

Visual Pivoting Unsupervised Multimodal Machine Translation in Low-Resource Distant Language Pairs

Turghun Tayir¹, Lin Li¹, Xiaohui Tao², Mieradilijiang Maimaiti³
Ming Li¹, Jianquan Liu⁴

¹Wuhan University of Technology, China, ²University of Southern Queensland, Australia

³Chinese Academy of Sciences, China, ⁴NEC Corporation, Japan

¹{hotpes, cathylilin, liming7677}@whut.edu.cn, ²xiaohui.tao@unisq.edu.au
³miradel_51@hotmail.com, ⁴jqliu@nec.com

Abstract

Unsupervised multimodal machine translation (UMMT) aims to leverage vision information as a pivot between two languages to achieve better performance on low-resource language pairs. However, there is presently a challenge: how to handle alignment between distant language pairs (DLPs) in UMMT. To this end, this paper proposes a visual pivoting UMMT method for DLPs. Specifically, we first construct a dataset containing two DLPs, including English-Uyghur and Chinese-Uyghur. We then apply the visual pivoting method for both to pre-training and fine-tuning, and we observe that the images on the encoder and decoder of UMMT have noticeable effects on DLPs. Finally, we introduce informative multi-granularity image features to facilitate further alignment of the latent space between the two languages. Experimental results show that the proposed method significantly outperforms several baselines on DLPs and close language pairs (CLPs). Our dataset Multi30k-Distant and code are available at: <https://github.com/WUT-IDEA/VP-UMMT>.

1 Introduction

Neural machine translation (MT) (Sutskever et al., 2014; Cho et al., 2014; Vaswani et al., 2017) has become a promising method for MT, which depends on the availability of large-scale parallel corpora. However, the preparation of such corpora in the low-resource language is extremely challenging, and existing studies (Zoph et al., 2016) have shown that neural MT achieves much worse translation quality than statistical MT with a small number of corpora. Therefore, developing methods to alleviate the need for annotation of large parallel corpora has attracted increasing attention from researchers.

To alleviate this problem, unsupervised MT (Lample et al., 2018; Artetxe et al., 2018) has been proposed, which relies on monolingual corpora and trains MT model in an unsupervised

manner. Since the alignment of the source-target sentence of the unsupervised MT is uncertain, it is highly subject to initialization. Therefore, researches (Su et al., 2019; Huang et al., 2020; Li et al., 2023b) have found that exploiting visual content for unsupervised MT while leveraging a language model pre-trained on large-scale monolingual data is a feasible way to improve translation quality. Visual content is qualified to improve alignment in the latent space of language because the physical visual perception of people who speak different languages is similar. However, previous works (Su et al., 2019; Huang et al., 2020; Li et al., 2023b,a; Huang et al., 2021) mainly consider high-resource CLPs, such as English-German and English-France. Because unsupervised MT aims to achieve high-quality translation results with low-resource language. The study of unsupervised MT solely on high-resource CLPs makes it challenging to assess its effectiveness in low-resource languages, diminishing its applicability and hindering the advancement of UMMT efficiency. In DLPs, even initialization with a monolingual pre-trained model does not yield significant improvements, as unsupervised MT performs well when the monolingual data in both languages belong to the same language family (Marchisio et al., 2020). Therefore, the UMMT task needs to be extended to translation between low-resource DLPs, which is beneficial for a more comprehensive exploration of the influence of linguistic distance and the contribution of visual content.

To address these challenges, we propose a visual pivoting UMMT method for DLPs. Specifically, we first manually translate the mainstream multimodal MT dataset Multi30k (Elliott et al., 2016), which primarily contains high-resource CLPs, into Chinese and Uyghur. Both Chinese and Uyghur belong to different language families from the languages of the Multi30k. Even their scripts and



Figure 1: Simple examples of CLPs from Multi30k and DLPs from our dataset. It consists of an image and its descriptions in four languages, English (En), German (De), Uyghur (Uy), and Chinese (Zh). Words with the same color have the same meaning in different language.

grammar structures are different, as shown in Figure 1; for example, Uyghur is a subject-object-verb language, while English and Chinese are subject-verb-object languages. Moreover, Uyghur is a low-resource language, hence the main data studied in this paper are the low-resource DLPs composed of English-Uyghur and Chinese-Uyghur sentences. We then extend MLM (Conneau and Lample, 2019) by leveraging visual information to generate a visual pre-training language modeling (VPLM) model, which is subsequently applied to initialize a UMMT model. Finally, we use images as a pivot to semantically align source-target languages into a shared latent space. Specifically, the image is introduced into the encoder to correct the pseudo-sentence, while the input image in the decoder is treated as a pivot between the source and target languages. We conducted experiments on DLPs and CLPs and the results show that the proposed method consistently outperforms several baselines.

Overall, we make the following contributions: (i) We construct a dataset with DLPs and the UMMT is implemented on this dataset. It provides a benchmark for further research on this challenging task. (ii) We find that visual content is more qualified to improve the alignment of DLPs latent space. (iii) The experimental results show that in unsupervised MT between gender and gender-neutral language, images contribute to improving gender accuracy.

