M³P: Towards Multimodal Multilingual Translation with Multimodal Prompt

Jian Yang♠, Hongcheng Guo♠, Yuwei Yin★, Jiaqi Bai♠, Bing Wang♠
Jiaheng Liu♠, Xinnian Liang♠, Linzheng Chai♠, Liqun Yang♠†, Zhoujun Li♠
♠ State Key Lab of Software Development Environment, Beihang University
★ Department of Computer Science, University of British Columbia
{jiaya, hongchengguo, bjq, bingwang}@buaa.edu.cn; yuweiyin@cs.ubc.ca
{liujiaheng, xnliang, challenging, lqyang, lizj}@buaa.edu.cn

Abstract

Multilingual translation supports multiple translation directions by projecting all languages in a shared space, but the translation quality is undermined by the difference between languages in the text-only modality, especially when the number of languages is large. To bridge this gap, we introduce visual context as the universal language-independent representation to facilitate multilingual translation. In this paper, we propose a framework to leverage the multimodal prompt to guide the **M**ultimodal **M**ultilingual neural **M**achine **T**ranslation (M³P), which aligns the representations of different languages with the same meaning and generates the conditional vision-language memory for translation. We construct a multilingual multimodal instruction dataset (`InstrMulti102`) to support 102 languages Our method aims to minimize the representation distance of different languages by regarding the image as a central language. Experimental results show that M³P outperforms previous text-only baselines and multilingual multimodal methods by a large margin. Furthermore, the probing experiments validate the effectiveness of our method in enhancing translation under the low-resource and massively multilingual scenario.

Keywords: Multimodal Multilingual Translation, Multimodal Instruction Tuning, Contrastive Learning

1. Introduction

Multilingual neural machine translation (MNMT) models relying on text data of multiple languages support diverse translation directions in a single shared model (Arivazhagan et al., 2019; Yang et al., 2021a). Beyond that, multimodal NMT captures the visual context from relevant images of the source sentences, bringing a further enhancement of multilingual translation (Zhang et al., 2020b; Li et al., 2021a, 2022; Fang and Feng, 2022; Guo et al., 2022b). As a language-agnostic semantic representation, the image plays a bridge role in translating sentences across different languages. It is intuitively promising that images can serve as a universal router in multilingual translation.

However, previous multimodal NMT works (Li et al., 2021a, 2022) mainly focus on the bilingual translation supervised by the image-sentence training data. In Figure 1(a), each bilingual model can only handle a single translation direction compared to existing thousands of languages in the world. MNMT involves more languages using available linguistic resources but only implicitly brings different languages together by sharing the same parameters. There still exists a gap between different translation directions. Some previous works (Pan et al., 2021; Yang et al., 2021b; Winata et al., 2021; Gong et al., 2021) propose to leverage the aligned augmentation and contrastive learning across multiple languages only on the language modality. Meanwhile, images are regarded as the universal language to communicate ideas and concepts effectively across linguistic and cultural barriers (Lu et al., 2019; Zeng et al., 2022). Hence, minimizing the difference across diverse directions by vision-language pair requires further exploration.

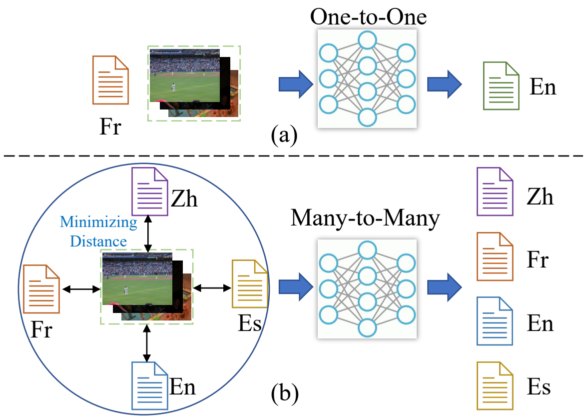


Figure 1: Comparison between (a) the bilingual translation baseline and (b) our proposed M³P.

To explicitly bridge the gap among the multiple languages, we propose a multimodal prompt-based framework for multimodal multilingual neural machine translation (M³P), which enables different translation directions between multiple source and target languages with the help of universal

† Corresponding author.
https://huggingface.co/datasets/CSJianYang/InstrMulti102

LREC-COLING 2024, pages 10858–10871

visual features. Specifically, we use the cross-lingual language encoder to extract the multilingual representation from the text data and the vision Transform encoder to derive the visual context. A designed multimodal prompt can be fed into the encoder-decoder model and decoder-only model (Llama2) (Liu et al., 2023b) to verify the motivation of our work. Multimodal multilingual contrastive learning (MMCL) with masked language/image augmentation is used to align two modalities into a common semantic space. Then, we consider language representations as the query based on visual features as key and value to attend the multi-head cross-attention to generate the conditional vision-language memory (CVLM) as the encoder states. Finally, the multilingual language decoder predicts the target translation given conditional vision-language memory.

Our method is effective for multilingual translation even for the massively translation of 102 languages. Experimental results on the supervised translation directions demonstrate that our method substantially outperforms previous text-only and multilingual multimodal methods by nearly +1∼+4 BLEU points. Our method is further evaluated on `InstrMulti102` to validate the essence of the multilingual multimodal contrastive learning (MMCL). Analytic experiments emphasize the importance of alignment in both the multilingual text and vision modality, leading to better performance.

2. Our Method

2.1. Overview

In Figure 2, our proposed model M^{3}P consists of the cross-lingual language encoder, vision Transformer encoder, and the multilingual language decoder. Specifically, given the k-th sentence pair (x^k, y^k) with the image z^k, we first use the cross-lingual pre-trained language model to encode the source concatenation, where the target language symbol is prefixed into the source sentence to indicate the direction. Meanwhile, we reshape the image $z^k \in \mathcal{R}^{H\times W\times C}$ into a sequence of flattened patches and extract the vision context s^k by the vision Transformer. To reduce the gap among different languages, the image is regarded as the central language to explicitly bring different languages to a shared semantic space using multilingual multimodal contrastive learning (MMCL). Then, we incorporate the language encoder states $s^k = \{s_u^k\}_{u=1}^U$ with U tokens and auxiliary vision encoder states $h^k = \{h_v^k\}_{v=1}^V$ with V tokens to generate the conditional vision-language memory (CVLM) as the final encoder states. Finally, $\{e_u^k\}_{u=1}^U$ is fed into multilingual language decoder $\mathcal{D}$ to predict the target translation y^k.

