision and language has been a long pursuit in the research of Artificial Intelligence (AI). Recently, the milestone work of CLIP (Radford et al., 2021) from OpenAI demonstrated impressive zero-shot performances across a number of tasks such as image classification on ImageNet (Deng et al., 2009), Image-to-Text and Text-to-Image retrieval on Flicker-30k (Young et al., 2014) and MSCOCO(Lin et al., 2014; Chen et al., 2015). There has been the pursuit of building contrastive language-image models in other languages such as Italian (Bianchi et al., 2021), Korean (Ko and Gu, 2022), Chinese (Changpinyo et al., 2021; Fei et al., 2021; Wang et al., 2022; Gu et al., 2022; Xie et al., 2022) or in a cross-lingual and multilingual setting (Aggarwal and Kale, 2020a).

Training a good language-image representation model often requires a huge amount of text-image pairs and vast computational resources. For instance, CLIP used 400M text-image pairs, and Taiyi (Wang et al., 2022), a recently proposed Chinese model, used 123M text-image pairs. To alleviate these problems, several works manage to take advantage of the existing text-image representation CLIP and extend its language capabilities to other languages (Portaz et al., 2019; Aggarwal and Kale, 2020a; Gu et al., 2022; Zhai et al., 2022).CN-CLIP (Yang et al., 2022) aligns a new Chinese text encoder to the CLIP vision encoder through 200M Chinese text-image pairs. More recently, M-CLIP (Carlsson et al., 2022) proposed to use Teacher Learning (a.k.a. Knowledge Distillation) on the text encoder of the CLIP model to learn a multilingual text-image representation model. This method only uses machine-translated data from English to a target language, without text-image pairs.

However, existing works in the cross-lingual or multilingual setting mainly focus on the model's retrieval performance and ignore their generalization ability. The dataset to evaluate retrieval performance is often small, e.g., $1,000$ images in test

[*] Equal contribution.

[†] Work done during internship with Beijing Academy of Artificial Intelligence.

[‡] Corresponding authors.

Findings of the Association for Computational Linguistics: ACL 2023, pages 8666–8682
July 9-14, 2023 ©2023 Association for Computational Linguistics

sets for Flickr-30k. The retrieval performance fluctuates acutely with the change in training data distribution. Although current methods achieve good performance in retrieval, these methods often do not perform well on the ImageNet classification tasks. The ability to accurately predict images over $1,000$ classes often indicates better generalization ability of the model.

To address the aforementioned problems, we propose a multilingual model named Alter ego CLIP (AltCLIP) which achieved strong performances on ImageNet and multimodal retrieval tasks across languages. Our proposed method AltCLIP learns a multilingual text-image representation under a two-stage framework (see Figure 1 for an overview). In the first stage, we use Teacher Learning on parallel text to distill the knowledge learned from CLIP and align different languages and images. In the second stage, we further improve the alignment of text and image via Contrastive Learning (Hadsell et al., 2006) on a moderate amount of multilingual text-image pairs. We employ this method to train a multilingual Vision-Language model that supports nine languages which we call AltCLIP$_{M9}$.

We present an extensive experimental comparison over a variety of benchmarks and baseline methods, to demonstrate the effectiveness of our method. We show that using recall-based parallel text data in teacher learning can learn well-aligned text-image representation in both English and extended languages, while contrastive learning with text-image pairs effectively aligns the multilingual language model to the CLIP vision encoder. The model trained by this two-step training strategy results in a very strong performance on a broad range of multilingual multimodal benchmarks, including the original English multimodal benchmarks studied in CLIP (Radford et al., 2021). AltCLIP$_{M9}$ sets new state-of-the-art results on multilingual image classification and retrieval tasks. Furthermore, AltCLIP$_{M9}$ achieves superior cross-modal performances in Chinese, Korean, Japanese, and Italian compared to methods trained from scratch with large-scale text-image pairs. Lastly, we apply AltCLIP$_{M9}$ to the task of text-to-image generations (Ramesh et al., 2021; Rombach et al., 2022) to show that it enables high-quality image generation from prompts in different languages.

2 Related Work

CLIP (Radford et al., 2021) provides a strong English Vision-Language representation. To expand the language of the CLIP model, there are prior studies on learning a bilingual text-image representation (Ko and Gu, 2022; Bianchi et al., 2021), and multilingual text-image representation (Aggarwal and Kale, 2020a). In the realm of multi-language models, MURAL(Jain et al., 2021), a dual-tower model, employs contrastive learning between multi-language text and text-image pairs to expand the paradigm of multi-modal learning. It was trained on large-scale private data obtained through web crawling, including more than 6 billion translation pairs and 1.8 billion image-caption pairs. Carlsson et al. (2022) proposed a way to utilize Teacher Learning (a.k.a. Knowledge Distillation) (Hinton et al., 2015) to train a new textual encoder from the original CLIP model with only machine-translated parallel data. Although this method achieves promising cross-lingual retrieval performances with only text data, its zero-shot classification performance in English drops significantly. In the domain of Chinese text-image pretraining models, prior work includes Taiyi (Wang et al., 2022), CN-CLIP (Yang et al., 2022), Wukong (Gu et al., 2022), R2D2 (Xie et al., 2022) and BriVL (Huo et al., 2021; Fei et al., 2021). These methods often need large-scale Chinese text-image pairs and suffer from a significant performance decline in English tasks.

XLM-R (Conneau et al., 2020) is a multilingual language model that achieves strong performances on a wide range of cross-lingual tasks. In our work, we use the XLM-R model as the underlying text encoder and align it with the image encoder trained in CLIP, to achieve competitive performances on cross-lingual and cross-modality tasks.

Knowledge distillation. In knowledge distillation, the teacher-student architecture is a generic carrier to form knowledge transfer. The model capacity gap between a large deep neural network and a small student neural network can degrade knowledge transfer.(Mirzadeh et al., 2020; Gao et al., 2021). To effectively transfer knowledge to student networks, a variety of methods have been proposed for a controlled reduction of the model complexity(Crowley et al., 2018; Liu et al., 2019; Wang et al., 2018). In this work, we use a multilingual model XLM-R as a student model for effectively transferring multilingual knowledge.

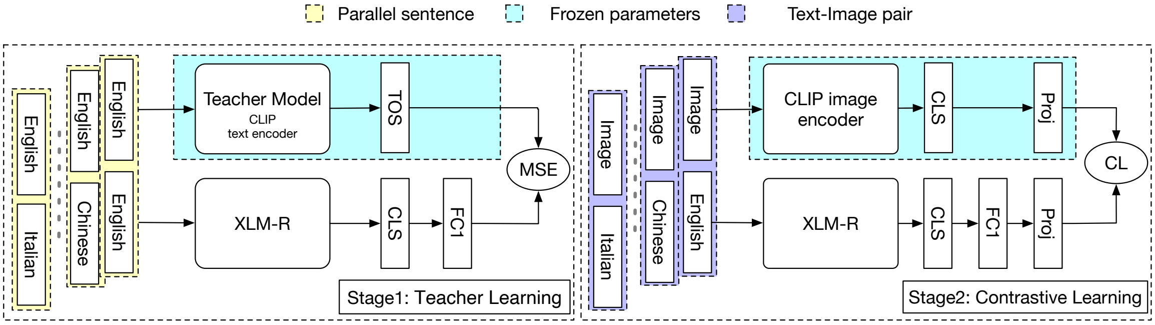


Figure 1: **An illustration of AltCLIP.**In the Teacher Learning stage, the student model (XLM-R) learns a well-aligned multilingual-image representation. The contrastive learning stage further improves alignment using only 7 million text-image pairs per language, making it more resource-efficient than training from scratch.

3 Methodology

We propose a two-stage method to learn a multilingual multimodal representation model. In the first stage, we follow the work of Carlsson et al. (2022) to use Teacher Learning to learn a multilingual text encoder from the CLIP text encoder. In this step, no image is needed in training and only language parallel data is used. In the second stage, we use text-image pairs to further fine-tune the model from contrastive learning. Our overall training procedure is summarized in Figure 1.

