LVP-M^3: Language-aware Visual Prompt for Multilingual Multimodal Machine Translation

Hongcheng Guo*[1], Jiaheng Liu*[1], Haoyang Huang[2], Jian Yang[1], Zhoujun Li✉[1], Dongdong Zhang[2] , Zheng Cui[2]
[1]Beihang University
[2]Microsoft Research Asia
{hongchengguo,liujiaheng,jiaya,lizj}@buaa.edu.cn
{haohua,dozhang,zhcui}@microsoft.com

Abstract

Multimodal Machine Translation (MMT) focuses on enhancing text-only translation with visual features, which has attracted considerable attention from both natural language processing and computer vision communities. Recent advances still struggle to train a separate model for each language pair, which is costly and unaffordable when the number of languages increases in the real world. In other words, the multilingual multimodal machine translation (**Multilingual MMT**) task has not been investigated, which aims to handle the aforementioned issues by providing a shared semantic space for multiple languages. Besides, the image modality has no language boundaries, which is superior to bridging the semantic gap between languages. To this end, we first propose the Multilingual MMT task by establishing two new Multilingual MMT benchmark datasets covering seven languages. Then, an effective baseline LVP-M^3 using visual prompts is proposed to support translations between different languages, which includes three stages (token encoding, language-aware visual prompt generation, and language translation). Extensive experimental results on our constructed benchmark datasets demonstrate the effectiveness of LVP-M^3 method for Multilingual MMT.

1 Introduction

Multimodal Machine Translation (**MMT**) extends the conventional text-based machine translation by taking corresponding images as additional inputs (Lin et al., 2020; Li et al., 2022) to mitigate the problems of data sparsity and ambiguity (Ive et al., 2019; Yang et al., 2022) when compared with purely text-based machine translation. Similar to other multimodal tasks (e.g., visual question answering (Antol et al., 2015; Shih et al., 2016), image captioning (Vinyals et al., 2015; Jia et al., 2015) and video-text retrieval (Liu et al., 2022d)), MMT aims to exploit the effectiveness of vision information for the machine translation task.

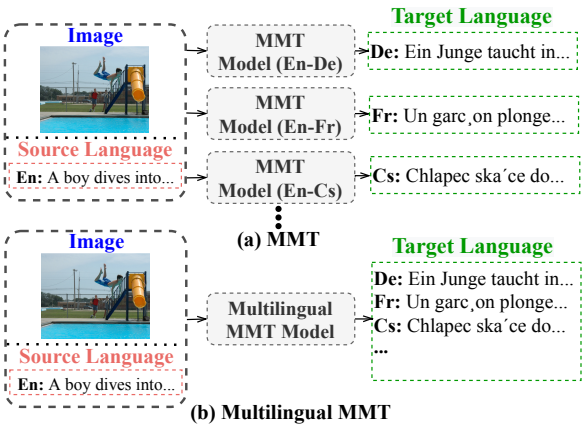

Figure 1: Comparison of MMT and Multilingual MMT. (a) For MMT, we need to train different MMT models to support translations between different language pairs (e.g., "En-De" represents to translate the English to German). (b). For Multilingual MMT, we only need one single model to translate the source language to different target languages.

Moreover, MMT has broad applications (Zhou et al., 2018), such as multimedia news and movie subtitles in different languages.

However, as shown in Fig. 1(a), previous MMT models (e.g., DCCN (Lin et al., 2020)) can handle a single language translation pair (e.g., English → German, English → French) well, but training a separate model for each language pair is unaffordable considering there are thousands of languages in the world. A straightforward solution to reduce computational cost is to use one model for handling the translations of multiple languages as shown in Fig. 1(b). Meanwhile, multilingual machine translation has been investigated for many years (Conneau et al., 2020), but these existing methods only consider the language as the input, where the vision context has been ignored. Therefore, in our work, we first propose the Multilingual Multimodal

* First two authors contributed equally.
✉ Corresponding author.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 2862–2872
December 7-11, 2022

Machine Translation (**Multilingual MMT**) task to achieve the translations for multiple languages using one single model.

To eliminate the above limitations, we propose a simple and effective **LVP-M^3** method, including Token Encoding, Language-aware Visual Prompt Generation (LVPG), and Language Translation. Specifically, in the token encoding stage, we use the pre-trained vision encoder to extract the visual tokens. Then, we follow (Johnson et al., 2017) to utilize the Transformer to encode the textual tokens. In LVPG, inspired by (Yang et al., 2019) and (Tian et al., 2020), a controller network in Fig. 3 is leveraged to dynamically generate the parameters of the mapping network conditioned on the target language. Further, the mapping network outputs the language-aware visual prompts. After that, during the language translation, following the works (e.g., ViLBERT (Lu et al., 2019)), we utilize co-Transformer to generate the vision-guided language tokens. Then the Transformer decoder is adopted to predict the translation results.

Extensive experiments are conducted on our proposed benchmark datasets for LVP-M^3. Results show that our model achieves the state-of-the-art performance in all translation directions, especially outperforming the text-only multilingual model by 4.3 BLEU scores on average.