2 Related Work

2.1 Multimodal Machine Translation Datasets

Existing commonly employed multimodal MT datasets include Multi30k (Elliott et al., 2016), IKEA (Grubinger et al., 2006), IAPR TC-12 (Elliott et al., 2016) and MS-COCO (Lin et al., 2014), and these datasets are all focused on high-resource DLPs such as English and German. Datasets IKEA

and IAPR TC-12 contain fewer images and description sentences. Multi30K dataset is not only immediately applicable to research on a wide range of tasks, it is collected from a wider range of fields. Moreover, Multi30k is the most commonly used dataset, and it contains 31k high-quality practical events. Therefore, we have manually translated it into Chinese and Uyghur and generated a low-resource DLPs dataset. MS-COCO dataset contains 164k images, each with five different English descriptive sentences. We automatically translate English sentences of MS-COCO into Chinese and Uyghur for the pre-training dataset.

2.2 Unsupervised Multimodal Machine Translation

While supervised MT relies on bilingual parallel corpora (Cho et al., 2014; Bahdanau et al., 2015), this approach often fails to effectively utilize monolingual corpora. To address this limitation, some recent studies (Lample et al., 2018; Artetxe et al., 2018) have proposed unsupervised MT that leverages monolingual corpora. However, the lack of target language supervision information poses a challenge, making it difficult for unsupervised MT to achieve the same high-quality translation as supervised MT. Therefore, improving model performance by incorporating visual information into unsupervised MT has gained significant attention from researchers (Su et al., 2019; Wang et al., 2021; Huang et al., 2021; Li et al., 2023a, 2022b).

UMMT investigates the possibility of using image disambiguation and improving unsupervised MT. Its core assumption, intuitively based on the immutability of images, suggests that descriptions of the same visual content in different languages should remain largely similar. However, existing research has primarily focused on high-resource CLPs, limiting the practical application of UMMT. To address this limitation and investigate UMMT in the context of DLPs, we construct a dataset containing DLPs. Furthermore, we incorporate image information into both pre-training and fine-tuning to improve translation performance.

3 Our Dataset

3.1 Distant Language Pairs

Language pairs are generally divided into CLPs and DLPs (Sun et al., 2021). Language similarity is determined by whether two languages belong to the same language family, whether they share

Table 1: Corpus-level statistics about Multi30k-Distant.

Splits	Sentences	Uyghur		Chinese		English	
		Tokens	Avg-length	Tokens	Avg-length	Tokens	Avg-length
Train	29,000	343,342	11.83	391,903	13.51	357,172	11.9
Validation	1,014	12,077	11.91	13,855	13.66	13,308	13.1
Test(Test2016)	1,000	11,834	11.83	13,566	13.57	12,968	13.0

words and sentences with the same word order, and so on. Most languages belong to different language families, and many of them suffer from a lack of resources. As shown in Figure 1, there are some gaps in the DLPs, such as English and Uyghur, they are written from different directions, and their scripts and word order are not the same, which also exists in Chinese-Uyghur. Moreover, Uyghur is a low-resource language, thus, English-Uyghur and Chinese-Uyghur are creating low-resource DLPs.

3.2 Data Collection

Multimodal MT mainstream corpus Multi30k (Elliott et al., 2016) contains 31k images and their descriptions in CLPs, e.g. English-German. To study UMMT on DLPs, we manually translate English sentences from Multi30k into Chinese and Uyghur. For Chinese, three native Chinese speakers with good English skills on our team, who are master students, are involved in the translation. For Uyghur, three native speakers with good English skills, all with bachelor’s degrees, participate in the translation. During the translation, the translator can access both the image and the English sentence, which facilitates the correct translation according to the image. To ensure the quality of translation, each translation sentence is further reviewed by another translator. It took about three months to complete the translation work. Statistics about our dataset Multi30k-Distant are shown in Table 1.

4 Method

In this section, we first detail the visual pivoting for UMMT and multimodal alignment, and then introduce the UMMT model and the training strategy.

4.1 Visual Pivoting for UMMT

Unsupervised MT assumes the availability of a monolingual corpus during training. It defines the input $T = [t_1, \dots, t_l]$ as a l -length sentence. Our model extends unsupervised MT by adding visual features $Z = [z_1, \dots, z_j]$, where j is the number of the most confident regions of an image. As shown

in Figure 2, the image is input in two forms, which are input and output to the encoder.

4.1.1 Encoder Input

We assume the availability of the sentence and image binary and redefine the input as:

$$M = [t_1, \dots, t_l, z_1, \dots, z_j] \quad (1)$$

As shown in Figure 2, each input to the encoder consists of a sentence and its corresponding image features. Specifically, for the source, the input is a concatenation of the source language sentence and its corresponding image features, denoted as $M_x = [x_1, \dots, x_n, z_{x1}, \dots, z_{xj}]$. Similarly, for the target, the input is a concatenation of the target language sentence and its corresponding image features, denoted as $M_y = [y_1, \dots, y_m, z_{y1}, \dots, z_{yj}]$. Where, x and y ($\{x\} \cap \{y\} = \phi$) represent source and target sentences, and z_x and z_y ($\{z_x\} \cap \{z_y\} = \phi$) represent their corresponding image features.

In this method, images not only improve the alignment of language latent spaces but also enhance the quality of incomplete pseudo-sentences through correction.