3.1 Teacher Learning Stage

In this stage, we perform Teacher Learning (Hinton et al., 2015) on text encoders. We use the text encoder from CLIP (Radford et al., 2021) as the teacher text encoder, and the XLM-R (Conneau et al., 2020) model pretrained on multilingual data as the student text encoder. A fully-connected layer is added to transform the output of the XLM-R model into the same output dimension as the teacher encoder. We use parallel text data between English and other languages * to distill the knowledge of text-image alignment.

Given parallel text input $(sent_1, sent_2)$, the teacher text encoder generates the learning target from input $sent_1$, which is the embedding of the [TOS] token, denoted by x^t_{tos}. The student text encoder generates embedding x^s_{cls} from input $sent_2$. We minimize Mean Squared Error (MSE) between x^t_{tos} and x^s_{cls}. After such training, the student text encoder can keep most of its multilingual capability and obtain text-image alignment capability in both languages. Note that the teacher encoder is only used at training time. At inference time, only the student encoder is used as the text encoder.

To show that our method is extensible in including more languages, we build a multilingual version (AltCLIP$_{M9}$) and a bilingual version (AltCLIP$_{M2}$). AltCLIP$_{M9}$ supports nine different languages: English(EN), Chinese(CN), Spanish(ES), French(FR), Russian(RU), Arabic(AR), Japanese(JA), Korean(KO), and Italian(IT). For the bilingual version (AltCLIP$_{M2}$), we align Chinese with English, with the same concept and architecture as in the multilingual version.

3.2 Contrastive Learning Stage

This stage of training aims to further improve text-image alignment by contrastive learning on multilingual text-image pairs. As illustrated in Figure 1, here we use the image encoder from CLIP which is based on Vision Transformer (ViT) (Dosovitskiy et al., 2020) as our image encoder, and use the student text encoder learned from the Teacher Learning Stage as our text encoder.

We use Contrastive Loss (Hadsell et al., 2006) between the output projection of the image encoder and text encoder, as done similarly in previous work (Radford et al., 2021). We follow LiT (Zhai et al., 2022) to freeze the image encoder at training time and only update the parameters in the text encoder. We observe that this stage of training further improves the model's performance on various evaluation benchmarks, as presented in Section 5.

4 Model Training

4.1 Training Datasets

In this section, we describe the training datasets used in our two-stage training schema.

Teacher Learning Stage We only use the parallel corpus to align the original CLIP text

[*] We also include English-English text pairs as parallel text

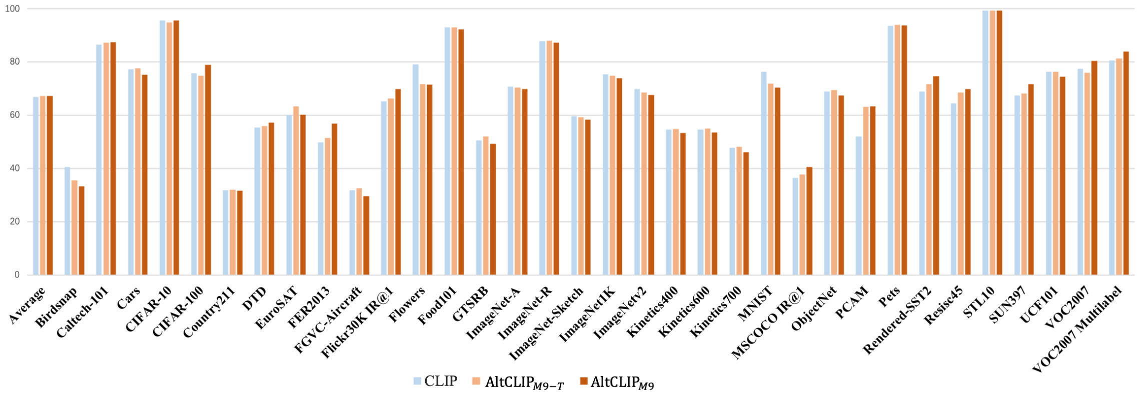


Figure 2: **Experimental results on CLIP Benchmark.** AltCLIP$_{M9-T}$ denotes our model after Teacher Learning Stage while AltCLIP$_{M9}$ denotes our model after Contrastive Learning Stage. All image encoders are CLIP$_{ViT-L14}$.

encoder and XLM-R text encoder. The parallel corpus consists of a recall-based corpus and a machine-translated corpus translated by MBART(Tang et al., 2020). We use the same amount of data for each language, which contains 5M recall-based parallel data collected from OPUS(Tiedemann, 2012)[†], 10M machine-translated data from LAION(Schuhmann et al., 2021)[‡] and 3M machine-translated data from Conceptual Captions (CC3M)(Sharma et al., 2018). We use TSL2019(5M)(Xu, 2019) as parallel data for the training of AltCLIP$_{M2}$.

Contrastive Learning Stage We use unfiltered text-image pair data in this stage. For AltCLIP$_{M9}$, we randomly selected 7 million text-image pairs for each language from the LAION2B-Multi(Schuhmann et al., 2022). For AltCLIP$_{M2}$, we only employed half a million text-image pairs for each language in training.

4.2 Implementation details

We initialize our text encoder from XLM-R$_{Large}$ and use the text encoder from CLIP$_{ViT-L14}$ as the teacher text encoder. We use the image encoder from CLIP$_{ViT-L14}$ as our image encoder. In the Teacher Learning stage, we trained for 27 hours using 11×8 NVIDIA A100-SXM4-40GB GPUs. In the Contrastive Learning stage, we continued training for an additional 12 hours using 8 NVIDIA A100-SXM4-40GB GPUs. Detailed training settings can be found in Appendix A.3.

5 Experiments

We present experimental results in this section. In Section 5.1, we introduce the datasets and metrics used. We comprehensively validate our model through multilingual multimodal benchmarks in Section 5.2. In Section 5.3, we conduct an ablation study on the effects of various design choices in Teacher Learning and Contrastive Learning. Finally, in Section 5.4, we apply AltCLIP to text-image generation, and show that our model is capable to align text in different languages.

5.1 Evaluation Datasets and Metrics

In this section, we describe the datasets and metrics used. We use ImageNet (Deng et al., 2009) and its four out-of-distribution test variants, i.e. ImageNet Sketch (Wang et al., 2019), ImageNet-A (Hendrycks et al., 2021b), ImageNet-R (Hendrycks et al., 2021a), ImageNetV2 (Recht et al., 2019), to evaluate zero-shot image classification performances in English(Radford et al., 2021), Chinese, Japanese, Italain and Korean[§]. We adapt templates of manual prompts from CLIP for English and the corresponding machine translation templates for Chinese and Korean. For Japanese and Italian, the templates are collected from the same sources with the translated class names.

For cross-modal retrieval, we evaluate AltCLIP$_{M9}$ on the XTD (Aggarwal and Kale, 2020b) dataset and Multi30k (Elliott et al., 2016).

[†] https://opus.nlpl.eu

[‡] Randomly sampled from LAION and translated into multiple languages by machine translation.