The contributions of this work are summarized as follows:

- We first propose the Multilingual Multimodal Machine Translation (Multilingual MMT) to handle the translations for multiple language pairs, which investigates the effect of vision modality for multilingual translation and reduces the computation costs of existing MMT methods for multiple languages.
- For Multilingual MMT, we propose an effective language-aware visual prompt generation strategy to produce different visual prompts for different target languages based on the vision modality and type of the target language.
- We establish two Multilingual MMT benchmark datasets to nourish the further research on Multilingual MMT, and extensive experiments on these datasets demonstrate the effectiveness of our proposed LVP-M^3 method.

2 Related Works

Multimodal Machine Translation. The multimodal context plays a key role in Multimodal Machine Translation (MMT). Recent MMT methods can be divided into three categories: (1) Using global visual features directly (Calixto and Liu, 2017). For instance, Huang et al. (2016) proposes to concatenate global and regional visual features with source sequences. (2) Exploiting visual features via attention scheme (Libovickỳ and Helcl, 2017; Helcl et al., 2018). Calixto et al. (2017) introduces the visual features into the MMT model by using an independent attention module. (3) Combining other vision tasks with the translation task by multitask learning (Calixto et al., 2019; Yin et al., 2020). Elliott and Kádár (2017) decomposes multimodal translation into two sub-tasks (i.e., translation and visual grounding). Recently, (Huang et al., 2020) focuses on unsupervised setting for MMT, which utilizes pseudo visual pivoting and visual content to improve the cross-lingual alignments in latent space. In contrast, LVP-M^3 considers fully-supervised multilingual setting by mapping vision embeddings into different feature spaces and achieving the purpose of using one MT model for handling translations of multiple languages. Besides, reducing computation cost is vital for many tasks (Liu et al., 2021, 2022c,a) and we focus on the Multilingual MMT task by using one single model for efficiency.

Multilingual Language Models. Pre-trained multilingual Transformer-based language models (e.g., mBERT (Kenton and Toutanova, 2019) and XLM-R (Conneau et al., 2020)) utilize the same pre-training strategies as their respective monolingual counterparts (e.g., BERT (Kenton and Toutanova, 2019) and RoBERTa (Liu et al., 2019)). They are pre-trained via the masked language modeling objective (MLM) Strategy. Artetxe et al. (2020) proposes a method to transfer monolingual representations to new languages in an unsupervised fashion and provide new insights into the generalization abilities of multilingual models. Hu et al. (2020) introduces the Cross-lingual Transfer Evaluation of Multilingual Encoders (XTREME) benchmark to evaluate the cross-lingual generalization capabilities, Karthikeyan et al. (2020) also provides a comprehensive study of the contribution of different components in M-BERT to its cross-lingual ability. Rust et al. (2021) shows that monolingually adapted tokenizers can robustly improve the mono-

lingual performance of multilingual models. Overall, when compared with these methods, we focus on the multilingual setting for MMT, which has not been investigated before.

Vision-Language Models. The success of vision-language models can be credited to the following three important reasons: Transformers (Liu et al., 2022b; Vaswani et al., 2017), contrastive representation learning (Radford et al., 2021; Li et al., 2020), and large-scale training datasets (Sharma et al., 2018; Miech et al., 2019). Previous Transformer-based multimodal models ((Tan and Bansal, 2019; Chen et al., 2020; Gan et al., 2020; Bugliarello et al., 2021)) jointly encode text tokens and image region features by preprocessing images using object detection models. The image region features are projected into the joint embedding space of the multimodal Transformer, and then the multi-head attention attends to all text and image inputs to learn a joint representation of both modalities. Besides, Kamath et al. (2021) avoids using object detectors as a black box for pre-extracting these region features and incorporates the object detector end-to-end with the multimodal Transformer to achieve flexibility and better representation capacity. Recently, a representative approach CLIP (Radford et al., 2021) is proposed, which trains two neural network-based encoders using a contrastive loss to match pairs of images and texts. After consuming 400 million data pairs, the CLIP model demonstrates a remarkable zero-shot image recognition capability, and has been applied to many downstream tasks. For example, Shen et al. (2022) proposes to leverage the CLIP model for different vision-language models across various tasks (e.g., image caption, visual question answering). In our work, we aim to investigate the effectiveness of the multimodal information for Multilingual MMT.

3 Datasets

We introduce two Multilingual MMT benchmark datasets (i.e., M^3-Multi30K, M^3-AmbigCaps) using Multi30K (Elliott et al., 2016) and AmbigCaps (Li et al., 2021). Here, we descried the details of the M^3-Multi30K and M^3-AmbigCaps.

Data Construction. The widely-used Multi30K dataset for the MMT task is based on the Flickr30K Entities dataset (Plummer et al., 2017). For each image of Multi30K, one of the English (En) descriptions is selected in Flickr30K Entities. Currently,

En: A child is splashing in the water
De: Ein Kind plantscht im Wasser
Cs: Dítě se šplouchá ve vodě
Tr: Bir çocuk suya sıçratıyo
Hi: एक बच्चा पानी में छींटे मार रहा है
Lv: Bērns plunčājas ūdenī
Fr: Un enfant éclabousse dans l'eau

Figure 2: Example of an image with its descriptions of seven different languages.