2.2. Multilingual Multimodal Translation

Given M bilingual corpora with images $D_{all} = \{D_m\}_{m=1}^M$, where M denote the number of the training corpora of N languages $L_{all} = \{L_n\}_{n=1}^N$ and L_n denote the n-th language. Each bilingual corpus with images $D_m = \{x^k, y^k, z^k\}_{k=1}^K$ from D_{all} consists of the source sentences, target sentences, and corresponding images. The training objective of multilingual multimodal translation can be described as:

$$\mathcal{L}_m = -\sum_{m=1}^{M} \mathbb{E}_{x^k,y^k,z^k \in D_m}\left[\log P(y^k|x^k, z^k; \Theta)\right] \tag{1}$$

where the multimodal multilingual model employ complete shared parameters Θ for all translation directions. We adopt Transformer as the backbone model for language and vision encoding, where the multilingual pre-trained model XLM-R (Conneau et al., 2020) and the pre-trained model CLIP (Radford et al., 2021) are used to initialize the language and vision encoder. The target symbol (e.g., `[En]` or `[De]`) is prefixed to the source sentence to indicate the direction.

Multilingual Multimodal Prompt. Given the source sentence x^k, image z^k, and its translation y^k, we construct the multimodal prompt as the whole input for the decoder-only model (e.g. Llama2) in Figure 3(a), where L_i and L_j are the source and target language. For the raw image z^k, we use the vision model to encode the image into U visual tokens $h^k = \{h_v^k\}_{v=1}^V$. For the encoder-decoder setting in Figure 3(b), we separately fed the source tokens x^k into the text encoder and image tokens z^k into the vision encoder.

2.3. Multimodal Encoding

For the encoder-decoder setting, given the text prompt, we encode the concatenation of U tokens with the language Transformer encoder to obtain the language representations s^k:

$$s^k = \{s_u^k\}_{u=1}^U = \mathcal{S}(t_{L_j}, x^k) \tag{2}$$

where $\mathcal{S}$ denotes the language encoder and the $s^k = \{s_u^k\}_{u=1}^U$ are language features.

Similarly, to encode the image $z^k \in \mathcal{R}^{H\times W\times C}$ with H height, W width, and C channels, we reshape the image $z^k \in \mathcal{R}^{H\times W}$ into a sequence of flattened patches $h \in \mathcal{R}^{V\times(P^2\times C)}$, where P is the resolution of the each patch and $V = \frac{H\times W}{P^2}$ is the number of patches. Given the original image z^k, based on the Transformer encoder $\mathcal{H}$, the source language tokens $\{s_f\}_{f=1}^F$ are extracted as:

$$h^k = \{h_v^k\}_{v=1}^V = \mathcal{H}(z^k) \tag{3}$$

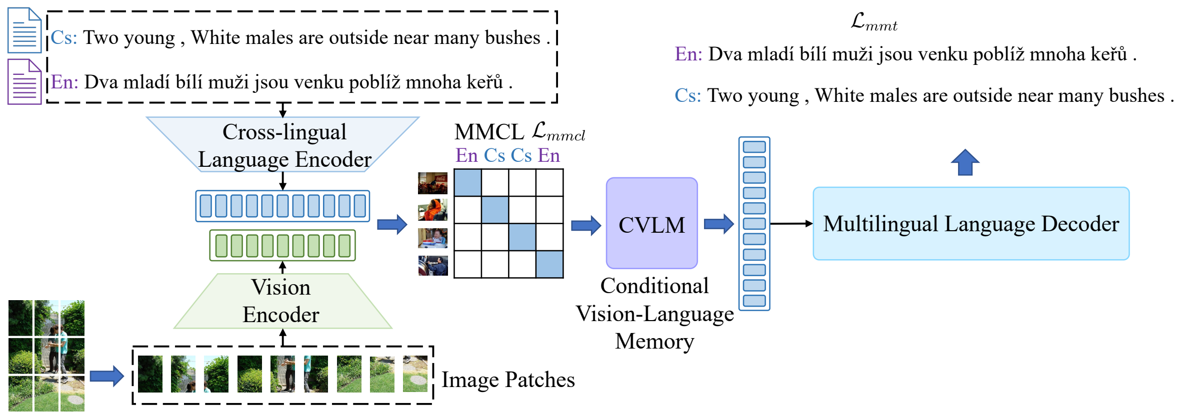


Figure 2: Overview of our method. $s^k = \{s_u^k\}_{u=1}^U$ denotes the representations of the source sentence of U tokens. We reshape the original image $z^k \in \mathcal{R}^{H \times W \times C}$ into V patches and then encoded as $h^k = \{s_v^k\}_{v=1}^V$ with the vision Transformer. Given the source and visual representations s^k and h^k, the multilingual multimodal contrastive learning (MMCL) adopted to minimize the distance between s^k of different languages and h^k, which greatly encourages multilingual multimodal agreement in a shared space. Conditioned on the image tokens as (key,value), the language features as the query attend the multi-head attention to generate final encoder states $e^k = \{e_u^k\}_{u=1}^U$ as conditional vision-language for multilingual translation.

(a) *Decoder-only Prompt:*
{Decoder}
Below is an instruction that describes a task. Write a response that appropriately completes the request.
Instruction:
Please translate the following sentence from
$\{L_i\}$ to $\{L_j\}$: $\{z^k\}$ $\{x^k\}$
Response:
$\{y^k\}$
(b) *Encoder-Decoder Prompt:*
{Text Encoder}:
Please translate the following sentence from
$\{L_i\}$ to $\{L_j\}$: $\{x^k\}$
{Vision Encoder}:
$\{z^k\}$
{Vision Decoder}:
$\{y^k\}$

Figure 3: Multimodal prompt for LLM.

where $\mathcal{H}$ denotes the vision Transformer encoder and the $h^k = \{h_v^k\}_{v=1}^V$ are vision representations.