[§] The corresponding translations of class names are respectively collected from - Chinese:https://github.com/ningbonb/imagenet_classes_chinese, Japanese: https://github.com/rinnakk/japanese-clip, Italian: Italian CLIP(Bianchi et al., 2021), Korean: machine translation

Lan.	Method	Txt-Img Data	IN-Adv.	IN-Ren.	IN-Ske.	IN-1K	IN-V2	avg.
English	M-CLIP	-	59.1	81.6	44.2	52.3	47.4	56.9
	OpenCLIP	50×	53.9	**87.5**	**63.3**	**75.3**	**67.7**	69.5
	AltCLIP$_{M9}$	1×	**69.8**	87.2	58.4	74.0	67.6	**71.4**$_{(+1.9)}$
Chinese	M-CLIP	-	50.9	68.4	36.2	43.0	39.6	47.6
	CN-CLIP	25×	43.3	78.1	47.3	53.3	48.1	54.0
	AltCLIP$_{M9}$	1×	**61.2**	**82.4**	**48.4**	**59.6**	**54.0**	**61.1**$_{(+7.1)}$
Japanese	M-CLIP	-	21.8	44.5	24.6	26.9	24.2	28.4
	JA-CLIP†	NA	21.2	50.9	25.1	50.7	43.5	38.3
	AltCLIP$_{M9}$	1×	**52.7**	**75.6**	**46.7**	**55.0**	**50.3**	**56.1**$_{(+17.8)}$
Italian	M-CLIP	-	51.8	72.9	38.3	43.0	38.9	49.0
	IT-CLIP†	0.7×	10.5	27.2	16.5	21.9	19.4	19.1
	AltCLIP$_{M9}$	1×	**56.7**	**78.2**	**45.9**	**55.3**	**50.4**	**57.3**$_{(+8.3)}$
Korean	M-CLIP	-	20.9	39.3	22.1	25.2	22.8	26.0
	KELIP†	100×	19.4	53.1	26.6	33.7	30.3	32.6
	AltCLIP$_{M9}$	1×	**51.1**	**72.9**	**44.8**	**55.2**	**50.5**	**54.9**$_{(+22.5)}$

Table 1: **Results on multilingual Image Classification benchmarks.** We compare AltCLIP$_{M9}$ with the M-CLIP and a model trained from scratch in five languages. For a fair comparison, the ones with ViT-L are chosen as default except for the ones with the mark †. The metric reported is zero-shot classification accuracy. We also build datasets and evaluate our model in the rest four languages with machine translation, details are in Appendix A.1

XTD is built from selecting 1K images from COCO (Lin et al., 2014), and translating the corresponding English Captions into 11 languages. [¶]. The Multi30k dataset is a collection of multilingual image captions that provides translations of captions in English, German, French, and Czech for 29,000 images. We select Flickr30k (Young et al., 2014), COCO, as well as their corresponding Chinese datasets, Flickr30k$_{\mathrm{CN}}$ (Lan et al., 2017), COCO$_{\mathrm{CN}}$ [‖] (Li et al., 2019), to evaluate zero-shot image-to-text retrieval and text-to-image retrieval performances on Chinese.

We further evaluated our model on a wide range of English tasks to compare its performance with the original CLIP model. We used datasets introduced in CLIP and the Open CLIP benchmark[**] and "Image Classification in the Wild (ICinW)" dataset from the ELEVATER benchmark (Li et al., 2022), including Birdsnap (Berg et al., 2014), Caltech-101 (Fei-Fei et al., 2006), Stanford Cars (Krause et al., 2013), CIFAR-10 (Krizhevsky et al., 2009), CIFAR-100 (Krizhevsky et al., 2009), Country211 (Radford et al., 2021), DTD (Cimpoi et al., 2014), EuroSAT (Helber et al., 2019), Facial Emotion Recognition 2013 (Goodfellow et al., 2013), FGVC Aircraft (Blaschko et al., 2012), Oxford Flowers 102 (Nilsback and Zisserman, 2008), Food-101 (Bossard et al., 2014), GTSRB (Stallkamp et al., 2011), Kinetics400 (Kay et al., 2017), Kinetics600 (Carreira et al., 2018), MNIST (Cireşan et al., 2011), PatchCamelyon (Veeling et al., 2018), ObjectNet (Barbu et al., 2019), Oxford-IIIT Pets (Parkhi et al., 2012), Rendered SST2 (Radford et al., 2021), RESISC45 (Cheng et al., 2017), STL-10 (Coates et al., 2011), SUN397 (Xiao et al., 2010), UCF101 (Soomro et al., 2012), Pascal VOC 2007 Classification (Everingham, 2007), Pascal VOC 2007 Multilabel Classification (Everingham, 2007), KITTI-Distance (Fritsch et al., 2013) and hateful-memes (Kiela et al., 2020).

The evaluation metrics for image classification benchmarks are accuracy (default), mean per class (the average of recall obtained on each category, for imbalanced datasets, such as FGVC Aircraft, Oxford-IIIT Pets, Caltech-101, Oxford Flowers 102), 11-point mAP (mean average of 11-pt interpolated precision for each class, for VOC 2007), and mean(top1, top5) (the mean of acc@1 and acc@5, for Kinetics400 and Kinetics600). For cross-modal retrieval benchmarks, we use Recall@K where $K \in \{1, 5, 10\}$, and Mean Recall (the average of Recall@K) for both image-to-text retrieval and text-to-image retrieval tasks, which are the same as the setups in CLIP (Radford et al., 2021).

5.2 Zero-shot performance

Image Classification We first present evaluation results of zero-shot image classification on the ImageNet dataset and its four out-of-distribution

[¶] English(EN), German(DE), French(FR), Chinese(CN), Japanese(JA), Italian(IT), Spanish(ES), Russian(RU), Polish(PL), Turkish(TR), Korean(KO)

[‖] There are two versions: texts in the 1k version(COCO$_{CNa}$) are manually written captions while in the 5k version (COCO$_{CNb}$) are manually translated captions

[**] `https://github.com/LAION-AI/CLIP_benchmark`

Model	XTD								Multi30K			
	En	Es	Fr	Zh	It	Ko	Ru	Jp	En	Fr	De	Cs
Base Model												
CLIP$_{ViT-B32}$	90.3	-	-	-	-	-	-	-	-	-	-	-
mUSE PATR	83.6	75.6	76.9	76.1	73.4	64.3	73.6	69.4	-	-	-	-
mUSE m3	85.3	78.9	78.9	76.7	73.6	67.8	76.1	70.7	-	-	-	-
UC^2	65.2	56.5	59.7	60.1	57.7	50.2	50.9	50.5	66.6	60.4	62.5	55.1
MLA$_{ViT-B16}$	76.0	62.8	72.9	73.8	64.7	57.3	58.1	67.2	86.4	80.9	80.8	72.9
ALIGN$_{BASE}$	-	88.8	-	86.5	87.9	76.6	82.3	-	84.3	78.3	78.9	71.1
MURAL$_{BASE}$	-	89.6	-	88.3	88.4	82.4	83.6	-	82.4	75.0	76.2	64.6
M-CLIP$^{\ddagger}_{ViT-B32}$	91.8	89.1	89.4	89.3	89.8	82.1	86.1	81.0	80.4	71.1	71.4	67.7
Large Model												
CLIP$_{ViT-L14}$	91.8	-	-	-	-	-	-	-	87.7	-	-	-
M-CLIP$^{\ddagger}_{ViT-L14}$	92.4	91	90	89.7	91.1	85.2	85.8	81.9	87.8	82.5	**83.1**	**81.3**
MURAL$_{LARGE}$	-	92.9	-	89.7	91.8	88.1	87.2	-	89.2	83.1	83.5	77.0
AltCLIP$_{M9}$	**93.3**	**92.2**	**91.1**	**92.2**	**91.9**	**91.5**	**89.2**	**89.1**	**89.9**	**85.2**	65.5†	36.6†

Table 2: **Results on the multilingual cross-modal retrieval dataset.** Recall@10 is reported for Text-to-Image on XTD and average recall for Text-to-Image and Image-to-Text on Multi30K. † denotes the unseen language in training AltCLIP$_{M9}$, such as German and Czech. ‡ denotes the reproduced results. Numbers denotes the good results of MURAL$_{LARGE}$ comes from the large-scale private data: 6 billion translation pairs (up to 100 million per language) in 109 languages and 1.8 billion image-caption pairs.