4.1.2 Encoder Output

This method employs an attention-gate structure (AGS) to fuse text and image features. The encoded sequence E and image features Z are integrated through an attention mechanism. Subsequently, a gate structure further combines E with the attention output H . This process can be represented as:

$$H = \text{Softmax} \left(\frac{EZ^T}{\sqrt{d}} \right) Z \quad (2)$$

$$g = \text{Sigmoid} (W_e E + W_h H) \quad (3)$$

$$H_f = (1 - g) \cdot E + g \cdot H \quad (4)$$

where W_e and W_h are trainable matrices. In this approach, images enable the model to avoid losing information during the encoding process, thus improving the prediction of the VPLM. Whereas in translation, the images serve as an approximate pivot point that connects the non-parallel sentences and thus improves the quality of the translation.

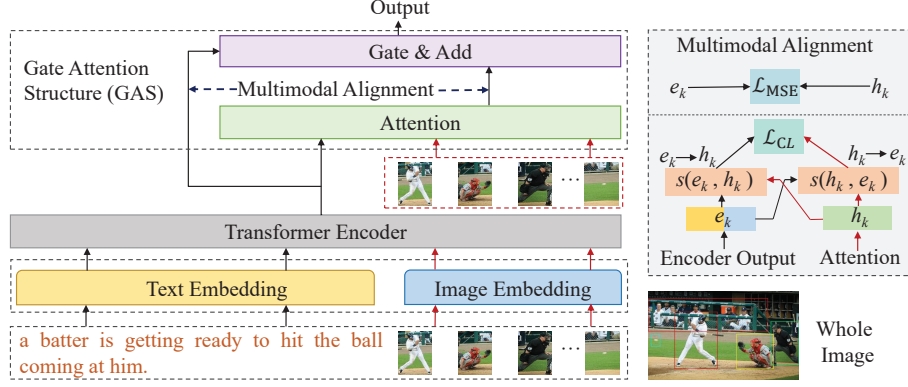


Figure 2: The framework of our multimodal fusion model.

4.2 Multimodal Alignment

We employ contrastive learning (Sohn, 2016) in cross-modal retrieval to align inputs in shared multilingual semantic space, where inputs are close when they are semantically related or paired. Specifically, we first generalize the encoding output E and the attention (Eq(2)) output H . The fine-grained alignment is then obtained by the cosine similarity $s(e_k, h_k)$ between the k -th token-level of E and H , where $e_k, h_k \in \mathbb{R}^d$. Finally, to bring the visual and textual modalities closer, we use noise contrastive estimation (van den Oord et al., 2018).

$$\begin{aligned} \mathcal{L}_{\text{CL}}^{e \rightarrow h} &= -\frac{1}{K} \sum_{k=1}^K \log \frac{\exp(s(e_k, h_k))}{\sum_{l=1}^K \exp(s(e_k, h_l))} \\ \mathcal{L}_{\text{CL}}^{h \rightarrow e} &= -\frac{1}{K} \sum_{k=1}^K \log \frac{\exp(s(h_k, e_k))}{\sum_{l=1}^K \exp(s(h_k, e_l))} \\ \mathcal{L}_{\text{CL}} &= \frac{1}{2} (\mathcal{L}_{\text{CL}}^{e \rightarrow h} + \mathcal{L}_{\text{CL}}^{h \rightarrow e}) \end{aligned} \quad (5)$$

where K is the sum of the sentence length and the number of regions in an image. We also utilize mean square error (MSE) losses to further minimize the distance between e_k and h_k .

$$\mathcal{L}_{\text{MSE}} = \frac{1}{2K} \sum_{k=1}^K \|e_k - h_k\|_2^2 \quad (6)$$

Finally, the multimodal alignment loss function is:

$$\mathcal{L}_{\text{MA}} = \mathcal{L}_{\text{CL}} + \lambda_1 \mathcal{L}_{\text{MSE}} \quad (7)$$

where the hyper-parameter λ_1 is set to 1.

4.3 Unsupervised Multimodal Machine Translation

Our UMMT model consists of a multimodal denoising auto-encoding (MDA) and a multimodal back-translation (MBT) model.

4.3.1 Multimodal Denoising Auto-encoding

MDA is extended by incorporating image features into denoising auto-encoding (Vincent et al., 2008). MDA is constructed by connecting the Transformer decoder to the output of Figure 2. It aims to improve the model learning ability by reconstructing noisy sentences in the same language. We create it separately for the unpaired source sentence x and target sentence y . The process in x is:

$$\text{Dec}_x(\text{Enc}_x(N(x), z_x), z_x) \rightarrow \hat{x} \quad (8)$$

where $N(\cdot)$ is the artificial noise function, which includes random deletion, swapping, and blanking. Firstly, the noisy source sentence $N(x)$ and its corresponding image feature z_x are introduced into the source language encoder $\text{Enc}_x(\cdot)$. The encoded output and image are then introduced into the source language decoder $\text{Dec}_x(\cdot)$, and the reconstructed sentence \hat{x} of $N(x)$ is obtained. Finally, supervised training is performed between x and \hat{x} . The reconstruction process on the target is similar to that on the source. The total MDA loss in x and y is:

$$\mathcal{L}_{\text{MDA}} = CE(\hat{x}, x) + CE(\hat{y}, y) \quad (9)$$

where $CE(\cdot, \cdot)$ represents cross-entropy loss.