For the decoder-only setting, we first leverage the vision extractor to obtain the visual tokens and fill them into the prompt in Figure 3. All tokens are concatenated as a whole into a large language model for the final representations. Then, we can similarly get the text representations s^k and h^k for the following operations.

2.4. Multilingual Multimodal Alignment

To effectively fuse the multilingual text and vision features, the image can be regarded as the universal language to bridge the gap among different languages. We introduce multilingual multimodal contrastive learning (MMCL) to further improve text-image alignment and multilingual text-text alignment. We use the InfoNCE objective (van den Oord et al., 2018) to learn the correspondence between image and text. In particular, we minimize the sum of two multi-modal contrastive losses:

$$\mathcal{L}_c = \sum_{x^k, z^k \in D_{all}} \left(f(x^k, z^k) + f(z^k, x^k)\right) \quad (4)$$

where D_{all} is the multilingual dataset that contains sampled multilingual image-text pairs. $f(z^k, x^k)$ and $f(x^k, z^k)$ are the contrastive loss on image-to-text similarity and text-to-image similarity. Specifically, the image-to-text contrastive loss is:

$$f(x^k, z^k) = -\log \frac{\exp\left(z^k \cdot x^k / \tau\right)}{\sum_{x \in \{x^k, x^-\}} \exp\left(z^k \cdot z / \tau\right)} \quad (5)$$

where τ is a temperature hyper-parameter, x^k are *positive* embedded text clips overlapping with image clip embedding z^k, and x^- are *negative* embedded text clips that are implicitly formed by other text clips in the training batch B. Symmetrically, the text-to-image loss $f(z_t, z_v)$ is defined as:

$$f(z^k, x^k) = -\log \frac{\exp\left(z^k \cdot x^k / \tau\right)}{\sum_{x \in \{x^k, x^-\}} \exp\left(z^k \cdot z / \tau\right)} \quad (6)$$

To construct the multilingual text in the training batch B and balance multiple bilingual corpora, we adopt a temperature-based sampling method to collect sentences of different languages in a single batch using sampling probabilities $q_1, \ldots, q_M$:

$$q_m = \frac{(|D_m|/|D_{all}|)^{\frac{1}{\tau}}}{\sum_{i=1}^{M}(|D_i|/|D_{all}|)^{\frac{1}{\tau}}} \quad (7)$$

where $|D_m|$ is the size of training dataset D_m. The temperature gradually increases to the peak value for several epochs. The temperature is calculated by $\tau_i = \min(\tau, \tau_0 + \frac{i}{\mathcal{W}}(\tau - \tau_0))$, where τ_0 and τ separately denote the initial and peak temperature, and $\mathcal{W}$ is the number of warming-up epochs.

2.5. Multilingual Multimodal Augmentation

Our goal is to learn to model multilingual image-text alignment by using difficult examples in the multilingual multimodal contrastive objective. We construct negatives in our training batch by using masked language/image modeling, which are semantically similar to the original sentence.

For image augmentation, we leverage the function $\mathcal{I}(\cdot)$ to augment the original image by cropping, resizing, rotation, cutout, color distortion, Gaussian blur, and Sobel filtering. Then, we divide an image into regular non-overlapping patches and mask the chosen patches sampling from a uniform distribution as masked image modeling.

For the multilingual text, we randomly mask some random spans of contiguous tokens. For each sentence, we adopt the multilingual data augmentation $\mathcal{T}(\cdot)$ to augment the original sentence of different languages. The augmented source sentence and the image $\{\mathcal{I}(x^k), \mathcal{T}(z^k)\}$ with multilingual multimodal augmentation (MMA) is used to enhance the contrastive learning to learn the specific representational invariances.

2.6. Conditional Vision-Language Memory

Given the source concatenation $s^k = \{s^k\}_{u=1}^{U}$ from the language encoder, the language is regarded as the main input (key) with the auxiliary by the multi-head cross-attention as:

$$e^k = \Big\Vert_{a=1}^{A} \sigma\left(\frac{(W_Q^a h^k)(W_Q^a s^k)^\top}{\sqrt{C}}\right)(W_V^a s^k) \quad (8)$$

where $\Vert_{a=1}^{A}$ is the concatenation operator of A attention heads and σ denotes the softmax operation. W_K^a, W_Q^a, and W_V^a are respectively the corresponding linear projection matrix of the query, key, and value for a-th head. C denotes the number of feature channels. $e^k = \{e^k\}_{u=1}^{U}$ are encoder representations, which will be fed into the decoder.

2.7. Multilingual Generation

Our method can be split into the text translation and the image caption task. To effectively train the text encoder, our model predicts the target words only based on the source language as below:

$$y_t^k = \mathcal{D}(y_{1:t-1}^k, s^k; \theta) \quad (9)$$

where $\mathcal{D}$ denotes the standard Transformer decoder and the y_t^k is the t-th target word conditioned on the previous $t-1$ tokens $y_{1:t-1}^k$.

Similarly, to align the image and target language, we adopt the training objective of image caption only based on the image context as below:

$$y_t^k = \mathcal{D}(y_{1:t-1}^k, h^k; \theta) \quad (10)$$

Given the conditional vision-language memory $e^k = \{e^k\}_{u=1}^{U}$ containing language and vision information, we adopt the standard Transformer decoder to predict the target words sequentially as:

$$y_t^k = \mathcal{D}(y_{1:t-1}^k, e^k; \theta) \quad (11)$$

where $\mathcal{D}$ is the language Transformer decoder and the y_t^k is the t-th target word conditioned on the previous $t-1$ tokens $y_{1:t-1}^k$. In our work, our model is separately trained on the objective Eq 9 and Eq 10 for 25% time and on Eq 11 for 50%, which is denoted as the Multimodal DropNet (MDropNet).

2.8. Training Objective

During training, M^{3}P is optimized by jointly minimizing the multilingual multimodal contrastive training objective from Equation 1 and translation objective from Equation 9∼11:

$$\mathcal{L}_{all} = \mathcal{L}_m + \lambda \mathcal{L}_c \quad (12)$$

where λ is the coefficient to balance the translation objective and multilingual contrastive objective.