	Average	Caltech-101	Cars	CIFAR-10	CIFAR-100	Country211	DTD	EuroSAT	FER2013	FGVC-Aircraft	Flowers	Food101	GTSRB	hateful-memes	KITTI-Distance	MNIST	PCAM	Pets	Rendered-SST2	RESISC45	VOC2007
M-CLIP	53.5	81.4	53.5	93.8	72.6	22.5	41.2	**62.0**	47.7	7.3	26.3	68.8	42.5	53.0	**28.7**	60.1	51.3	49.9	65.6	62.0	79.7
CLIP	64.9	86.6	**77.3**	95.6	75.8	**31.9**	55.4	60.0	49.9	**31.9**	**79.1**	**93.1**	**50.6**	56.0	21.8	**76.4**	52.0	93.6	69.0	64.5	77.4
AltCLIP-M9	**66.1**	**87.5**	75.2	**95.7**	**79.0**	31.7	**57.3**	60.3	**56.8**	29.6	71.5	92.3	49.2	**57.2**	25.5	70.5	**63.3**	**93.8**	**74.7**	**69.8**	**80.5**

Table 3: The results on Image Classification in the Wild (ICinW)

variants. For baselines, we compare our model with OpenCLIP (Radford et al., 2021), CN-CLIP (Yang et al., 2022), KELIP (Ko and Gu, 2022), IT-CLIP (Bianchi et al., 2021), JA-CLIP (, 2022) and multilingual CLIP (M-CLIP) (Carlsson et al., 2022). As illustrated in Table 1, AltCLIP$_{M9}$ outperforms OpenCLIP in English and sets new state-of-the-art results on ImageNet, ImageNet-A, ImageNet-R, and ImageNet V2 in Chinese, Japanese, Korean and Italian. These results demonstrate the effectiveness of our method in expanding the language ability of CLIP. Compared to Chinese/Korean baseline models where hundreds of millions of text-image pairs are used in pretraining, we only use 18M parallel text data and 7M text-image pairs (per language) in training.

Multilingual Cross-modal Retrieval We compare our model with CLIP, M-CLIP (Carlsson et al., 2022), mUSE (Yang et al., 2020), UC^2 (Zhou et al., 2021), MLA (Zhang et al., 2022), ALIGN (Jia et al., 2021) and MURAL (Jain et al., 2021). The results of the comparison on Multi30k(Elliott et al., 2016) and XTD (Aggarwal and Kale, 2020b) are shown in Table 2, where AltCLIP$_{M9}$ achieves state-of-the-art results in 7 languages and outperforms the original CLIP model in English. This superior performance of our model is likely due to the use of higher-quality parallel corpora during the Teacher Learning stage, which effectively eliminates potential bias from machine translation. Additionally, we utilize contrastive learning to further align the text and image representation, which is crucial for downstream tasks. We will discuss this in more detail in Section 5. We also provide additional cases in Appenix A.4.

Dataset	Method	Text-to-Image Retrival			Image-to-Text Retrival			MR
		R@1	R@5	R@10	R@1	R@5	R@10	
Flickr30k	CLIP	65.0	87.1	92.2	85.1	97.3	99.2	87.6
	Taiyi	25.3	48.2	59.2	39.3	68.1	79.6	53.3
	CN-CLIP	49.5	76.9	83.8	66.5	91.2	96.0	77.3
	AltCLIP-M2	**72.5**	**91.6**	**95.4**	86.0	**98.0**	99.1	**90.4**
	AltCLIP-M9	69.8	90.8	94.2	**86.6**	97.8	**99.2**	89.7
COCO	CLIP	36.5	61.1	71.1	56.4	79.5	86.5	65.2
	Taiyi	11.7	27.8	37.4	19.8	42.1	54.3	32.2
	CN-CLIP	26.1	50.0	61.3	40.9	65.8	76.3	53.4
	AltCLIP-M2	**42.9**	**68.0**	**77.4**	58.6	80.6	87.8	**69.2**
	AltCLIP-M9	40.5	65.2	74.9	**58.7**	**81.2**	**88.3**	68.2
Flickr30k$_{CN}$	CLIP	0	2.4	4.0	2.3	8.1	12.6	5.0
	Taiyi	53.7	79.8	86.6	63.8	90.5	95.9	78.4
	Wukong†	51.7	78.9	86.3	76.1	94.8	97.5	80.9
	R2D2†	60.9	86.8	92.7	77.6	96.7	98.9	85.6
	CN-CLIP	68	89.7	94.4	80.2	96.6	98.2	87.9
	AltCLIP-M2	**69.8**	**89.9**	**94.7**	84.8	97.4	98.8	**89.2**
	AltCLIP-M9	68.6	89.4	94.5	**85.8**	**98.2**	**99.0**	**89.2**
COCO$_{CNa}$	CLIP	0.6	4.1	7.1	1.8	6.7	11.9	5.4
	Taiyi	52.0	80.2	89.6	46.6	76.3	88.6	72.2
	Wukong†	55.2	81.0	90.6	53.4	80.2	90.1	75.1
	R2D2†	63.3	**89.3**	**95.7**	56.4	85.0	93.1	80.5
	CN-CLIP	63.7	88.7	94.4	61.0	84.7	93.6	81.0
	AltCLIP-M2	**63.9**	87.2	93.9	62.8	88.8	95.5	**82.0**
	AltCLIP-M9	60.6	86.3	93.4	**66.2**	**88.9**	**96.2**	81.9
COCO$_{CNb}$	CLIP	0.8	3.9	5.8	3.5	8.9	14.4	6.2
	Taiyi	46.1	74.9	85.1	58.1	83.9	91.7	73.3
	CN-CLIP	58.6	85.3	92.7	72.1	90.9	94.7	82.4
	AltCLIP-M2	**61.3**	**86.0**	**93.2**	**77.8**	**94.4**	97.5	**85.0**
	AltCLIP-M9	58.9	84.5	92.5	77.7	94.3	**97.7**	84.3

Table 4: **Experimental results on English and Chinese retrieval tasks.** All image encoders used in these models are ViT-L for a fair comparison.† represents we report original results from papers.

Full CLIP benchmark We present the evaluation results for a range of tasks in English in Figure 2. We compare the effectiveness of multilingual AltCLIP$_{M9}$ and AltCLIP$_{M9-T}$ with the original CLIP. AltCLIP$_{M9}$ outperforms CLIP, indicating that our method effectively fuses the abilities of CLIP and XLMR. We observed that at the Teacher Learning stage, the model already learns a good representation of text-image representation, as it achieves better average results than the original CLIP model on a range of zero-shot benchmarks. The Contrastive Learning stage further improves the model's performance, particularly on retrieval tasks such as Flickr30k.

Task-level transferability We evaluated the transferability of AltCLIP for zero-shot image classification on the "Image Classification in the Wild (ICinW)" dataset from the ELEVATER benchmark (Li et al., 2022). ICinW is a publicly available benchmark to evaluate the large-scale task-level transferability of Vison Language models. ICinW consists of a series of image classification datasets such as KITTI-Distance (Fritsch et al., 2013) and hateful-memes (Kiela et al., 2020). As shown in Table 3, AltCLIP$_{M9}$ achieved an average score of 66.1, outperforming the original CLIP and achieving a 23.6% improvement compared to M-CLIP, demonstrating the effectiveness of our training strategy.

Method	Multilingual			English			
	INs	IRs	TRs	INs	IRs	TRs	CinW
MT	47.8	54.2	63.5	71.5	57.6	73.2	**66.1**
RB	50.2	51.8	60.8	67.2	55.1	71.2	61.5
MT+RB	56.2	56.2	65.6	**72.2**	57.7	73.1	65.8
MT+RB+CL	**58.4**	**60.6**	**68.7**	71.4	**60.8**	**74.8**	**66.1**

Table 5: **Ablation Experiments.** For a fair comparison, all models were trained for 10 epochs and evaluated using the average results over nine-language ImageNet series tasks (INs), image-retrieval tasks (IRs), and text-retrieval tasks (TRs) on XTD and Multi30K for eight languages (excluding Arabic).