Language	ISO	Family	Speakers
English	En	Germanic	400M
German	De	Germanic	95M
French	Fr	Romance	250M
Czech	Cs	Slavic	11M
Hindi	Hi	Indo-Aryan	800M
Turkish	Tr	Turkic	65M
Latvian	Lv	Baltic	2M

Table 1: Languages covered by our proposed M^3-Multi30K and M^3-AmbigCaps datasets.

the English description of each image is translated into German (De), French (Fr), and Czech (Cs) (Elliott et al., 2017; Barrault et al., 2018). To support more languages from different language families and various language distributions for Multilingual MMT, we extend the existing Multi30K dataset with additional three languages as shown in Table 1, where one sample of the M^3-Multi30K dataset is shown in Fig. 2.

Specifically, in the annotation process, based on the recent state-of-the-art multilingual machine translation model XLM-R (Conneau et al., 2020), we first translate the English description into Hindi (Hi), Turkish (Tr), and Latvian (Lv) for each image in Multi30K. Then, we hire independent native speakers to verify and improve the quality of the translation results of different languages. In addi-

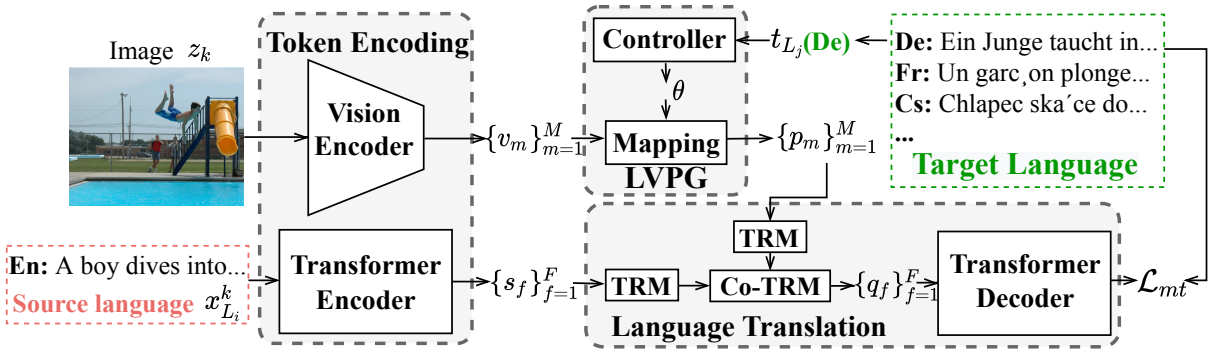

Figure 3: The overall framework of our proposed LVP-M^3 method for Multilingual MMT task, which includes three stages (i.e., token encoding and language-ware visual prompt generation (LVPG) and language translation). Here, we take an example by translating English (En) to German (De). "TRM" and "Co-TRM" represent the Transformer and co-Transformer models, respectively.

tion, as the original AmbigCaps (Li et al., 2021) dataset only contains two types of languages, we use a similar way to extend AmbigCaps into additional five languages in M^3-AmbigCaps.

Data Splits. In M^3-Multi30K, the number of image-translation pairs for training and testing data are 29000, 1000, respectively. In M^3-AmbigCaps, the number of image-translation pairs for training and testing data are 89600, 1000, respectively. We will released these datasets.

4 Method

4.1 Multilingual MMT

Supposing we have M languages $\{L_m\}^M_{m=1}$ and N bilingual corpora $\{D_n\}^N_{n=1}$ under the multilingual setting, the dataset D_n consists of K parallel sentences $\{(x^k_{L_i}, x^k_{L_j})\}^K_{k=1}$ between language L_i and L_j, where K is the number of training instances and each instance has the corresponding image z_k. Given the corpora, we can train a Multilingual MMT model that enables the translation among different languages with the help of image modality. The training objective of the Multilingual MMT is learnt with a combination of different languages:

$$\mathcal{L}_{mt} = -\sum_{i,j,k} \log \mathbf{P}(x^k_{L_i}; x^k_{L_j}; z_k), \qquad (1)$$

where the Multilingual MMT model uses a complete shared model for all translation directions. In this work, we adopt Transformer as the backbone model for language encoding and pre-trained vision branch of the CLIP model (Radford et al., 2021) for vision modality. A target language token t_{L_j} is prefixed to each source sentence to indicate the translation direction (Johnson et al., 2017).

4.2 LVP-M^3

As shown in Fig. 3, our proposed LVP-M^3 method includes three stages: token encoding, language-aware visual prompt generation and language translation. Specifically, in training, give each image z_k, the parallel sentences $\{(x^k_{L_i}; x^k_{L_j})\}$ from source language L_i and target language L_j, and the target language token embedding t_{L_j}, in token encoding stage, we first use the vision encoder to extract the visual token features $\{v_m\}^M_{m=1}$ based on z_k, where M is the number of visual tokens. Then, we utilize the Transformer encoder to extract the source language tokens $\{s_f\}^F_{f=1}$, where F is the number of language tokens. In language-aware visual prompt generation (LVPG) stage, we map the $\{v_m\}^M_{m=1}$ into the language-aware visual prompt $\{p_m\}^M_{m=1}$ conditioned on t_{L_j} to generate different visual prompts for different target languages, where we propose to adopt the controller network to dynamically generate the parameters of the mapping network. In language translation stage, we first adopt the co-attention strategy to generate the vision-guided language tokens $\{q_f\}^F_{f=1}$ based on $\{p_m\}^M_{m=1}$ and $\{s_f\}^F_{f=1}$. Then, we use the $\{q_f\}^F_{f=1}$ as the input of the Transformer decoder to predict the translation results and compute the loss in Eq. 1 using the predicted translation results and the ground-truth target language $x^k_{L_j}$.