4.3.2 Multimodal Back-Translation

MDA's training inputs and outputs still involve only one language, even though MT goal is to map input sentences from the source/target language to the target/source language. For cross-language training, we use MBT which is extended by adding image features to back-translation (Sennrich et al., 2016a). It explicitly guarantees that the model has translation ability without paired sentences. The MBT is carried out on the source sentence x and target sentence y respectively, and we analyze the source in detail here. As shown in Figure 3, first

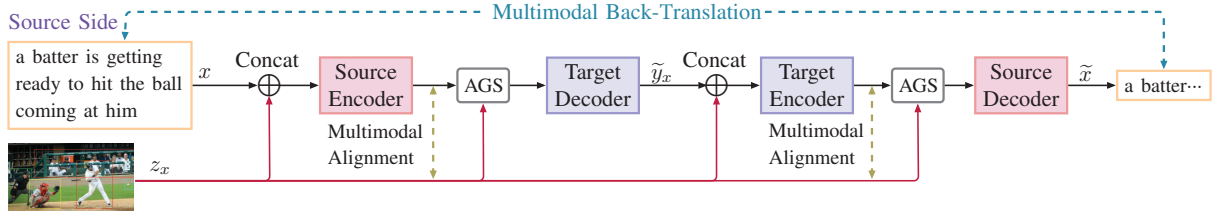


Figure 3: MBT framework in the source sentence x , since the framework in the target language is similar, we show only on the source language. Its encoder and decoder are from Transformer. AGS represents the fusion of text and image through the attention-gate structure, as shown in Figure 2.

given x and its corresponding image z_x , we apply the source language encoder $\text{Enc}_x(\cdot)$ and target language decoder $\text{Dec}_y(\cdot)$ trained in MDA to translate x into target sentence \tilde{y}_x . $\text{Enc}_x(\cdot)$ and $\text{Dec}_y(\cdot)$ are frozen, and they are involved in inferring:

$$\text{Dec}_y(\text{Enc}_x(x, z_x), z_x) \rightarrow \tilde{y}_x \quad (10)$$

x is the high-quality input, and z_x supplements the information lost during encoding and decoding, thereby improving the \tilde{y}_x . \tilde{y}_x and z_x are then fed to the target language encoder $\text{Enc}_y(\cdot)$ and the source language decoder $\text{Dec}_x(\cdot)$ translates \tilde{y}_x into \tilde{x} :

$$\text{Dec}_x(\text{Enc}_y(\tilde{y}_x, z_x), z_x) \rightarrow \tilde{x} \quad (11)$$

The total process $(x, z_x) \rightarrow (\tilde{y}_x, z_x) \rightarrow \tilde{x}$:

$$\text{Dec}_x(\text{Enc}_y([\text{Dec}_y(\text{Enc}_x(x, z_x), z_x)], z_x), z_x) \rightarrow \tilde{x} \quad (12)$$

Pseudo-input \tilde{y}_x is a corrupted version of unknown y_x , and the noisy inputs result in degraded translation performance. Therefore, z_x is introduced into $\text{Enc}_y(\cdot)$ to correct the pseudo-sentence and eliminate the noise. Whereas, the input z_x in $\text{Dec}_x(\cdot)$ is treated as a pivot between y_x and x . This is the process of translating two language sentences into each other, so they correspond to one image.

Training on the target side is similar to the source, the training process in target side:

$$\text{Dec}_y(\text{Enc}_x([\text{Dec}_x(\text{Enc}_y(y, z_y), z_y)], z_y), z_y) \rightarrow \tilde{y} \quad (13)$$

The total MBT loss in x and y is:

$$\mathcal{L}_{\text{MBT}} = CE(\tilde{x}, x) + CE(\tilde{y}, y) \quad (14)$$

4.4 Training Strategy

4.4.1 Pre-training

VPLM extends MLM (Conneau and Lample, 2019) by adding image features $\mathcal{Z} = [z_1, \dots, z_j]$ in domain \mathcal{Z} . The framework of VPLM is built by adding a prediction layer to the output of Figure 2.

Similar to MLM, 15% of the text and image region are randomly selected for masking. The objective function of VPLM is a combination of MLM loss and masked region classification loss. It is the masking text \tilde{t} and region \tilde{z} against ground truth text target \tilde{t} and region label \tilde{z} :

$$\begin{aligned} \mathcal{L}_{\text{VPLM}} = & \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} -\log p(\tilde{t} | \tilde{t}; \theta_p) \\ & + \frac{1}{|\mathcal{Z}|} \sum_{z \in \mathcal{Z}} -\log p(\tilde{z} | \tilde{z}; \theta_p) \end{aligned} \quad (15)$$

where θ_p is the model parameter. VPLM is trained with multimodal alignment:

$$\mathcal{L}_{\text{Pre}} = \mathcal{L}_{\text{VPLM}} + \lambda_2 \mathcal{L}_{\text{MA}} \quad (16)$$

where the hyper-parameter λ_2 is set to 1.

4.4.2 Fine-tuning

MDA and MBT are initialized with VPLM and then fine-tuned, and they are also trained with multimodal alignment:

$$\mathcal{L}_{\text{Fin}} = \mathcal{L}_{\text{MDA}} + \lambda_3 \mathcal{L}_{\text{MBT}} + \lambda_4 \mathcal{L}_{\text{MA}} \quad (17)$$

where the hyper-parameter λ_3 and λ_4 are set to 1. MDA and MBT are cycle-trained, and their parameters are fully shared between them.

5 Experiments

5.1 Experimental Setup

For pre-training, we use the training and validation set of the MS-COCO (Lin et al., 2014) dataset. To construct the monolingual data, this dataset is randomly split into two disjoint subsets. Each set contains 64,542 images and five English descriptive sentences for each image. Then we apply Lingvanex¹ translator to translate English sentences into German, Chinese, and Uyghur.