Comparison with models trained from scratch. We compare our model with the ones trained with hundreds of millions of text-image pairs: CLIP in English and R2D2 (Xie et al., 2022), Wukong

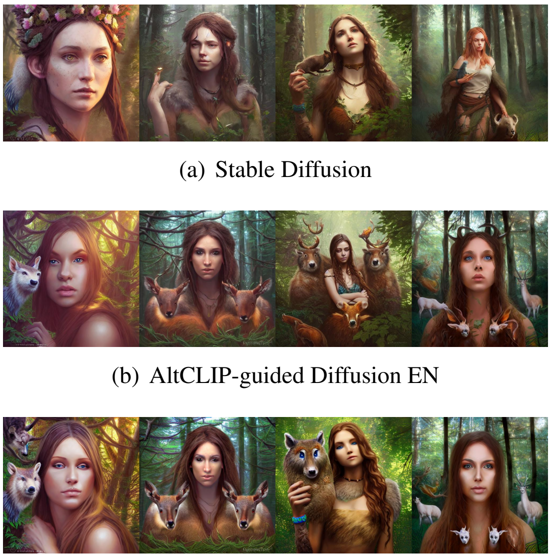
(Gu et al., 2022), Taiyi (Wang et al., 2022) and CN-CLIP (Yang et al., 2022) in Chinese. The results are shown in Table 4. AltCLIP$_{M9}$ outperforms all baseline models including models trained with large-scale text-image pairs on most datasets and tasks. We notice that AltCLIP$_{M2}$ outperforms CLIP on both text-to-image and image-to-text retrieval. This could be due to the following reasons: 1). We used a small subset (less than 1M) of LAION 5B at the Contrastive Learning stage, which is in a different distribution of the pretraining data used in CLIP; 2). Our language encoder initialized from XLM-R provides better language understanding ability. We elaborate on the detailed results of Bilingual settings in Appendix A.2.

5.3 Ablation study

We evaluate the effectiveness of our AltCLIP$_{M9}$ by analyzing its major components in this section. We use CL to denote the Contrastive Learning stage, and MT and RB to denote the Machine-Translated and Recall-Based parallel data used in the Teacher Learning stage. We evaluate the variations of our models in English-only and in multilingual settings. We use the average score on ImageNet series (INs), Image Retrieval tasks (IRs), and Text Retrieval tasks (TRs) as evaluation metrics. Results in Table 5 show that excluding machine-translated data has a significant impact on performance, except for the multilingual ImageNet series tasks. Combining machine-translated and recall-based parallel data leads to a significant improvement in most tasks, indicating that the quality and diversity in training data are both important. Additionally, the Contrastive Learning stage significantly improves the model's performances on multilingual tasks, achieving 58.4 on multilingual INs, a 3.9% improvement.

5.4 Examples of text-to-image generation

In this section, we apply our model to the task of text-to-image generation to enable multilingual image generation, and to show the effect of language alignment in our model. We use the text encoder of AltCLIP$_{M9}$ to fine-tune a Stable Diffusion model (Rombach et al., 2022). We use stable-diffusion v1-4[††] as initialization and AltCLIP$_{M9}$ as the language encoder, and we freeze all parameters in the diffusion model except for the key and value projection layers of the cross-attention block during fine-tuning. The dataset used for fine-tuning is the same one used for the Contrastive Learning stage as described in Section 4.1. As demonstrated in Fig. 3, our model generates high-quality images comparable to those generated by Stable Diffusion. This is likely due to the reason that AltCLIP$_{M9}$ achieves competitive performance in English with CLIP, where the latter is used in the original Stable Diffusion model. Additionally, we observe that our model generates similar images for translated English and Chinese prompts, demonstrating the effect of language alignment. More examples with images generated from different languages can be found in Appendix A.5.



(a) Stable Diffusion

(b) AltCLIP-guided Diffusion EN

(c) AltCLIP-guided Diffusion CN

Figure 3: Examples of text-to-image generation. Text prompt: *"a pretty female druid surrounded by forest animals, digital painting, photorealistic, in the style of greg rutkowski, highly detailed, realistic."*, *"一个由森林动物环绕的漂亮的女德鲁伊,数字绘画,摄影现实,格雷格·鲁特科夫斯基风格,高度详细,现实"*

6 Conclusion

In this work, we propose an effective two-stage training method for learning multilingual multimodal representation models, through teacher learning and contrastive learning. The effectiveness is demonstrated through extensive experiments on a wide range of tasks in multilingual multimodal benchmarks. AltCLIP$_{M9}$ outperforms the original CLIP model on many tasks in English and sets new state-of-the-art zero-shot results on multiple image classification tasks in Chinese/Korean/Italian/Japanese and multilingual re-

[††] https://huggingface.co/CompVis/stable-diffusion-v-1-4-original

trieval tasks. Meanwhile, our method is highly data-efficient, which consumes only around 1% text-image pairs compared to the hundreds of millions of text-image pairs used by prior work on vision-language pretraining models.

7 Limitations

It's worth noting that this study has certain limitations. One of the limitations is the limited scope of the training data employed. The AltCLIP model is trained on open-source parallel corpora and publicly available unfiltered text-image pairs. A more careful study of the training data, i.e. filtering text-image pairs by relevance and text/image quality may help to further improve the overall performance of the model. Another limitation is the challenge of evaluating the model in a multilingual setting. Despite our best efforts to include as many benchmarks as possible and to translate from English datasets, the evaluation of the model's performance in other languages is not as comprehensive as it is in English. For example, there may be fewer tasks available such as OCR or action recognition in videos in other languages. In addition, the use of machine translation may introduce biases that could affect performance. Future research should focus on creating a more robust and scientifically rigorous multilingual evaluation framework.

8 Ethics Statement

The AltCLIP approach presents an innovative way of building robust multilingual multimodal representation models while minimizing the need for energy-intensive GPU training, promoting a more sustainable approach. Additionally, it allows for greater accessibility as it does not require extensive computational resources to implement. Furthermore, our model was trained using open-sourced data and our model is open-sourced to promote transparency and reproducibility. However, we have not carefully investigated the training data we used, such as LAION (Schuhmann et al., 2022). The data may contain unsafe or biased text and/or images. It is important to note that models pretrained on it have the potential to reproduce sensitive training data. It is crucial to use this method responsibly and ethically to ensure it contributes to safe applications.

References

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

Pranav Aggarwal and Ajinkya Kale. 2020b. Towards zero-shot cross-lingual image retrieval.

Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. 2019. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. *Advances in neural information processing systems*, 32.

Thomas Berg, Jiongxin Liu, Seung Woo Lee, Michelle L Alexander, David W Jacobs, and Peter N Belhumeur. 2014. Birdsnap: Large-scale fine-grained visual categorization of birds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2011–2018.

Federico Bianchi, Giuseppe Attanasio, Raphael Pisoni, Silvia Terragni, Gabriele Sarti, and Sri Lakshmi. 2021. Contrastive language-image pretraining for the italian language. *arXiv preprint arXiv:2108.08688*.

Matthew Blaschko, Ross B Girshick, Juho Kannala, Iasonas Kokkinos, Siddarth Mahendran, Subhransu Maji, Sammy Mohammed, Esa Rahtu, Naomi Saphra, Karen Simonyan, et al. 2012. Towards a detailed understanding of objects and scenes in natural images.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. 2014. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*.

Fredrik Carlsson, Philipp Eisen, Faton Rekathati, and Magnus Sahlgren. 2022. Cross-lingual and multilingual clip. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6848–6854.

Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. 2018. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*.

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 the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Gong Cheng, Junwei Han, and Xiaoqiang Lu. 2017. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. 2014. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613.

Dan C Cireşan, Ueli Meier, Jonathan Masci, Luca M Gambardella, and Jürgen Schmidhuber. 2011. High-performance neural networks for visual object classification. *arXiv preprint arXiv:1102.0183*.

Adam Coates, Andrew Ng, and Honglak Lee. 2011. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223. JMLR Workshop and Conference Proceedings.

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 *ACL*.

Elliot J Crowley, Gavin Gray, and Amos J Storkey. 2018. Moonshine: Distilling with cheap convolutions. *Advances in Neural Information Processing Systems*, 31.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee.