3. Experiments

3.1. Datasets

Multi30k. We conducted experiments on the widely used Multi30k benchmark (Elliott et al., 2016). The training and valid sets contain 29K and 1K sentences, respectively. The dataset contains four languages and each sentence pair has a corresponding image, including English (En), German (De), French (fr), and Czech (Cs). We reported the results on the Flickr2016, Flickr2017, Flickr2018, and MSCOCO test sets (Lin et al., 2014; Barrault et al., 2018), where MSCOCO is the out-of-domain dataset with ambiguous verbs.

		En→Fr	En→Cs	En→De	Fr→En	Cs→En	De→En	Avg_6
Only Trained on Text Data								
1→1	BiNMT (Vaswani et al., 2017)	63.3	33.4	39.9	54.0	41.1	43.8	45.9
N→N	MNMT (Fan et al., 2021)	63.8	34.0	40.2	52.0	41.3	42.5	45.6
Trained on Text and Vision Data								
1→1	BiNMT (Vaswani et al., 2017)	63.5	33.0	40.3	55.1	41.8	44.1	46.3
N→N	MNMT (Gated Fusion) (Li et al., 2021a)	63.8	34.4	41.0	51.5	41.1	43.3	45.8
	MNMT (Concatenation) (Li et al., 2021a)	63.0	33.8	38.8	53.3	43.6	44.0	46.1
	mRASP2 (Pan et al., 2021)	63.8	34.4	41.3	53.2	44.0	44.5	46.9
	Selective Attn (Li et al., 2022)	63.5	34.4	41.3	53.2	44.0	44.5	46.8
	LVP-M^3 (Guo et al., 2022b)	63.4	34.1	41.4	53.2	44.0	44.5	46.8
	M^3P (Encoder-Decoder)	**64.8**	**35.2**	**41.8**	**53.8**	**44.8**	**45.0**	**47.6**
	M^3P (Decoder-only)	**66.4**	**38.1**	**43.5**	**56.7**	**46.9**	**48.1**	**49.9**

Table 1: X→En and En→X evaluation results for bilingual (1→1) and many-to-many ($N \to N$) models on the Flickr2016 test set.

		En→Fr	En→De	De→En	Fr→En	Avg_4	En→Fr	En→De	Fr→En	De→En	Avg_4
		Flick2017					MSCOCO				
Only Trained on Text Data											
1→1	BiNMT (Vaswani et al., 2017)	55.4	34.1	39.2	43.4	43.0	45.8	32.1	40.6	34.3	38.2
N→N	MNMT (Fan et al., 2021)	56.8	34.9	40.3	44.6	44.2	45.9	31.9	41.6	34.6	38.5
Trained on Text and Vision Data											
1→1	BiNMT (Vaswani et al., 2017)	55.8	34.6	39.6	43.6	43.4	45.8	32.3	41.6	34.4	38.5
N→N	MNMT (Gated Fusion) (Li et al., 2021a)	56.8	34.3	40.3	44.2	43.9	46.8	32.5	42.2	34.5	39.0
	MNMT (Concatenation) (Li et al., 2021a)	56.4	34.0	39.4	43.8	43.4	46.4	32.6	42.4	34.1	38.9
	mRASP2 (Pan et al., 2021)	57.0	35.1	39.6	44.1	43.9	47.1	32.7	42.3	34.8	39.2
	Selective Attn (Li et al., 2022)	56.6	34.2	40.3	44.4	43.9	46.8	32.5	42.5	34.3	39.0
	LVP-M^3 (Guo et al., 2022b)	57.4	34.4	40.4	44.7	44.2	46.8	32.5	42.6	34.5	39.1
	M^3P (Encoder-Decoder)	**57.4**	**35.3**	**41.0**	**45.6**	**44.8**	**46.8**	**33.1**	**43.2**	**35.2**	**39.6**
	M^3P (Decoder-only)	**58.3**	**37.2**	**42.2**	**46.5**	**46.1**	**47.4**	**34.2**	**44.5**	**36.2**	**40.6**

Table 2: X→En and En→X evaluation results for bilingual (1→1) and many-to-many ($N \to N$) models on the Flickr2017 test set and MSCOCO test set.

3.2. Baselines

Text-only Methods. **BiNMT** (Conneau et al., 2020) adopt the Transformer backbone initialized by XLM-R and then only trained on single translation direction. **MNMT** (Fan et al., 2021) is jointly trained on all multilingual data, where the target language symbol is prefixed to the input sentence.

Multimodal Methods. **BiNMT** (Vaswani et al., 2017) is the bilingual Transformer model concatenating the language and visual feature. **MNMT (Gated Fusion)** and **MNMT (Concatenation)** (Li et al., 2021a) use the visual context using the gated fusion and concatenation unit, respectively. We apply the **mRASP2** (Pan et al., 2021) on the multimodal translation with the text-only contrastive learning. **Selective Attn** (Li et al., 2022) use a single-head attention network to correlate words with image patches. **LVP-M^3** uses the language-aware visual prompt to guide the multimodal translation. For a fair comparison, all the language encoders are initialized by XLM-R (Conneau et al., 2020) and the vision encoders are initialized by CLIP (Radford et al., 2021).

Model	Zh→En	Hi→En	Th→En	Avg_{101}
Text-only MNMT	14.3	13.5	11.1	14.3
MNMT (Gated Fusion)	15.2	14.3	12.1	15.4
MNMT (Concatenation)	15.1	14.6	13.1	15.8
M^3P (Encoder-Decoder)	**16.8**	**15.2**	**14.8**	**18.2**
M^3P (Decoder-only)	**18.2**	**16.4**	**16.5**	**21.2**

Table 3: Massively multilingual translation average results (101 translation directions) on `Instr-Multi102`.

Model	En→Fr	En→De	Fr→En	De→En
Text-only MNMT	63.8	40.2	52.0	42.5
ResNet50	64.2	40.6	52.3	43.1
ResNet101	64.4	40.8	52.4	43.4
ViT-B/32	64.8	41.6	**53.8**	45.0
ViT-B/16	65.1	41.8	53.6	44.8
ViT-B/14	**65.2**	**41.9**	53.4	**45.2**

Table 4: Comparison of different vision backbones (e.g., CNN and Transformer backbones) on the Flickr2016 test set.