For fine-tuning, we performed experiments on both the Multi30k and Multi30k-Distant. To ensure

¹<https://lingvanex.com/translate>

Table 2: Results for DLPs translation. Uyghur and Chinese are not supported by METEOR.

	En → Uy			Uy → En			Zh → Uy			Uy → Zh		
	RIBES↑	BLEU↑	TER↓	RIBES↑	BLEU↑	TER↓	RIBES↑	BLEU↑	TER↓	RIBES↑	BLEU↑	TER↓
XLM(Text-only)	53.2	2.6	96.4	54.4	3.1	87.1	51.9	2.6	92.3	58.8	3.9	89.4
UMNMT	65.1	7.4	83.2	65.9	8.0	74.9	67.4	10.6	75.8	71.0	14.1	74.7
M-Transformer	70.4	11.5	76.1	70.2	11.3	75.6	70.7	17.2	71.0	73.7	21.2	68.8
IVTA	69.8	13.2	74.9	71.0	13.7	69.8	76.9	22.4	63.1	77.5	24.5	61.3
VUMMT	73.3	15.7	72.3	75.1	16.0	75.5	81.9	28.7	53.1	79.8	33.2	52.4
Ours	76.4	20.9	66.1	81.1	20.6	64.8	86.5	32.2	50.4	85.9	37.0	46.7

that the model avoids learning from parallel sentences, as with the pre-training data, the training set of a language is randomly divided into two, and two non-parallel corpora with 14,500 samples of training set are produced.

For Chinese, we use the tokenizer of Chang et al. (2008). For all other languages, Moses (Koehn et al., 2007) toolkit is used to tokenize all sentences. We use byte pair encoding (BPE) (Sennrich et al., 2016b) and use fastBPE² to learn the BPE code and split words into sub-word units.

Model dimension and feedforward dimension are set to 512 and 2,048. The Adam (Kingma and Ba, 2015) optimizer with a learning rate of 1×10^{-4} is used for optimization. Experiments are implemented on a machine with a single 12GB TITAN Xp GPU. For image features, we follow (Caglayan et al., 2021), using Faster R-CNN (Ren et al., 2015) models to extract features $[z_1, \dots, z_j]$, where the number of regions j is set at 36 and $z_i \in \mathbb{R}^{1536}$.

5.2 Baselines and Evaluation Metrics

Baselines To verify the performance of our model, We mainly compare with the existing models: (1) **XLM** (Conneau and Lample, 2019) is a monolingual text-only unsupervised MT based on MLM. (2) **UMNMT** (Su et al., 2019) is established on multimodal monolingual data by two training paths, such as auto-encoding loss and cycle-consistency loss. (3) **M-Transformer** (Huang et al., 2021) uses additional visual modalities to recover sentences that have previously masked some words. (4) **IVTA** (Li et al., 2023a) is semi-supervised MT and includes both unsupervised and supervised training components. (5) **VUMMT** (Tayir and Li, 2024) studies the effect of practice measures for UMMT. (6) **Game-MMT** (Chen et al., 2018) is a reinforcement learning-based UMMT. (7) **Knwl.** (Huang et al., 2023) introduces knowledge entities as an additional modality to enhance the representation.

²<https://github.com/glample/fastBPE>

Since Game-MMT, Progressive, and Knwl. are only translated between English, German, and French, they also take advantage of pre-trained models and knowledge entities that exist only in those languages. Therefore, they cannot be reproduced using our dataset.

Evaluation metrics We apply MT metrics such as RIBES (Isozaki et al., 2010), BLEU (Papineni et al., 2002), METEOR (Lavie and Agarwal, 2007) and TER (Snover et al., 2006) to evaluate translation quality. RIBES is mainly utilized to evaluate the translation quality between DLPs, while METEOR is not supported for Uyghur and Chinese.

Moreover, Uyghur is a gender-neutral language (e.g. “he”, “she” and “it” are all translated into “u”), whereas others are gendered languages. Almost 14% of the sentences in the fine-tuned data contain gender pronouns, which affects the translation between Uyghur and other languages. This paper argues that the image provides information to correct the **gender accuracy** of the translation. We scored the correctness of the gender pronoun by examining the gender pronoun in the translation and its reference sentence (wrong: 0, correct: 1). Then, the gender accuracy is obtained by dividing the whole test set score by the number of gender pronouns in the reference set.

5.3 Overall Performance

5.3.1 Results on Distant Language Pairs

In Table 2, the baselines are reproduced using our datasets. IVTA is semi-supervised model with 300 parallel corpora, it is trained from scratch.

Text-only models The results of XLM show that the translation quality of the model deteriorates significantly when using text-only data. Although XLM is initialized with a pre-trained model trained on 322,710 monolingual sentences, it fails to translate complete sentences. Compared to this, XLM on CLPs provides quite satisfactory experimental results, as shown in Table 4.

Table 3: Experimental results (BLEU) of different multimodal inputs and alignments. Gender Accuracy (introduced in Section 5.2) is the average of the accuracy in both Uy→En and Uy→Zh. Eq(1): text and image concat inputs, Eq(2): text and image inputs via AGS and \mathcal{L}_{MA} : multimodal alignment.