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

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30k: Multilingual english-german image descriptions. *arXiv preprint arXiv:1605.00459*.

Mark Everingham. 2007. The pascal visual object classes challenge,(voc2007) results. *http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2007/index.html*.

Nanyi Fei, Zhiwu Lu, Yizhao Gao, Guoxing Yang, Yuqi Huo, Jingyuan Wen, Haoyu Lu, Ruihua Song, Xin Gao, Tao Xiang, et al. 2021. Wenlan 2.0: Make ai imagine via a multimodal foundation model. *arXiv preprint arXiv:2110.14378*.

Li Fei-Fei, Robert Fergus, and Pietro Perona. 2006. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 28(4):594–611.

Jannik Fritsch, Tobias Kuehnl, and Andreas Geiger. 2013. A new performance measure and evaluation benchmark for road detection algorithms. In *16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013)*, pages 1693–1700. IEEE.

Mengya Gao, Yujun Wang, and Liang Wan. 2021. Residual error based knowledge distillation. *Neurocomputing*, 433:154–161.

Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. 2013. Challenges in representation learning: A report on three machine learning contests. In *International conference on neural information processing*, pages 117–124. Springer.

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

Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. 2019. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. 2021a. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. 2021b. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *stat*, 1050:9.

Yuqi Huo, Manli Zhang, Guangzhen Liu, Haoyu Lu, Yizhao Gao, Guoxing Yang, Jingyuan Wen, Heng Zhang, Baogui Xu, Weihao Zheng, et al. 2021.

Wenlan: Bridging vision and language by large-scale multi-modal pre-training. *arXiv preprint arXiv:2103.06561*.

Aashi Jain, Mandy Guo, Krishna Srinivasan, Ting Chen, Sneha Kudugunta, Chao Jia, Yinfei Yang, and Jason Baldridge. 2021. Mural: multimodal, multitask retrieval across languages. *arXiv preprint arXiv:2109.05125*.

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 *International Conference on Machine Learning*, pages 4904–4916. PMLR.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in Neural Information Processing Systems*, 33:2611–2624.

Byungsoo Ko and Geonmo Gu. 2022. Large-scale bilingual language-image contrastive learning. *arXiv preprint arXiv:2203.14463*.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 2013. 3d object representations for fine-grained categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, Sydney, Australia.

Alex Krizhevsky, Geoffrey Hinton, et al. 2009. Learning multiple layers of features from tiny images.

Weiyu Lan, Xirong Li, and Jianfeng Dong. 2017. Fluency-guided cross-lingual image captioning. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1549–1557.

Chunyuan Li, Haotian Liu, Liunian Harold Li, Pengchuan Zhang, Jyoti Aneja, Jianwei Yang, Ping Jin, Yong Jae Lee, Houdong Hu, Zicheng Liu, et al. 2022. Elevater: A benchmark and toolkit for evaluating language-augmented visual models. *arXiv preprint arXiv:2204.08790*.

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 *European conference on computer vision*, pages 740–755. Springer.

Iou-Jen Liu, Jian Peng, and Alexander G Schwing. 2019. Knowledge flow: Improve upon your teachers. *arXiv preprint arXiv:1904.05878*.

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. 2020. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 5191–5198.

Maria-Elena Nilsback and Andrew Zisserman. 2008. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729. IEEE.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. 2012. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE.

Maxime Portaz, Hicham Randrianarivo, Adrien Nivaggioli, Estelle Maudet, Christophe Servan, and Sylvain Peyronnet. 2019. Image search using multilingual texts: a cross-modal learning approach between image and text. *arXiv preprint arXiv:1903.11299*.

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

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2019. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pages 5389–5400. PMLR.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*.

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

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. 2011. The german traffic sign recognition benchmark: a multi-class classification competition. In *The 2011 international joint conference on neural networks*, pages 1453–1460. IEEE.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218.

Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. 2018. Rotation equivariant cnns for digital pathology. In *International Conference on Medical image computing and computer-assisted intervention*, pages 210–218. Springer.

Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. 2019. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32.

Hui Wang, Hanbin Zhao, Xi Li, and Xu Tan. 2018. Progressive blockwise knowledge distillation for neural network acceleration. In *IJCAI*, pages 2769–2775.

Junjie Wang, Yuxiang Zhang, Lin Zhang, Ping Yang, Xinyu Gao, Ziwei Wu, Xiaoqun Dong, Junqing He, Jianheng Zhuo, Qi Yang, Yongfeng Huang, Xiayu Li, Yanghan Wu, Junyu Lu, Xinyu Zhu, Weifeng Chen, Ting Han, Kunhao Pan, Rui Wang, Hao Wang, Xiaojun Wu, Zhongshen Zeng, Chongpei Chen, Ruyi Gan, and Jiaxing Zhang. 2022. Fengshenbang 1.0: Being the foundation of chinese cognitive intelligence. *CoRR*, abs/2209.02970.

J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3485–3492.

Chunyu Xie, Heng Cai, Jianfei Song, Jincheng Li, Fanjing Kong, Xiaoyu Wu, Henrique Morimitsu, Lin Yao, Dexin Wang, Dawei Leng, et al. 2022. Zero and r2d2: A large-scale chinese cross-modal benchmark and a vision-language framework. *arXiv preprint arXiv:2205.03860*.

Bright Xu. 2019. Nlp chinese corpus: Large scale chinese corpus for nlp.

An Yang, Junshu Pan, Junyang Lin, Rui Men, Yichang Zhang, Jingren Zhou, and Chang Zhou. 2022. Chinese clip: Contrastive vision-language pretraining in chinese. *arXiv preprint arXiv:2211.01335*.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 2020. Multilingual universal sentence encoder for semantic retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 87–94, Online. Association for Computational Linguistics.

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. 2022. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18123–18133.

Liang Zhang, Anwen Hu, and Qin Jin. 2022. Generalizing multimodal pre-training into multilingual via language acquisition. *arXiv preprint arXiv:2206.11091*.

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

, . 2022. . In *The 25th Meeting on Image Recognition and Understanding*.

A Appendix

A.1 Classification on ImageNet series

Lan.	Method	IN-Adv.	IN-Ren.	IN-Ske.	IN-1K	IN-V2	avg.
ES	M-CLIP	54.4	75.4	39.3	41.3	45.7	51.2
	Our	**58.1**	**76.8**	**46.6**	**52.9**	**57.9**	**58.5**
	Imp.	+3.7	+1.4	+7.3	+11.6	+12.2	+7.3
FR	M-CLIP	50.3	71.6	38.3	40.8	44.8	49.2
	Our	**58.6**	**78.1**	**47.9**	**53.3**	**58.4**	**59.2**
	Imp.	+8.3	+6.5	+9.6	+12.5	+13.6	+10.0
RU	M-CLIP	47.4	72.9	36.9	39.5	42.7	47.9
	Our	**50.7**	**76.1**	**44.9**	**49.4**	**54.4**	**55.1**
	Imp.	+3.3	+3.2	+8.0	+9.9	+11.7	+7.2
AR	M-CLIP	46.2	61.7	31.2	32.4	35.7	41.4
	Our	**53.9**	**70.7**	**41.0**	**44.8**	**49.4**	**52.0**
	Imp.	+7.7	+9.0	+9.8	+12.4	+13.7	+10.6

Table 6: **Results on ImageNet variants.** Due to the lack of publicly available models or translation datasets for other languages, we used Google Translate to translate the class names and prompts of ImageNet from English to the target language, thus constructing variations of Imagenet for the other 4 languages.

As shown in Table 6, our proposed AltCLIP$_{M9}$ outperforms M-CLIP on Spanish, French, Russian, and Arabic on the ImageNet series datasets.