3.3. Training and Evaluation

For the encoder-decoder setting, our model comprises a language encoder initialized by the cross-lingual language pre-trained encoder XLM-R (Con-

ID	Flickr2016	En→De	De→En
①	M^{3}P (our method)	**41.6**	**45.0**
②	① - MMCL	41.2	44.6
③	② - CVLM	40.8	44.0
④	③ - MDropNet	40.5	43.8
⑤	④ - Multilingual Training	40.1	43.2

Table 5: Ablation study of the different modules on Flickr2016. M^{3}P is the final model of our method.

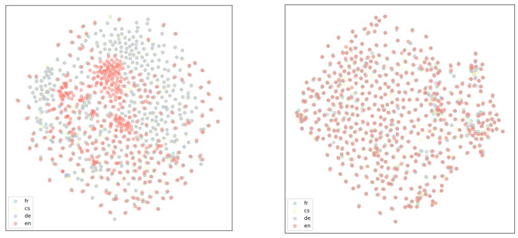
neau et al., 2020) and a vision encoder initialized by CLIP (Radford et al., 2021), We train multilingual models with Adam (Kingma and Ba, 2015) ($\beta_1 = 0.9$, $\beta_2 = 0.98$). For the decoder-only setting, we use the Llama2 (Liu et al., 2023b) for text generation and CLIP for vision extractor. The learning rate is set as 5e-4 with a warm-up step of 4,000. The models are trained with the label smoothing cross-entropy with a smoothing ratio of 0.1. Our model comprises a vision encoder, language encoder, and language decoder, which all consist of 12 layers with 768 hidden size and share the same embedding matrix. For the multilingual training, the batch size is 2048 tokens on 8 Tesla V100 GPUs. The evaluation metric is the case-sensitive detokenized sacreBLEU[1].

3.4. Results

Flickr Test Set. In Table 1 and 2, M^{3}P clearly improves multilingual baselines by a large margin in 6 translation directions. Previously, text-only MNMT underperforms bilingual translation on average. Further, **MNMT (Gated Fusion)** and **Concatenation** introduce the image as the auxiliary context to enhance translation, but these methods ignore the alignment of different languages. **mRASP2** further adopt the text-text contrastive learning scheme to close the gap among representations of different languages. M^{3}P extract the visual and language features with the Transformer encoder and fuse them for translation in a shared space by the MMCL and CVLM.

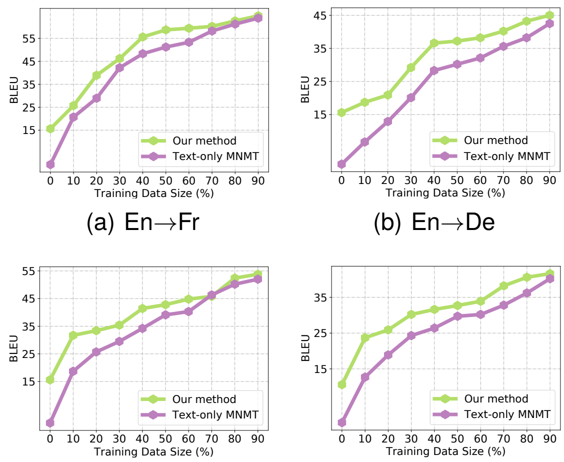
MSCOCO Test Set. In Table 2, we report the performance of the previous baselines and our method on the MSCOCO test set, which is more challenging for MMT models due to the out-of-domain instances with ambiguous verbs. Therefore, it more relies on the image context for disambiguation. Our method outperforms the bilingual baseline by a large margin due to the fusion of text and image.

[1] https://github.com/mjpost/sacrebleu



(a) (b)

Figure 4: Visualization of the sentence average encoder representations of all languages from the multilingual baseline (a) and our multilingual model supervised by the image context (b). Each color denotes one language.



(a) En→Fr (b) En→De

(c) Fr→En (d) De→En

Figure 5: The performance of our method on Flickr2016 (a) En→fr, (b) En→De, (c) Fr→En, and (d) De→En with different sizes of training data on Flickr2016.

4. Massively Multilingual Translation

Considering the existing multimodal translations are limited to only a few languages, we break the limits of multilingual multimodal machine translation by extending the number of used languages in the previous benchmark Multi30k.

Data Construction. We introduce a massive multilingual multimodal machine translation dataset, called `InstrMulti102`, originating from the previous dataset Multi30k (Elliott et al., 2016). Here, we describe the details of the `InstrMulti102`. We use the text-only multilingual Microsoft translator (Yang et al., 2021a) to construct `InstrMulti102` by translating the English data to other 101 languages (Please refer to Appendix A for more details). The many-to-one multilingual model are jointly trained on the expanded dataset of 102 languages and then evaluated on the test set.

(a) Original (b) En (c) De (d) Fr (e) Cs

Figure 6: Representative examples of vision-language alignment from the CVLM of four languages between image patches. Brighter colors represent a higher attention value.

Main Results. In Table 3, we can see that all multilingual models with visual context perform better than the text-only baselines in terms of average BLEU. This shows that image information as the auxiliary context brings more significant improvement in the massively multilingual translation by nearly +4 BLEU points. The visual features of different languages from ViT encoder is successfully projected into the shared semantic.

5. Ablation Study

Performance on Different Backbones. In Table 4, we compare the results of M^{3}P by using the different vision backbones, including ResNet and Transformer (Radford et al., 2021). In Table 4, we observe that M^{3}P with the Transformer backbone outperforms the counterpart with CNN network. It shows that our method can unify the two views of visual and language data in the Transformer backbone. Besides, the vision Transformer with smaller patch size (ViT-B/14) gets the better performance but generates longer visual tokens for computation compared to ViT-B/32 and ViT-B/16. Therefore, we recommend the ViT-B/32 for efficiency or ViT-B/16 for performance as the vision encoder backbone.