Eq(1)	Eq(2)	\mathcal{L}_{MA}	En-Uy		Zh-Uy		En-De		Gender Accuracy
			→	←	→	←	→	←	
			2.6	3.1	2.6	3.9	26.4	29.8	18.2
✓			15.7	16.0	28.7	33.2	29.4	33.2	59.8
	✓		15.3	15.6	28.1	32.9	28.0	32.6	58.2
✓	✓		19.3	19.2	31.4	35.5	28.1	32.3	65.5
	✓	✓	16.5	16.3	29.0	34.2	28.7	32.9	60.4
✓	✓	✓	20.4	19.9	31.8	36.7	28.4	32.5	67.2

Table 4: Results for CLP translation.

	En→De		De→En	
	BLEU	METEOR	BLEU	METEOR
XLM(Text-only)	26.4	45.2	29.8	29.9
Game-MMT	16.6	–	19.6	–
UMNMT	23.5	26.1	26.4	29.7
M-Transformer	26.7	–	29.8	–
Knwl.	28.9	–	31.8	–
IVTA	22.9	39.7	25.5	29.2
VUMMT	29.4	48.8	33.2	32.5
Ours	30.7	50.1	34.4	33.4

Table 5: Human evaluations on DLPs. Com., Amb., and Flu. stand for Completeness, Ambiguity, and Fluency. Results are averaged on En→Uy and Zh→Uy.

	Avg. BLEU	Human evaluations		
		Com.↑	Amb.↓	Flu.↑
M-Transformer	15.6	4.3	7.2	4.6
IVTA	17.1	4.5	7.1	4.9
GPT-4	17.8	5.1	6.8	5.1
Ours	25.5	5.6	6.1	5.7

Multimodal models The performance of the model is significantly improved when images are introduced for fine-tuning (XLM Vs. UMNMT, M-Transformer). For example, for En → Uy, the BLEU of UMNMT is 7.4, which is larger than that of XLM, i.e., 2.6. Meanwhile, for Uy → Zh, their BLEU gap is 10.2. IVTA achieves comparatively better results, which indicates that a small number of parallel corpora can significantly improve the translation. Among the baseline models, VUMMT yields the best results, while our model benefits from the outstanding image introduction method.

5.3.2 Results on Close Language Pairs

Table 4 shows the original paper and our experimental results. Our model yielded superior per-

formance, which outperformed the text-only and multimodal baselines. Notably, our model has a BLEU score of 1.3 and 1.2 higher than VUMMT. Compared to DLPs, XLM results on CLP yield better results because the relationship between the two languages can be learned without images.

5.4 Human Evaluation

For each model, we randomly sampled 100 sentences from its test translations and rated each sentence on a scale from 0 to 10 according to their quality. As listed in Table 5, our model shows the highest BLEU among the three manually evaluated models. We also compare ours with the translation results of GPT-4³. Our model BLEU reaches 25.5 with 43.2% to 63.4% improvements over GPT-4 and M-Transformer. In terms of FLU. to measure the translation cohesion and fluency, our model still shows best among three human evaluations.

5.5 Effects of Image Pivoting

5.5.1 Multimodal Inputs and Alignment

Compared to the text-only model (the first row), the model with connected image features (Eq (1)) has a significant improvement, and the score on BLEU increases by 3.0 to 29.3, as shown in Table 3. Whether it is the concatenated image introduction (Eq(1)) or the AGS image introduction (Eq(2)) method, they all bring great improvement to DLPs, and CLP achieves the best result in Eq(1). When these two methods are used together, the translation quality of DLPs continues to improve, while that of CLP decreases significantly. This means that, to some extent, richer images serve to bridge the gap between DLPs, thus improving translation performance. Moreover, the translation of both languages pairs has been improved with the addition

³<https://openai.com/gpt-4>

Table 6: Experimental results (BLEU) of images on different branch models. VPLM: visual pre-training language modeling, MDA: multimodal denoising auto-encodin model and MBT: multimodal back-translation model.

		VPLM	MDA	MBT	En-Uy		Zh-Uy		En-De	
					→	←	→	←	→	←
Image					2.6	3.1	2.6	3.9	26.4	29.8
			✓	✓	7.4	8.3	9.2	12.7	28.6	32.8
		✓			9.3	10.8	26.6	30.9	28.2	31.6
		✓	✓		8.6	9.5	16.9	18.7	17.6	25.4
		✓		✓	15.4	16.2	28.4	33.4	28.8	32.9
		✓	✓	✓	15.7	16.0	28.7	33.2	29.4	33.2

of multimodal alignment (\mathcal{L}_{MA}) methods.

In terms of gender accuracy, two image introduction methods and multimodal alignment gradually improved gender recognition, which validated our hypothesis that image fusion is conducive to the correct translation of gender pronouns.

5.5.2 Images on Different Branch Models

As described in Section 4, our model consists of three modules, VMLM, MDA, and MBT. As shown in Table 6, compared to the text-only model (the first row), the performance of the fine-tuned translation model containing images is improved on both language pairs. Images from pre-trained models provide a significant reinforcement to DLPs. Images have a positive effect on all branches, and the best translation results are achieved when they all are fused with the image. However, the introduction of images to the MDA without the inclusion of images for MBT impairs the model performance and the image in the MDA effect is not significant.