A.2 Effects of English-English data

We present results from ablation studies on AltCLIP$_{M2}$. We show the significance of including various parallel data in the Teacher Learning stage in Table 7. As illustrated in the 3rd and 5th lines, without English-to-English parallel data, the accuracy on English ImageNet drastically drops to 15.47 from 53.8. Similarly, excluding machine-translated English-to-Chinese data, has a great impact on the performances on Chinese benchmarks, i.e. Imagenet$_{CN}$ and Flickr30K$_{CN}$, due to influenced Chinese text-image representation. Moreover, empirical experiments show that introducing recall-based parallel data leads to a great improvement in Imagenet$_{CN}$ which may be related to the distribution of the data set. This indicates that the diversity of training data used for teacher learning can benefit the language model to gain more knowledge about entities or concepts.

A.3 Hyper-parameters

As shown in Table 8, we set the hyper-parameters for bilingual and multilingual AltCLIP training.

EN-EN	EN-CN$_{MT}$	EN-CN$_{RB}$	CL	Flickr30K$_{EN}$	Flickr30K$_{CN}$	ImageNet$_{EN}$	ImageNet$_{CN}$
✓	✓	✓	✓	**90.4**	**89.2**	74.5	**59.6**
✓	✓	✓		88.3	87.2	**74.7**	58.2
✓	✓			86.8	85.8	51.6	41.7
✓				86.6	53.9	53.8	12.8
	✓			61.9	85.4	15.5	42.5

Table 7: **Ablation Experiments.** CL indicates the use of the Contrastive Learning stage, while EN-EN, EN-CN$_{MT}$, EN-CN$_{RB}$ refers to parallel data used in the Teacher Learning stage. Specifically, EN-EN indicates the use of English-English text pairs; EN-CN indicates the use of English-Chinese parallel text, including EN-CN$_{MT}$ represents machine translated pairs while EN-CN$_{RB}$ stands for Recall-Based data, i.e TSL2019. All compared models are pre-trained for 10 epochs.

Hyper-paramters	TL	CL
Batch size	11264	1024
Optimizer (AdamW, β)	(0.99, 0.999)	(0.99, 0.999)
Learning rate	2e-4	2e-6
Weight decay	2e-1	5e-2
Eps	1e-8	1e-8
Warmup steps	500	2000
#Epochs	10	1
Gradient clipping	1.0	5.0
Steps	146500	2000

Table 8: Hyper-parameters setting in Teacher Learning Stage and Contrastive Learning Stage.

A.4 Examples for multilingual cross-modal retrieval

As illustrated in Tab. 9, our AltCLIP$_{M9}$ can recall the accurate results.

A.5 Examples for text-image generation

We show more examples generated from our AltCLIP$_{M9}$ guided diffusion model: we use the same prompt and translate it into different lanugages and present the results in Tab. 10. One can observe that the model generates similar images but with subtle differences for different languages.

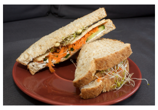
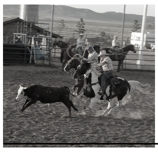
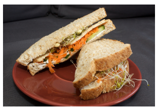
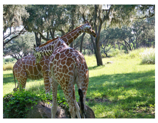
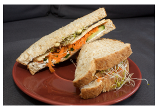
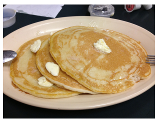
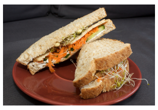
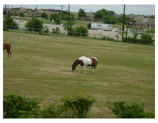
Image	En	Pred.
	**a vegetarian sandwich , cut in half is on a red plate**	**83.2**
	a hoagie sandwich with several vegetables and turkey on it	12.4
	the sandwich is in half on the table next to pickle slices	2.8
	a plate with salad , chips , and large white bread sandwiches with meat	0.9
	soup , a sandwich , a pickle , and some chips are all on a plate	0.4
	**a cow appears to run while two men on horses wearing hats are seen with lassos**	**48.9**
	a man pulling two cows by ropes with a lot of people gathered together	41.3
	horses are running with their faces very close to each other	2.7
	a man on a horse landing on the backside of an obstacle	2.3
	it is always fun to have a good friend along for the ride	1.1
	Fr	
	**un sandwich végétarien , coupé en deux est sur une plaque rouge .**	**84.6**
	un sandwich hoagie avec plusieurs légumes et dinde sur elle .	8.9
	le sandwich est dans la moitié sur la table à côté de tranches de cornichons .	5.7
	Une assiette avec de la salade , des chips et d'énormes sandwichs de pain blanc avec de la viande .	0.5
	soupe , un sandwich , un cornichon , et certaines puces sont tous sur une plaque .	0.1
	**two giraffes standing on all fours next to one another with grass , bushes and trees around them**	**63.7**
	Deux giraffes regardent autour pendant que l'autre se penche pour manger .	23.2
	une mère et un bébé girafe dans les arbres bordé d'un parc animalier .	7.6
	une girafe est debout à côté d'un arbre comme d'autres girafes marchent derrière eux .	1.5
	une girafe dans une enceinte tord sa tête pour manger un peu de feuillage sur un poteau .	1.4
	Es	
	**un sándwich vegetariano cortado por la mitad en un plato rojo**	**57.0**
	sándwich con verduras y pavo	36.5
	sándwich cortado a la mitad sobre la mesa junto a rebanadas de pepinillos	5.3
	sándwich con pepinillo, queso, mostaza, ketchup y mahonesa en un plato con un tenedor	0.4
	un planto con ensalada, patatas y grandes sándwiches con carne	0.3
	**tres tortitas con mantequilla en un plato amarillo con forma ovalada**	**98.2**
	la comida en el plato en la mesa ya está lista para comerse	1.1
	mesa con platos de desayuno y bebidas	0.2
	varias personas sentadas a una mesa con platos con comida	0.1
	mesa llena de platos de comida y dos vasos con bebida	0.1
	It	
	**sandwich vegetariano tagliato a metà su un piatto rosso**	**86.2**
	mezzo sandwich sul tavolo accanto a fettine di sottaceto	11.4
	piatto con insalata, patatine e grandi sandwich di pane bianco con carne	0.7
	panino imbottito con verdure varie e tacchino	0.6
	sandwich con sottaceto, formaggio, senape, ketchup e maionese su un piatto con una forchetta	0.5
	cavallo bianco e marrone che bruca un prato verde	89.2
	cavallo bianco e marrone in piedi su un prato	**10.5**
	mucca bianca e marrone in un pascolo	0.1
	cavallo marrone che bruca l'erba in mezzo a un bosco	0.1
	cavallo con paraocchi legato a un palo in un parcheggio	0.0

Table 9: **A Case Study of multi-lingual retrieval results.** We conduct a case study on the XTD dataset, using our proposed model AltCLIP-M9 for text retrieval in one of four languages (English, French, Italian, Spanish) for each randomly selected image. Our model demonstrated consistent and satisfactory performance in object recognition and understanding spatial relationships across languages, with top 5 most similar texts retrieved from the dataset for each image. Bold text indicates ground truth and prediction (Pred.) value is in %.