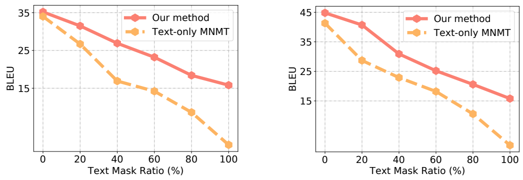
Effect of Different Modules. Table 5 summarizes the ablation study of our proposed modules, which shows that each approach has a significant contribution to the final model. Our multilingual model is first trained on the multilingual data, where the model is denoted as ④ in contrast to bilingual model ⑤. Given the sentence pair with image, we adopt the visual representations to enhance the translation. The performance of multimodal translation is improved by the alternative training strategy (MDropNet), where the model is randomly trained with visual or language tokens (③). Since the source sentences are more important for translation than images, CVLM uses the language tokens as query and visual tokens as (key, value) for cross-attention, which we denoted as ②. We further introduce MMCL to explicitly narrow the gap among different languages. Putting them all together, we obtain the final model ① **M^{3}P**, which proves the effectiveness of progressive learning that can gradually improve performance in different aspects.

6. Analysis

Distance of Different Languages. The image as a universal language is used to narrow the distance among multiple languages, we visualize the sentence representations of the last language encoder layer. We select 500 parallel sentences from the valid set of four languages, including English, German, French, and Czech. Then, we apply t-SNE (van der Maaten and Hinton, 2008) to reduce the 1024-dim representations to 2-dim. It is clear in Figure 4 that text-only MNMT cannot align the 4 languages. By contrast, M^{3}P draws the representations across 3 languages much closer.

Low-resource Setting. To further analyze the performance of M^{3}P given different sizes of downstream parallel data with image context, we randomly extract P percentage of the whole sentence pairs of different languages as the fine-tuned parallel data from the Multi30k dataset. We set $P = \{10\%, 20\%, \ldots, 100\%\}$ and compare our method with the text-only MNMT model. Figure 5 shows the BLEU points of our pre-trained multilingual model and the baseline on four directions, including En→De, En→Fr, De→En, and Fr→En. When the parallel data size is small, the baseline without pre-trained model produces unsatisfactory results. Similarly, in Figure 5(a), M^{3}P fine-tuned on nearly 90% data defeats the baseline trained on all pairs, exemplifying the effectiveness of our method in low-resource scenarios.

Vision-Language Alignment. The function of multimodal multilingual contrastive learning is used to align vision and language, which aims to project vision and language into the same space. In Figure 6, we visualize the conditional vision-language alignment (CVLM) between the source sentence of different languages and image patches. For example, Figure 6 plots the original sentence and Figure 6(b) shows cross-attention between English



(a) (b)

Figure 7: Comparison between the text-only MNMT and M^{3}P when the source sentence is masked with different ratios.

sentence "A young child is standing alone on some jagged rocks." and image patches. Similarly, Figure 6(c) describes the attention about the German counterpart "Ein kleines Kind steht allein auf einem zerklüfteten Felsen." We can oberserve that given diffenet sentences with the same meaning tends to pay attention to the similar image regions, such as jagged rocks and zerklüfteten Felsen. Figure 6(b)$\sim$6(e) indicate our method effectively force the model to learn the similar vision-language attention pattern and project different languages into the same semantic space using MMCL and CVLM.

Sanity Check on Visual Context. In Figure 7, we compare our M^{3}P with the text-only multilingual model to emphasize the necessity of visual context by masking source words with different mask ratios $\{0\%, 20\%, 40\%, 60\%, 80\%, 100\%\}$. When the source sentence is masked, the visual context provide the supplementary information to help translation correctly. When only receiving the source language, the performance of MNMT is obviously worse than M^{3}P, where the visual representations from vision encoder can compensate for the masked words. When the mask ratio is 0%, MNMT can not perform translation since the all words are masked while M^{3}P outperforms MNMT by a large margin (nearly 15 BLEU points). Despite that all source words are masked, M^{3}P can perform image caption under this extreme scenario due to MDropNet.

7. Related Work

Multilingual Multimodal Translation. Multilingual Neural Machine Translation (MNMT) aims to support multiple translation directions by sharing parameters. Recent works (Aharoni et al., 2019; Zhang et al., 2020a; Bapna et al., 2022; Yang et al., 2021a,b, 2022b,a, 2023; Gu et al., 2018; Wang et al., 2023c; Liu et al., 2022) scale to the massively multilingual setting to support more languages. Despite these benefits, the multilingual model tends to underperform its bilingual counterparts with worse translation performance (Arivazhagan et al., 2019). Multimodal machine translation (MMT) refers to the process of translating content that includes both text and images from one language to another, which is a challenging task that aims to enhance source-target translation extra visual context. Researchers propose different attention mechanisms to incorporate language and vision features based on the encoder-decoder architecture (Caglayan et al., 2018; Yao and Wan, 2020; Yin et al., 2020; Wu et al., 2021; Fang and Feng, 2022; Li et al., 2021a; Guo et al., 2022b, 2023) and decoder-only models (Zhu et al., 2023). Vision-language pre-trained models have the ability to process visual information and understand natural language jointly (Lu et al., 2019; Chen et al., 2020; Su et al., 2020; Huang et al., 2022; Radford et al., 2021; Xu et al., 2021). These vision-language models (Lu et al., 2019; Tan and Bansal, 2019; Radford et al., 2021) perform remarkably on various benchmarks and demonstrated to be effective in a range of tasks, including image and video captioning (Tang et al., 2021), visual question answering (Wang et al., 2021, 2022a,b), and multimodal machine translation (Yawei and Fan, 2021).

Large Language Model. Large language models (LLM) (Liu et al., 2024; Wang et al., 2023b) have emerged as a significant milestone in the field of natural language processing, such as GPT (OpenAI, 2023), OPT (Zhang et al., 2022), Llama (Touvron et al., 2023; Liu et al., 2023b), BLOOM (Scao et al., 2022). These models demonstrated remarkable proficiency in understanding and generating human language, offering the potential for a wide range of applications in fields such as natural language understanding, text generation, and conversational AI. Instruction tuning (IT) is proposed to align the LLM to follow instructions response (Liu et al., 2023a; Zhang et al., 2023; Shen et al., 2023; Wang et al., 2023a) and bridge the gap between the next-word prediction objective and the downstream tasks.