4.2.1 Token Encoding

For each image z_k, we directly use the vision backbone (e.g., the pre-trained vision branch of the widely-used CLIP model (Radford et al., 2021)) as the vision encoder to extract the visual tokens

for z_k as follows:

$$\{v_m\}_{m=1}^{M} = \mathcal{H}(z_k), \tag{2}$$

where $\mathcal{H}$ denotes the vision encoder and M is the number of visual tokens.

Similarly, given the source language $x_{L_i}^k$, based on the Transformer encoder $\mathcal{E}$, the source language tokens $\{s_f\}_{f=1}^{F}$ are extracted as follows:

$$\{s_f\}_{f=1}^{F} = \mathcal{E}(x_{L_i}^k), \tag{3}$$

where F is defined as the number of source language tokens.

4.2.2 Language-aware Visual Prompt Generation

In language-aware visual prompt generation stage of Fig. 3, motivated by recent works (e.g., dynamic filter networks (Jia et al., 2016) and CondConv (Yang et al., 2019)) based on conditional convolutions, where the filters of conditional convolutions are conditioned on the input and are dynamically generated by another network to improve the capacity of the neural network, we extend this idea to generate the visual prompt conditioned on the target language type t_{L_j} (e.g., German) to map the extracted the visual tokens into different embedding spaces for different target language. Specifically, in Fig. 3, based on the embedding of the target language token t_{L_j}, we utilize a controller network $\mathcal{C}$ implemented by two fully-connected layers with ReLU (Nair and Hinton, 2010) activation function to generate the parameters θ of the mapping network $\mathcal{M}$ as follows:

$$\theta = \mathcal{C}(t_{L_j}). \tag{4}$$

After that, we generate the language-aware visual prompt $\{p_m\}_{m=1}^{M}$ as follows:

$$\{p_m\}_{m=1}^{M} = \mathcal{M}(\{v_m\}_{m=1}^{M}, \theta). \tag{5}$$

θ is the generated parameters in Eq. 4, which is assigned to the mapping network $\mathcal{M}$. In this way, when translating source language into different target languages, the θ will be generated according to type of target language tokens, and the visual tokens $\{v_m\}_{m=1}^{M}$ can be mapped into different visual prompts according to the type of the target language.

4.2.3 Language Translation

In Fig. 3, based on the source language tokens $\{s_f\}_{f=1}^{F}$ and language-aware visual prompt $\{p_m\}_{m=1}^{M}$, we first generate the vision-guided source language tokens based on co-attention strategy, which are widely used for fusing the information from another modality in vision-language models (Lu et al., 2019). Then, we predict the translation results using the Transformer decoder.

Specifically, we utilize the Transformer module implemented by self-attention to fuse the information from other tokens within each modality for $\{s_f\}_{f=1}^{F}$ and $\{p_m\}_{m=1}^{M}$, respectively, and we represent the updated source language tokens and visual prompt as $\mathbf{S}$ and $\mathbf{P}$, respectively. Then, we take $\mathbf{S}$ as the query, and the $\mathbf{P}$ as the key and value in the co-attention module to generate the vision-guided source language tokens $\{q_f\}_{f=1}^{F}$ as follows:

$$\{q_f\}_{f=1}^{F} = \Big\Vert_{h=1}^{H} \texttt{SF}\left(\frac{\boldsymbol{\phi}_Q^h(\mathbf{S})\boldsymbol{\phi}_K^h(\mathbf{P})^\top}{\sqrt{C}}\right)\boldsymbol{\phi}_V^h(\mathbf{P}), \tag{6}$$

where $\Vert_{h=1}^{H}$ is the concatenation of the H attentive features along the channel dimension. SF represents the softmax operation. $\boldsymbol{\phi}_Q^h(\cdot), \boldsymbol{\phi}_K^h(\cdot)$ and $\boldsymbol{\phi}_V^h(\cdot)$ are the corresponding linear projection operations of the h-th head for the query, the key and the value, respectively. C denotes the number of feature channels. After the operation of Eq. 6, other operations (e.g., FFN, layer normalization (Ba et al., 2016)) of standard attention scheme (Vaswani et al., 2017) are used.

Finally, at inference, based on $\{q_f\}_{f=1}^{F}$, we use the Transformer decoder to predict the target language sequence in our LVP-M^3.

5 Experiments

We evaluate our proposed LVP-M^3 method on the multilingual dataset including 7 languages and 6 translation directions. In all experiments, English (En) is treated as the pivot language for Multilingual MMT setting.