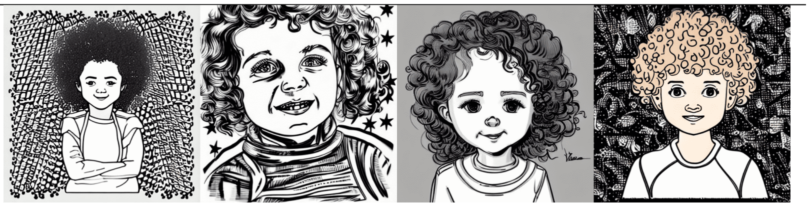
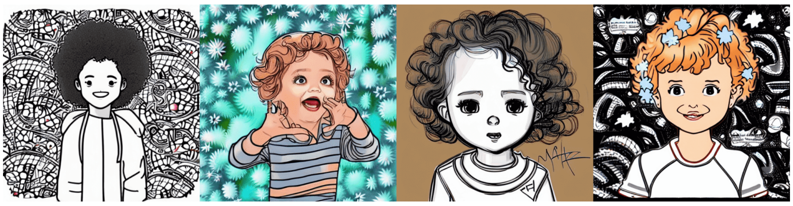
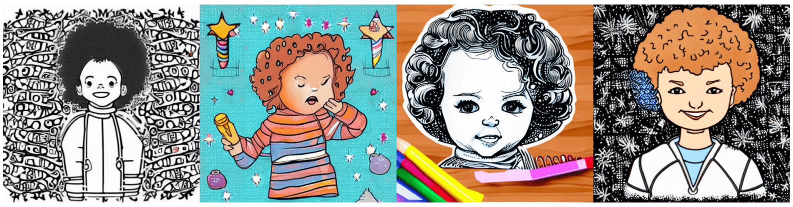
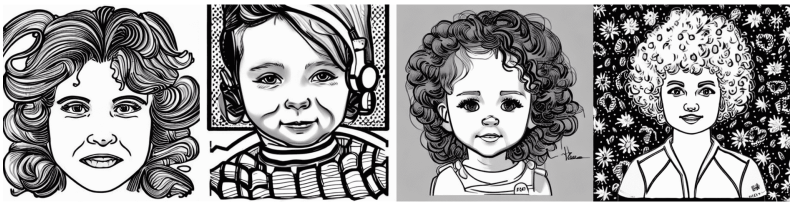
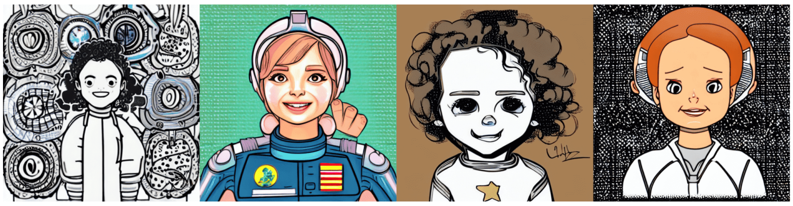
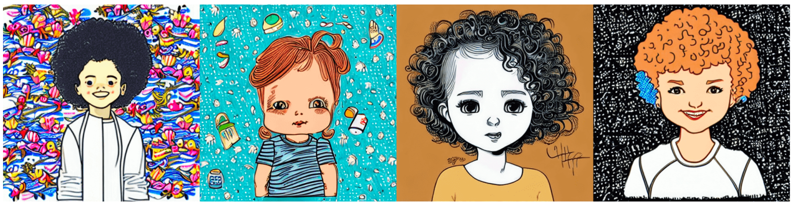
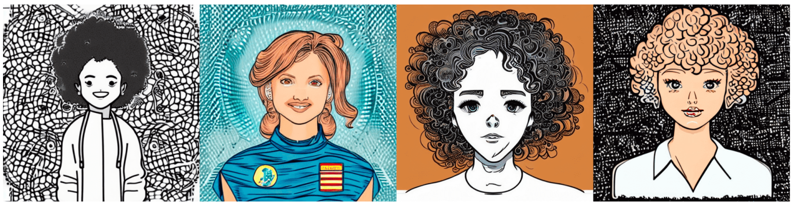
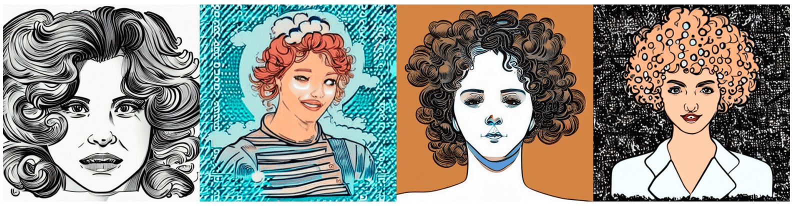
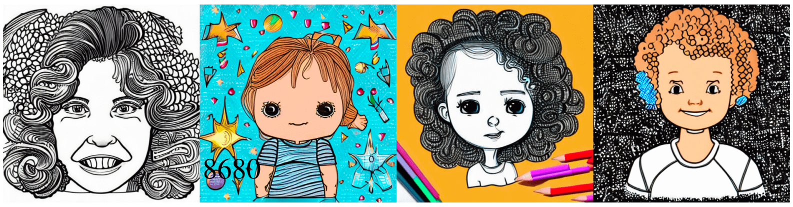
Prompts	Generated Image
EN:clean simple line art of a cute little girl with short wavy curly hair. she is dressed as an astronaut. no background. well composed, clean coloring book page, beautiful detailed face. coloring book line art by artgerm and greg rutkowski and johanna basford and alphonse mucha	
ZH:一个可爱的小女孩的干净简单的线条艺术，短波浪卷发。她打扮成宇航员。没有背景。构图良好，涂色书页面干净，面部细节优美。Artgerm和Greg Rutkowski以及Johanna Basford和Alphonse Mucha的着色书线条艺术	
FR:Nettoyer l'art au trait simple d'une petite fille mignonne avec des cheveux courts et bouclés ondulés. Elle est habillée en astronaute. Pas d' antécédents. Bien composé, page de livre à colorier propre, beau visage détaillé. Dessin de ligne de livre de coloriage par Artgerm et Greg Rutkowski et Johanna Basford et Alphonse Mucha	
SP:Arte de línea simple y limpia de una linda niña con cabello corto y rizado ondulado. Está vestida de astronauta. sin antecedentes. Bien compuesto, página de libro para colorear limpio, hermosa cara detallada. Coloring Book Line Art por Artgerm y Greg Rutkowski y Johanna Basford y Alphonse Mucha	
RU:чистая простая линия искусства милой маленькой девочки с короткими волнистыми вьющимися волосами. она одета как астронавт. нет предыстории. хорошо составленная, чистая раскраска, красивое детализированное лицо. раскраска линии артгерма и Грега Рутковски и Джоанны Басфорд и Альфонса Мухи	
AR: فن الخط النظيف البسيط لفتاة صغيرة لطيفة ذات شعر قصير مموج مجعد. كانت ترتد ي زي رائدة فضاء. أي خلفية. مؤلفة بشكل جي د ، صفحة كتاب تلوين نظيفة ، وجه مفصل ج Artgerm ميل. كتاب تلوين فن الخط بواسطة و Johanna Basford و Greg Rutkowski و Alphonse Mucha	
JA:短いウェーブのかかった巻き毛を持つかわいい女の子のきれいなシンプルなライン アート。 彼女は宇宙飛行士の格好をしています。 背景なし。 よく構成された、きれいな塗り絵ページ、美しい詳細な顔。 artgerm と greg rutkowski と johanna basford と alphonse mucha による塗り絵の線画	
KO: 짧은 물결 모양의 곱슬머리를 가진 귀여운 소녀의 깨끗하고 단순한 라인 아트. 그녀는 우주 비행사 옷을 입고 있습니다. 배경이 없습니다. 잘 구성되고 깨끗한 컬러링 북 페이지, 아름다운 디테일한 얼굴. artgerm 및 greg rutkowski 및 johanna basford 및 alphonse mucha의 색칠하기 책 라인 아트	
IT: linea semplice e pulita di una bambina carina con capelli ricci corti e ondulati. è vestita da astronauta. nessuno sfondo. pagina del libro da colorare ben composta, pulita, bel viso dettagliato. disegno al tratto del libro da colorare di artgerm e greg rutkowski e johanna basford e alphonse mucha	

Table 10: The images generated by AltCLIP$_{M9}$-guided Diffusion with the same prompt translated to nine languages and a fixed seed.

ACL 2023 Responsible NLP Checklist

A For every submission:

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

☒ A2. Did you discuss any potential risks of your work?
Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Section 1 (Introduction)

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

B ☑ Did you use or create scientific artifacts?

2, 3

☑ B1. Did you cite the creators of artifacts you used?
1,2,5

□ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.

☑ 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)?
5

☒ 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?
Left blank.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
2, 5

☑ 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.
5

C ☑ Did you run computational experiments?

2

□ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
No response.

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?
3

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
3

☑ 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.)?
3

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

Left blank.

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

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

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

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

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