8. Conclusion

In this work, we introduce M^{3}P, a state-of-the-art multilingual multimodal machine translation model, which supports multiple translation directions of 102 languages guided by image context. To narrow the gap among different languages, the image is operated as the central language by contrastive learning (MMCL) trained on the multilingual text-image pairs. Then, we incorporate the visual context into the language representations as the conditional vision-language memory (CVLM) for multilingual generation. Extensive experiments prove the effectiveness of M^{3}P on the Multi30k and the extended large-scale dataset `InstrMulti102` of 102 languages. The importance of visual signals in multilingual training has been further verified by a series of probing experiments.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (Grant Nos. U1636211, U2333205, 61672081, 62302025, 62276017), a fund project: State Grid Co., Ltd. Technology R&D Project (ProjectName: Research on Key Technologies of Data Scenario-based Security Governance and Emergency Blocking in Power Monitoring System, Proiect No.: 5108-202303439A-3-2-ZN), the 2022 CCF-NSFOCUS Kun-Peng Scientific Research Fund and the Opening Project of Shanghai Trusted Industrial Control Platform and the State Key Laboratory of Complex & Critical Software Environment (Grant No. SKLSDE-2021ZX-18).

9. Bibliographical References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George F. Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint arXiv:1907.05019*, abs/1907.05019.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*, abs/2309.16609.

Ankur Bapna, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi Baljekar, Xavier Garcia, Wolfgang Macherey, Theresa Breiner, Vera Axelrod, Jason Riesa, Yuan Cao, Mia Xu Chen, Klaus Macherey, Maxim Krikun, Pidong Wang, Alexander Gutkin, Apurva Shah, Yanping Huang, Zhifeng Chen, Yonghui Wu, and Macduff Hughes. 2022. Building machine translation systems for the next thousand languages. *arXiv preprint arXiv:2205.03983*, abs/2205.03983.

Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. 2018. Findings of the third shared task on multimodal machine translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 304–323, Belgium, Brussels. Association for Computational Linguistics.

Ozan Caglayan, Adrien Bardet, Fethi Bougares, Loïc Barrault, Kai Wang, Marc Masana, Luis Herranz, and Joost van de Weijer. 2018. LIUM-CVC submissions for WMT18 multimodal translation task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 597–602, Belgium, Brussels. Association for Computational Linguistics.

Linzheng Chai, Jian Yang, Tao Sun, Hongcheng Guo, Jiaheng Liu, Bing Wang, Xiannian Liang, Jiaqi Bai, Tongliang Li, Qiyao Peng, and Zhoujun Li. 2024. xcot: Cross-lingual instruction tuning for cross-lingual chain-of-thought reasoning. *arXiv preprint arXiv:2401.07037*.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: universal image-text representation learning. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX*, volume 12375 of *Lecture Notes in Computer Science*, pages 104–120. Springer.

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

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR*

2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net.

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

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22:107:1–107:48.

Qingkai Fang and Yang Feng. 2022. Neural machine translation with phrase-level universal visual representations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5687–5698, Dublin, Ireland. Association for Computational Linguistics.

Hongyu Gong, Xian Li, and Dmitriy Genzel. 2021. Adaptive sparse transformer for multilingual translation. *arXiv preprint arXiv:2104.07358*, abs/2104.07358.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2021. The FLORES-101 evaluation benchmark for low-resource and multilingual machine translation. *arXiv preprint arXiv:2106.03193*, abs/2106.03193.

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor O.K. Li. 2018. Universal neural machine translation for extremely low resource languages. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 344–354, New Orleans, Louisiana. Association for Computational Linguistics.

Hongcheng Guo, Jiaheng Liu, Haoyang Huang, Jian Yang, Zhoujun Li, Dongdong Zhang, and Zheng Cui. 2022a. LVP-M3: Language-aware visual prompt for multilingual multimodal machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2862–2872, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hongcheng Guo, Jiaheng Liu, Haoyang Huang, Jian Yang, Zhoujun Li, Dongdong Zhang, and Furu Wei. 2022b. LVP-M3: language-aware visual prompt for multilingual multimodal machine translation. *arXiv preprint arXiv:2210.15461*, abs/2210.15461.

Hongcheng Guo, Boyang Wang, Jiaqi Bai, Jiaheng Liu, Jian Yang, and Zhoujun Li. 2023. M2C: towards automatic multimodal manga complement. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 9876–9882. Association for Computational Linguistics.

Po-Yao Huang, Mandela Patrick, Junjie Hu, Graham Neubig, Florian Metze, and Alexander Hauptmann. 2021. Multilingual multimodal pre-training for zero-shot cross-lingual transfer of vision-language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2443–2459, Online. Association for Computational Linguistics.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. Layoutlmv3: Pre-training for document AI with unified text and image masking. In *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, pages 4083–4091. ACM.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Bei Li, Chuanhao Lv, Zefan Zhou, Tao Zhou, Tong Xiao, Anxiang Ma, and JingBo Zhu. 2022. On vision features in multimodal machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6327–6337, Dublin, Ireland. Association for Computational Linguistics.

Jiaoda Li, Duygu Ataman, and Rico Sennrich. 2021a. Vision matters when it should: Sanity checking multimodal machine translation models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8556–8562, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

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

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 *Computer Vision – ECCV 2014*, pages 740–755, Cham. Springer International Publishing.

Jiaheng Liu, Zhiqi Bai, Yuanxing Zhang, Chenchen Zhang, Yu Zhang, Ge Zhang, Jiakai Wang, Haoran Que, Yukang Chen, Wenbo Su, Tiezheng Ge, Jie Fu, Wenhu Chen, and Bo Zheng. 2024. E2-llm: Efficient and extreme length extension of large language models. *arXiv preprint arXiv:2401.06951*.

Jiaheng Liu, Tan Yu, Hanyu Peng, Mingming Sun, and Ping Li. 2022. Cross-lingual cross-modal consolidation for effective multilingual video corpus moment retrieval. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1854–1862, Seattle, United States. Association for Computational Linguistics.

Xuejing Liu, Wei Tang, Jinghui Lu, Rui Zhao, Zhaojun Guo, and Fei Tan. 2023a. Deeply coupled cross-modal prompt learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7957–7970, Toronto, Canada. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Zhengliang Liu, Yiwei Li, Peng Shu, Aoxiao Zhong, Longtao Yang, Chao Ju, Zihao Wu, Chong Ma, Jie Luo, Cheng Chen, Sekeun Kim, Jiang Hu, Haixing Dai, Lin Zhao, Dajiang Zhu, Jun Liu, Wei Liu, Dinggang Shen, Tianming Liu, Quanzheng Li, and Xiang Li. 2023b. Radiology-llama2: Best-in-class large language model for radiology. *arXiv preprint arXiv:2309.06419*, abs/2309.06419.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13–23.