5.1 Experimental Setting

Implementation Details. Our implementation is based on the Fairseq (Ott et al., 2019) toolbox. We utilize Sentencepiece tokenizer. The model in Fig. 3 consists of 6 Transformer encoder/decoder layers. The number of attention heads in all Transformer layers is set as 12. For training, we take the Adam optimizer (Kingma and Ba, 2015) with

Model (En→X)	Fr	Cs	De	Lv	Hi	Tr	Avg_{all}
Text-only Multilingual MT Systems							
Text Transformer (Fan et al., 2021)	61.8	32.8	40.6	51.2	59.0	53.8	49.8
Multilingual MMT Systems							
Vision Matters (Gated fusion) (Li et al., 2021)	62.5	32.9	41.2	52.1	59.6	54.2	50.4
Vision Matters (Concatenation) (Li et al., 2021)	59.7	33.1	39.8	50.3	57.6	51.4	48.6
LVP-M^3 (w/o LVPG)	62.2	33.4	40.9	51.6	59.3	54.0	50.2
LVP-M^3 (Our method)	**63.7**	**34.6**	**43.2**	**53.5**	**61.4**	**55.6**	**52.0**

Table 2: The BLEU scores of different methods on M^3-Multi30K test set. Five multilingual baselines are compared by us. The bottom part shows the results of the multilingual models trained with text and vision modalities. The best results are highlighted.

Model (En→X)	Fr	Cs	De	Lv	Hi	Tr	Avg_{all}
Text-only Multilingual MT Systems							
Text Transformer (Fan et al., 2021)	62.3	47.8	49.0	46.6	52.4	35.9	49.0
Multilingual MMT Systems							
Vision Matters (Gated fusion) (Li et al., 2021)	64.3	50.3	51.2	48.5	54.1	38.7	51.2
Vision Matters (Concatenation) (Li et al., 2021)	62.6	47.6	48.7	45.9	52.7	36.0	48.9
LVP-M^3 (w/o LVPG)	63.4	49.2	50.3	47.9	52.4	37.1	50.1
LVP-M^3 (Our method)	**65.7**	**52.9**	**53.7**	**51.6**	**56.3**	**42.7**	**53.8**

Table 3: The BLEU scores of different methods on M^3-AmbigCaps test set. Five multilingual baselines are compared by us. The bottom part shows the results of the multilingual models trained with text and vision modalities.The best results are highlighted.

$\beta_1 = 0.9$ and $\beta_2 = 0.98$. The learning rate warms up from 1e-7 to 1e-4 in 2000 steps and then decays based on the inverse square root of the update number. The maximum number of tokens in each mini-batch is 4096. Dropout and label-smoothing rate are set as 0.3 and 0.1, respectively. For vision encoder, by default, we adopt the vision branch of CLIP based on the ViT-L/14 model. The effect of different vision backbones is discussed in our ablation study. All models are trained for 30 epochs and evaluated on one single linux machine with 8 NVIDIA A100 GPUs (80G).

Evaluation. We compute the cumulative 4-gram BLEU scores to evaluate the quality of translation. During inference, the beam search strategy is performed with a beam size of 5 for the target sentence generation. We set the length penalty as 1.0.

Baseline Methods. As we are the first multilingual method in this area, we reproduce methods including Text Transformer (Fan et al., 2021), the Vision Matters (Gated fusion) (Li et al., 2021), and the Vision Matters (Concatenation) (Li et al., 2021) in the multilingual translation setting for a fair comparison. Besides, we also report the results of LVP-M^3 (w/o LVPG), where we directly adopt the co-attention strategy in Lu et al. (2019) to generate the vision-guided language tokens using the source tokens with the visual features.

5.2 Results on M^3-Multi30K

To demonstrate the effectiveness of LVP-M^3, we compare our method with baseline methods on M^3-Multi30K under the multilingual MMT setting in Table 2. It should be mentioned that the Vision Matters (Gated fusion) and the Vision Matters (Concatenation) are originally proposed in the bilingual setting, and we reproduce these methods in the multilingual setting for a fair comparison. In Table 2, our LVP-M^3 achieves the best BLEU scores in all translation directions. Specifically, first, when compared with text Transformer with only text information, LVP-M^3 outperforms by +2.2 BLEU scores on average, which demonstrates the effectiveness of visual context for Multi-

Model (En→X)	Fr	Cs	De	Lv	Hi	Tr	Avg$_{all}$
LVP-M^3 (Static)	62.0	33.1	41.1	51.7	59.6	54.2	50.3
LVP-M^3 (LVPG)	**63.7**	**34.6**	**43.2**	**53.5**	**61.4**	**55.6**	**52.0**

Table 4: Comparison of different vision prompt generation methods with BLEU scores.

Model (En→X)	Fr	Cs	De	Lv	Hi	Tr	Avg$_{all}$
LVP-M^3+ResNet50	62.3	33.3	41.7	52.3	61.1	54.0	50.8
LVP-M^3+ResNet101	62.8	33.8	42.1	52.5	60.7	54.2	51.1
LVP-M^3+ViT-L/14	**63.7**	**34.6**	**43.2**	**53.5**	**61.4**	**55.6**	**52.0**

Table 5: Comparison different visual backbones with BLEU scores.

lingual MMT. Second, when compared with baseline method LVP-M^3 (w/o LVPG), LVP-M^3 also achieves better performance on all settings, which verifies the effectiveness of our proposed language-aware prompt generation module for Multilingual MMT. Among all translation directions, the task of En→De achieves the most improvement. Because English and German are from the same language family, both languages can share the similar semantic knowledge by cross-lingual transfer.

5.3 Results on M^3-AmbigCaps

Results of M^3-AmbigCaps are presented in Table 3. When compared with other baseline methods, we observe that our proposed LVP-M^3 method also achieves significant performance improvements in all translation directions. In Table 3, we observe that our proposed method LVP-M^3 outperforms by $+4.8$ BLEU scores on average when the visual modality is used, which is larger than that in M^3-Multi30K.