5.6 Experimental Results on Image Features with Different Granularity

We argue that regional features are extracted based on object confidence, which may ignore the relationships between objects and their background information. Therefore, we also discuss the grid features $G \in \mathbb{R}^{49 \times 2,048}$ extracted by using Resnet-101 (He et al., 2016). As shown in Table 7, we conducted experiments on the region and grid features individually and together. Two image features are combined in a concatenated manner. The experimental results show that DLPs are significantly improved in both features, while CLP is in grid features. This validates our hypothesis that DLPs require richer image features.

Table 7: Experimental results (BLEU) of image features with different granularity. The experiments are based on the best model of Table 3 and 6. Reg. and Gri. represent region and grid features, and Reg.&Gri. indicates the both features.

		En-Uy		Zh-Uy		En-De	
		→	←	→	←	→	←
Reg.		20.4	19.9	31.8	36.7	29.4	33.2
Gri.		20.2	20.3	29.9	33.5	30.7	34.4
Reg&Gri.		20.9	20.6	32.2	37.0	29.0	32.3

6 Analyses

6.1 Case Study


To further demonstrate the validity of our model, we show the translation results generated by different models, as shown in Table 8. We can see that our added image information is more helpful when the translated object encounters ambiguity. XLM translates the source word “blue” to “säriq renglik” (yellow), and the model with the added image translates correctly. It is interesting to observe that our model, as defined in Eq(1), extracts more information from images in complex scenes and translates information that is not present in the reference sentence but is present in the image, e.g., “öz’ara paranglishiwatidu” (talking to each other).

Moreover, we also compare ours with GPT-4. The translation result indicates that although the objects in the input sentence are correctly translated, the relations between them are not.

6.2 Bucketed Analysis

To find significant differences between translated sentences, we used the compare-mt toolkit (Neubig et al., 2019) for analysis. We provide an example of sentence hierarchy analysis, as shown in Figure 4. In the number of translated sentences with different BLEU values, it can be found that

Table 8: Case study. Eq(1) represents the model in the second row of Table 3

	SRC(En):	a group of men in blue uniforms are standing together.
	REF(Uy):	bir top kök renglik forma kiygen erler bille turidu.
	XLM(Text-only):	bir top sèriq renglik kiyim kiygen.
	Eq(1):	bir top kök renglik kiyim kiygen erler öz'ara paranglishiwatidu.
	Ours:	bir top kök renglik forma kiygen erler bille turdi.
GPT-4:		kök uniformadiki bir gurup er adamlar birge turdu.

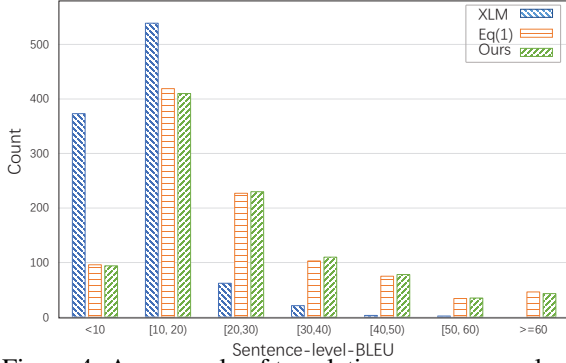


Figure 4: An example of translation accuracy analysis in the En→Uy task.

when the BLEU value is less than 20, the relationship between the number of sentences from small to large is inversely related to the output quality of our model for the whole test set. However, when the BLEU value is greater than 20, the relationship between the number of sentences is consistent with the model’s score on the whole test set. XLM has no more than 60 BLEU sentences in translation.

6.3 Supervised Case

While this paper primarily focuses on unsupervised MT with images as pivots, we are also interested in exploring supervised translation on our dataset. As shown in Table 9, we conducted supervised MT experiments by switching from back-translation to a transformer-based framework with additional image features. We benchmarked recent supervised MT models, including **Transformer**(text-only) (Vaswani et al., 2017), **Selective-attn** (Li et al., 2022a), **RG-MMT-EDC** (Tayir et al., 2024), and **VTLM** (Caglayan et al., 2021). VTLM and our model are both pre-trained and fine-tuned on our dataset Multi30k-Distant.

It can be seen from the experimental results that our method shows the best than other baselines. Compared with VTLM, our fusion method is more effective for supervised MT. It is noteworthy that images provide marginal improvements in supervised DLPs translation.

Table 9: Supervised results (BLEU) on Multi30K-Distant.

	En-Uy		Zh-Uy	
	→	←	→	←
Transformer	40.4	36.0	61.9	61.2
Selective-attn	41.2	36.6	62.1	61.2
RG-MMT-EDC	41.7	36.5	62.4	62.1
VTLM	42.5	38.2	64.5	64.1
Ours	44.8	39.8	65.3	64.9

7 Conclusions

In this work, we first create a dataset containing two DLPs to investigate UMMT on low-resource DLPs. We then found that cross-language alignment in shared latent spaces can be improved by incorporating visual content in both pre-trained and fine-tuned models. Compared to the baseline model, our model has 5.2 and 4.6 BLEU score improvements in English-Uyghur translation, and 3.5 and 3.8 BLEU score improvements in Chinese-Uyghur translation. Moreover, the experimental results show that images contribute to improving gender accuracy in translation between gender and gender-neutral languages.

8 Limitations

Although our method achieves good results, it also has some limitations in dealing with DLPs. As can be seen from Figure 4, incorporating more image features may hurt the accuracy of a high-score translated sentence. More persons are needed to join our human evaluations since translation is subjective to some degree. The differences among different persons could be analyzed in detail.

Acknowledgments

This work is partially supported by NSFC, China (No.62276196).