Shuming Ma, Jian Yang, Haoyang Huang, Zewen Chi, Li Dong, Dongdong Zhang, Hany Hassan Awadalla, Alexandre Muzio, Akiko Eriguchi, Saksham Singhal, Xia Song, Arul Menezes, and Furu Wei. 2020. XLM-T: scaling up multilingual machine translation with pretrained cross-lingual transformer encoders. *arXiv preprint arXiv:2012.15547*, abs/2012.15547.

OpenAI. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, abs/2303.08774.

Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. 2021. Contrastive learning for many-to-many multilingual neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 244–258, Online. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, abs/2211.05100.

Timo Schick and Hinrich Schütze. 2021. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online. Association for Computational Linguistics.

Sheng Shen, Le Hou, Yanqi Zhou, Nan Du, Shayne Longpre, Jason Wei, Hyung Won Chung, Barret Zoph, William Fedus, Xinyun Chen, Tu Vu, Yuexin Wu, Wuyang Chen, Albert Webson, Yunxuan Li, Vincent Zhao, Hongkun Yu, Kurt Keutzer, Trevor Darrell, and Denny Zhou. 2023. Flan-moe: Scaling instruction-finetuned language models with sparse mixture of experts. *arXiv preprint arXiv:2305.14705*, abs/2305.14705.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. VL-BERT: pre-training of generic visual-linguistic representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.

Mingkang Tang, Zhanyu Wang, Zhenhua LIU, Fengyun Rao, Dian Li, and Xiu Li. 2021. Clip4caption: Clip for video caption. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, page 4858–4862, New York, NY, USA. Association for Computing Machinery.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, abs/2302.13971.

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

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022a. OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 23318–23340. PMLR.

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

Wenhui Wang, Hangbo Bao, Li Dong, and Furu Wei. 2021. Vlmo: Unified vision-language pretraining with mixture-of-modality-experts. *arXiv preprint arXiv:2111.02358*, abs/2111.02358.

Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. 2019. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 4580–4590. IEEE.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023a. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhu Chen, Jie Fu, and Junran Peng.

2023b. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *arXiv preprint arXiv: 2310.00746*.

Zixiang Wang, Linzheng Chai, Jian Yang, Jiaqi Bai, Yuwei Yin, Jiaheng Liu, Hongcheng Guo, Tongliang Li, Liqun Yang, Hebboul Zine El Abidine, and Zhoujun Li. 2023c. Mt4crossoie: Multi-stage tuning for cross-lingual open information extraction. *arXiv preprint arXiv:2308.06552*, abs/2308.06552.

Genta Indra Winata, Samuel Cahyawijaya, Zihan Liu, Zhaojiang Lin, Andrea Madotto, and Pascale Fung. 2021. Are multilingual models effective in code-switching? In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 142–153, Online. Association for Computational Linguistics.

Zhiyong Wu, Lingpeng Kong, Wei Bi, Xiang Li, and Ben Kao. 2021. Good for misconceived reasons: An empirical revisiting on the need for visual context in multimodal machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6153–6166, Online. Association for Computational Linguistics.

Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. 2021. VideoCLIP: Contrastive pre-training for zero-shot video-text understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6787–6800, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jian Yang, Shuming Ma, Li Dong, Shaohan Huang, Haoyang Huang, Yuwei Yin, Dongdong Zhang, Liqun Yang, Furu Wei, and Zhoujun Li. 2023. Ganlm: Encoder-decoder pre-training with an auxiliary discriminator. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9394–9412, Toronto, Canada. Association for Computational Linguistics.

Jian Yang, Shuming Ma, Haoyang Huang, Dongdong Zhang, Li Dong, Shaohan Huang, Alexandre Muzio, Saksham Singhal, Hany Hassan, Xia Song, and Furu Wei. 2021a. Multilingual machine translation systems from Microsoft for WMT21 shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 446–455, Online. Association for Computational Linguistics.

Jian Yang, Shuming Ma, Dongdong Zhang, Shuangzhi Wu, Zhoujun Li, and Ming Zhou. 2020. Alternating language modeling for cross-lingual pre-training. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9386–9393. AAAI Press.

Jian Yang, Yuwei Yin, Shuming Ma, Haoyang Huang, Dongdong Zhang, Zhoujun Li, and Furu Wei. 2021b. Multilingual agreement for multilingual neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 233–239, Online. Association for Computational Linguistics.

Jian Yang, Yuwei Yin, Shuming Ma, Dongdong Zhang, Zhoujun Li, and Furu Wei. 2022a. High-resource language-specific training for multilingual neural machine translation. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4461–4467. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Jian Yang, Yuwei Yin, Shuming Ma, Dongdong Zhang, Shuangzhi Wu, Hongcheng Guo, Zhoujun Li, and Furu Wei. 2022b. Um4: Unified multilingual multiple teacher-student model for zero-resource neural machine translation. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4454–4460. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Shaowei Yao and Xiaojun Wan. 2020. Multimodal transformer for multimodal machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4346–4350, Online. Association for Computational Linguistics.

Kong Yawei and Kai Fan. 2021. Probing multimodal machine translation with pre-trained language model. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3689–3699, Online. Association for Computational Linguistics.

Yongjing Yin, Fandong Meng, Jinsong Su, Chulun Zhou, Zhengyuan Yang, Jie Zhou, and Jiebo Luo. 2020. A novel graph-based multi-modal fusion encoder for neural machine translation. In *Proceedings of the 58th Annual Meeting of the*

Association for Computational Linguistics, pages 3025–3035, Online. Association for Computational Linguistics.

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

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020a. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. 2023. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*, abs/2308.10792.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, abs/2205.01068.

Zhuosheng Zhang, Kehai Chen, Rui Wang, Masao Utiyama, Eiichiro Sumita, Zuchao Li, and Hai Zhao. 2020b. Neural machine translation with universal visual representation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, abs/2304.10592.