5.4 Ablation Study

In this section, we conduct comprehensive ablation study to demonstrate the effectiveness of different components in our proposed LVP-M^3 method on the test set of M^3-Multi30K.

Effect of LVPG. In Table 2 and Table 3, we observe that our language-aware visual prompt generation (LVPG) brings significant improvements for Multilingual MMT. To demonstrate the effectiveness of LVPG, we further propose two alternative methods (i.e., LVP-M^3 (Static) and LVP-M^3 (CoCoOp)) to generate the visual prompts in Table 4. Specifically, in LVP-M^3 (Static), we directly generate visual prompts by mapping the visual tokens $\{v_m\}_{m=1}^{M}$ using the mapping network, where the parameters of the mapping network are static after training and not conditioned on the target language token embedding t_{L_j}. In Table 4, we observe that our LVP-M^3 outperforms these alternative methods a lot, which guides the visual clues to bridge the semantic gap between multiple languages.

Effect of Different Vision Backbones. In Table 5, we compare the results of LVP-M^3 by using the visual tokens extracted by different vision backbones (He et al., 2016; Dosovitskiy et al., 2020) in CLIP. In Table 5, we observe that our LVP-M^3 achieves best results when using ViT-L/14 as the vision encoder. Thus, we use ViT-L/14 as the vision encoder by default. Moreover, we observe that the performance is better when the capacity of the vision backbone is better. It is also reasonable because the quality of the visual tokens is better when using more powerful vision backbones.

5.5 Further Analysis

Visualization of Different Masking Ratios. As shown in Fig. 4, we compare our LVP-M^3 method with the alternative method LVP-M^3 (w/o vision) to analyze the effectiveness of visual context when using different masking ratios on the source language. Specifically, in LVP-M^3 (w/o vision), we only use the Transformer encoder to process the source language with the target language embedding and then adopt the Transformer decoder to predict the target language for multilingual MT, where the vision encoder and LVPG are both not used in LVP-M^3 (w/o vision).

In Fig. 4, we report the results of these methods by translating from English (En) to French (Fr) and Turkish (Tr). First, when the ratio of masking increases, BLEU scores drop whether the visual contents are added or not, and our LVP-M^3 still

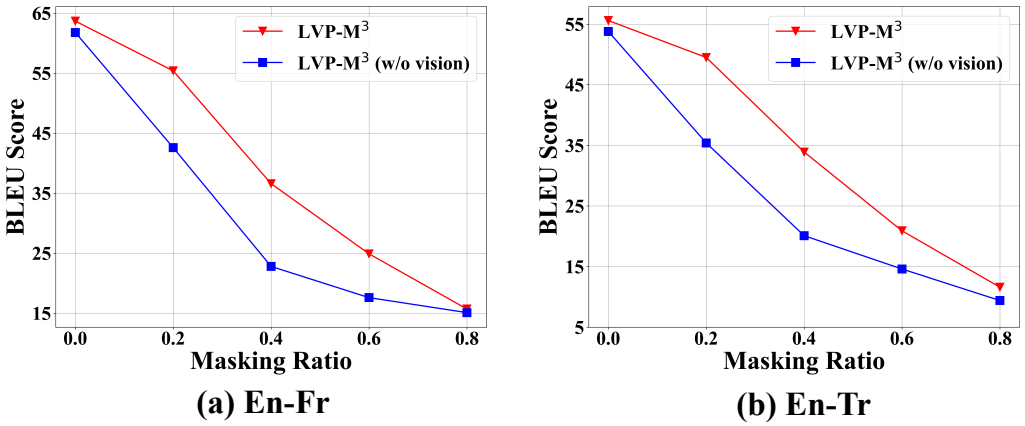

Figure 4: Translation results of LVP-M^3 under different masking ratios on the source language. Results are evaluated on the M^3-Multi30K test set by translating English (En) to other languages (i.e., Fr and Tr).

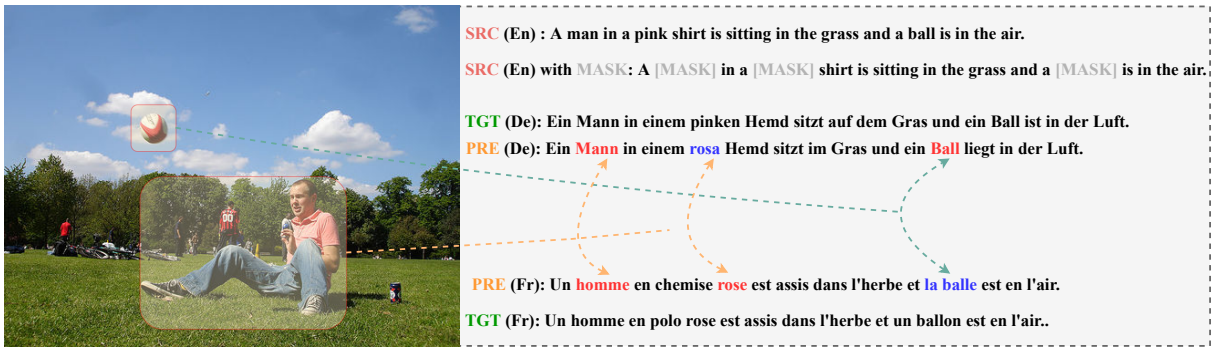

Figure 5: A qualitative example by translating English (En) to German (De) and French (Fr) with the help of vision modality. Tokens in red denotes correct translation. Tokens in blue denotes good synonyms, which have the similar meaning with the ground-truth of target language. **SRC** denotes the source language. **MASK** means the masked contents in the source language. **PRE** and **TGT** represent the predicted translation results and the ground-truth of the target language, respectively.