References

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural

- machine translation. In *International Conference on Learning Representations*, pages 1–12.
- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*, pages 1–15.
- Ozan Caglayan, Menekse Kuyuu, Mustafa Sercan Amac, Pranava Madhyastha, Erkut Erdem, Aykut Erdem, and Lucia Specia. 2021. Cross-lingual visual pre-training for multimodal machine translation. In *Conference of the European Chapter of the Association for Computational Linguistics*, pages 1317–1324.
- Pi-Chuan Chang, Michel Galley, and Christopher D. Manning. 2008. Optimizing chinese word segmentation for machine translation performance. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 224–232.
- Yun Chen, Yang Liu, and Victor O. K. Li. 2018. Zero-resource neural machine translation with multi-agent communication game. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th Innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence*, pages 5086–5093.
- Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. In *Proceedings of SSST@EMNLP 2014, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111.
- Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, pages 7057–7067.
- 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.
- Michael Grubinger, Paul Clough, Henning Müller, and Thomas Deselaers. 2006. The iapr tc-12 benchmark: A new evaluation resource for visual information systems. In *International workshop ontoImage*, pages 1–11.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.
- Ping Huang, Shiliang Sun, and Hao Yang. 2021. Image-assisted transformer in zero-resource multi-modal translation. In *International Conference on Acoustics, Speech and Signal Processing*, pages 7548–7552.
- Ping Huang, Jing Zhao, Shilinag Sun, and Yichu Lin. 2023. Knowledge enhanced zero-resource machine translation using image-pivoting. *Appl. Intell.*, 53(7):7484–7496.
- Po-Yao Huang, Junjie Hu, Xiaojun Chang, and Alexander G. 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.
- Hideki Isozaki, Katsuhito Sudoh, Hajime Tsukada, and Kevin Duh. 2010. Head finalization: A simple reordering rule for SOV languages. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 244–251.
- Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Computer Vision and Pattern Recognition*, pages 1–15.
- Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 177–180.
- Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc’Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *6th International Conference on Learning Representations*, pages 1–14.

- Alon Lavie and Abhaya Agarwal. 2007. METEOR: an automatic metric for MT evaluation with high levels of correlation with human judgments. In *The Second Workshop on Statistical Machine Translation*, pages 228–231.
- Bei Li, Chuanhao Lv, Zefan Zhou, Tao Zhou, Tong Xiao, Anxiang Ma, and Jingbo Zhu. 2022a. On vision features in multimodal machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 6327–6337.
- Lin Li, Kaixi Hu, Turghun Tayir, Jianquan Liu, and Kong Aik Lee. 2022b. Noise-robust semi-supervised multi-modal machine translation. In *Trends in Artificial Intelligence - 19th Pacific Rim International Conference on Artificial Intelligence*, pages 155–168.
- Lin Li, Turghun Tayir, Yifeng Han, Xiaohui Tao, and Juan D. Velásquez. 2023a. Multimodality information fusion for automated machine translation. *Information Fusion*, 91:352–363.
- Mingjie Li, Po-Yao Huang, Xiaojun Chang, Junjie Hu, Yi Yang, and Alex Hauptmann. 2023b. Video pivoting unsupervised multi-modal machine translation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(3):3918–3932.
- Tsung-Yi Lin, Michael Maire, Serge J. 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 - 13th European Conference*, pages 740–755.
- Kelly Marchisio, Kevin Duh, and Philipp Koehn. 2020. When does unsupervised machine translation work? In *Proceedings of the Fifth Conference on Machine Translation*, pages 571–583.
- Graham Neubig, Zi-Yi Dou, Junjie Hu, Paul Michel, Danish Pruthi, and Xinyi Wang. 2019. compare-mt: A tool for holistic comparison of language generation systems. In *Proceedings of NAACL-HLT 2019*, pages 35–41.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.
- Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: towards real-time object detection with region proposal networks. In *Conference on Neural Information Processing Systems*, pages 91–99.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 86–96.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725.
- Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*, pages 223–231.
- Kihyuk Sohn. 2016. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems*, pages 1849–1857.
- Yuanhang Su, Kai Fan, Nguyen Bach, C.-C. Jay Kuo, and Fei Huang. 2019. Unsupervised multimodal neural machine translation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 10482–10491.
- Haipeng Sun, Rui Wang, Masao Utiyama, Benjamin Marie, Kehai Chen, Eiichiro Sumita, and Tiejun Zhao. 2021. Unsupervised neural machine translation for similar and distant language pairs: An empirical study. *ACM Trans. Asian Low Resour. Lang. Inf. Process.*, 20(1):1–17.
- Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.
- Turghun Tayir and Lin Li. 2024. Unsupervised multimodal machine translation for low-resource distant language pairs. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, pages 1–22.

- Turghun Tayir, Lin Li, Bei Li, Jianquan Liu, and Kong Aik Lee. 2024. Encoder-decoder calibration for multimodal machine translation. *IEEE Trans. Artif. Intell.*, 5(8):3965–3973.
- Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748.
- 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*, pages 5998–6008.
- Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. In *Machine Learning, Proceedings of the Twenty-Fifth International Conference*, pages 1096–1103.
- Yijun Wang, Tianxin Wei, Qi Liu, and Enhong Chen. 2021. Unpaired multimodal neural machine translation via reinforcement learning. In *Database Systems for Advanced Applications - 26th International Conference*, pages 168–185.
- Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575.