outperforms LVP-M^3 (w/o vision) a lot. Second, the performance gap between LVP-M^3 and LVP-M^3 (w/o vision) is larger when the mask ratio is between 20% and 40%, which shows that the visual information improves the robustness of the translation model. Third, when the mask ratio is larger, the results of these methods on all settings degrade. When the mask ratio is set as 80%, the results of LVP-M^3 (w/o vision) are close to those of LVP-M^3. It is also reasonable, because most tokens in each source language are masked and it is difficult to translate well for both methods under these extreme scenarios.

Qualitative Analysis. To further explore the necessity of visual modality for machine translation, we compare the predictions results (i.e., De and Fr) of a sample source language (i.e., En) with the ground truth of these target languages in Fig. 5. Specifically, the "man" (noun), "pink" (adjective), and "ball" (noun) are masked, and these masked tokens describe the saliency regions in the corresponding left image. We have the following observations. First, we observe that even though the "man" is masked, the prediction results of German and French on this token are still right, which means that visual modality is complementary rather than redundant if the text is insufficient. Second, our model translates some tokens to their synonyms in the target language. For example, although the translated word "rosa" in German is evaluated as a bad translation for the masked token "pink" in English, it represents the same meaning as the word "pinken" in German. Besides, "la balle" in French is also the synonym of "ball" in English, which further demonstrates the effectiveness of the vision modality.

5.6 Discussion on LVP-M^3

In our proposed LVP-M^3 method, first, both encoders (vision and text) and decoder are shared for all language pairs, while previous methods on MMT usually adopt different models for different language pairs. Second, to generate different visual prompts for different language pairs with minimal additional parameters, we just use controller network to generate the parameters of mapping network to map the vision embeddings. Third, different language translation directions are used in training, where the target language token is also prefixed to each source sentence for denoting the translation direction. Last, training separated models will result in huge training costs when compared with the multilingual models as discussed in many multilingual methods.

6 Conclusion

In our work, we first propose the Multilingual MMT task to support the multilingual multimodal machine translations between different language pairs using one single model. Then, we propose an effective LVP-M^3 baseline method for the Multilingual MMT task, where a language-aware prompt generation module is proposed to generate visual prompts for different target languages dynamically. Comprehensive experimental results on our established Multilingual MMT benchmark datasets demonstrate the effectiveness of our proposed LVP-M^3 method for Multilingual MMT.

7 Limitations

Although our proposed LVP-M^3 method has achieved substantial improvements for Multilingual MMT, we find that there still exists some hyper-parameters (e.g., the number of encoder and decoder layers,) to tune for better results, which may be time-consuming. Besides, in our established datasets, we support seven languages currently, and we will extend to support more languages and more translation directions for Multilingual MMT in the future work.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grant Nos. 62276017, U1636211, 61672081), the 2022 Tencent Big Travel Rhino-Bird Special Research Program, and the Fund of the State Key Laboratory of Software Development Environment (Grant No. SKLSDE-2021ZX-18).

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *ICCV 2015*, pages 2425–2433.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *ACL 2020*, pages 4623–4637.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

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 *WMT 2018*, pages 304–323.

Emanuele Bugliarello, Ryan Cotterell, Naoaki Okazaki, and Desmond Elliott. 2021. Multimodal pretraining unmasked: A meta-analysis and a unified framework of vision-and-language berts. *Transactions of the Association for Computational Linguistics*, 9:978–994.

Iacer Calixto and Qun Liu. 2017. Incorporating global visual features into attention-based neural machine translation. In *EMNLP 2017*, pages 992–1003.

Iacer Calixto, Qun Liu, and Nick Campbell. 2017. Doubly-attentive decoder for multi-modal neural machine translation. In *ACL 2017*, pages 1913–1924.

Iacer Calixto, Miguel Rios, and Wilker Aziz. 2019. Latent variable model for multi-modal translation. In *ACL 2019*, pages 6392–6405.

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 *ECCV 2020*, 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 *ACL 2020*.

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. In *ICLR 2020*.

Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. 2017. Findings of the second shared task on multimodal machine translation and multilingual image description. In *WMT 2017*, pages 215–233.

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.

Desmond Elliott and Ákos Kádár. 2017. Imagination improves multimodal translation. In *IJCNLP 2017*, pages 130–141.

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. *J. Mach. Learn. Res.*, 22:107:1–107:48.

Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. 2020. Large-scale adversarial training for vision-and-language representation learning. *Advances in Neural Information Processing Systems*, 33:6616–6628.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR 2016*, pages 770–778.

Jindřich Helcl, Jindřich Libovickỳ, and Dusan Varis. 2018. Cuni system for the wmt18 multimodal translation task. In *WMT 2018*, pages 616–623.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *ICML 2020*, pages 4411–4421. PMLR.

Po-Yao Huang, Junjie Hu, Xiaojun Chang, and Alexander Hauptmann. 2020. Unsupervised multimodal neural machine translation with pseudo visual pivoting. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8226–8237, Online. Association for Computational Linguistics.

Po-Yao Huang, Frederick Liu, Sz-Rung Shiang, Jean Oh, and Chris Dyer. 2016. Attention-based multimodal neural machine translation. In *WMT 2016*, pages 639–645.

Julia Ive, Pranava Madhyastha, and Lucia Specia. 2019. Distilling translations with visual awareness. In *ACL 2019*, pages 6525–6538.

Xu Jia, Bert De Brabandere, Tinne Tuytelaars, and Luc V Gool. 2016. Dynamic filter networks. *Advances in neural information processing systems*, 29.

Xu Jia, Efstratios Gavves, Basura Fernando, and Tinne Tuytelaars. 2015. Guiding the long-short term memory model for image caption generation. In *ICCV 2015*, pages 2407–2415.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *ACL 2017*, 5:339–351.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. 2021. Mdetr-modulated detection for end-to-end multi-modal understanding. In *ICCV 2021*, pages 1780–1790.

K Karthikeyan, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual bert: An empirical study. In *ICLR 2020*.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT 2019*, pages 4171–4186.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR 2015*.

Bei Li, Chuanhao Lv, Zefan Zhou, Tao Zhou, Tong Xiao, Anxiang Ma, and JingBo Zhu. 2022. On vision features in multimodal machine translation. In *ACL 2022*, pages 6327–6337.

Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. 2020. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *AAAI 2020*, volume 34, pages 11336–11344.

Jiaoda Li, Duygu Ataman, and Rico Sennrich. 2021. Vision matters when it should: Sanity checking multimodal machine translation models. In *EMNLP 2021*, pages 8556–8562.

Jindřich Libovickỳ and Jindřich Helcl. 2017. Attention strategies for multi-source sequence-to-sequence learning. In *ACL 2017*, pages 196–202.

Huan Lin, Fandong Meng, Jinsong Su, Yongjing Yin, Zhengyuan Yang, Yubin Ge, Jie Zhou, and Jiebo Luo. 2020. Dynamic context-guided capsule network for multimodal machine translation. In *ACM MM 2020*, pages 1320–1329.

Jiaheng Liu, Jinyang Guo, and Dong Xu. 2022a. Apsnet: Toward adaptive point sampling for efficient 3d action recognition. *IEEE Transactions on Image Processing*, 31:5287–5302.

Jiaheng Liu, Jinyang Guo, and Dong Xu. 2022b. Geometrymotion-transformer: An end-to-end framework for 3d action recognition. *IEEE Transactions on Multimedia*, pages 1–13.

Jiaheng Liu, Haoyu Qin, Yichao Wu, Jinyang Guo, Ding Liang, and Ke Xu. 2022c. Coupleface: Relation matters for face recognition distillation. In *Proceedings of the European Conference on Computer Vision*.

Jiaheng Liu, Tan Yu, Hanyu Peng, Mingming Sun, and Ping Li. 2022d. Cross-lingual cross-modal consolidation for effective multilingual video corpus moment retrieval. In *NAACL 2022*, pages 1854–1862.

Jiaheng Liu, Shunfeng Zhou, Yichao Wu, Ken Chen, Wanli Ouyang, and Dong Xu. 2021. Block proposal neural architecture search. *IEEE Transactions on Image Processing*, 30:15–25.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. *ICCV 2019*, pages 2630–2640.

Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *ICML 2010*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *NAACL-HLT 2019*, pages 48–53.

Bryan A. Plummer, Liwei Wang, Christopher M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2017. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *IJCV 2017*, 123:74–93.

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 *ICML 2021*, pages 8748–8763.

Phillip Rust, Jonas Pfeiffer, Ivan Vulic, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. In *ACL 2021*.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL 2018*, pages 2556–2565.

Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. 2022. How much can CLIP benefit vision-and-language tasks? In *ICLR 2022*.

Kevin J Shih, Saurabh Singh, and Derek Hoiem. 2016. Where to look: Focus regions for visual question answering. In *CVPR 2016*, pages 4613–4621.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. In *EMNLP-IJCNLP 2019*, pages 5100–5111.

Zhi Tian, Chunhua Shen, and Hao Chen. 2020. Conditional convolutions for instance segmentation. In *ECCV 2020*, pages 282–298. Springer.

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 *NIPS 2017*.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *CVPR 2015*, pages 3156–3164.

Brandon Yang, Gabriel Bender, Quoc V Le, and Jiquan Ngiam. 2019. Condconv: Conditionally parameterized convolutions for efficient inference. *Advances in Neural Information Processing Systems*, 32.

Jian Yang, Yuwei Yin, Shuming Ma, Dongdong Zhang, Shuangzhi Wu, Hongcheng Guo, Zhoujun Li, and Furu Wei. 2022. UM4: unified multilingual multiple teacher-student model for zero-resource neural machine translation. In *IJCAI 2022*, pages 4454–4460.

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 *ACL 2020*, pages 3025–3035.

Mingyang Zhou, Runxiang Cheng, Yong Jae Lee, and Zhou Yu. 2018. A visual attention grounding neural model for multimodal machine translation. In *EMNLP 2018*, pages 3643–3653.