



November 12 –16  
Miami, Florida



University of  
Southern  
Queensland  
Australia



NEC

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

Turghun Tayir<sup>1</sup>, Lin Li<sup>1</sup>, Xiaohui Tao<sup>2</sup>, Mieradilijiang Maimaiti<sup>3</sup>, Ming Li<sup>1</sup>, Jianquan Liu<sup>4</sup>

<sup>1</sup>Wuhan University of Technology, China

<sup>2</sup>University of Southern Queensland, Australia

<sup>3</sup>Chinese Academy of Sciences, China

<sup>4</sup>NEC Corporation, Japan

{hotpes, cathylin, liming7677}@whut.edu.cn, xiaohui.tao@unisq.edu.au miradel\_51@hotmail.com, jqliu@nec.com

# Introduction

- **Problem:**

Unsupervised machine translation (UMMT) struggles with distant language pairs (DLPs) due to limited linguistic similarity and data scarcity.

- **Motivation:**

Leveraging visual information as a pivot can improve translation performance, especially for DLPs.


- **Goal:**

This paper presents a visual pivoting approach for UMMT to enhance alignment between DLPs.

# Visual Pivoting for UMMT

- **Dataset:**

We manually translated the English (En) sentences in Multi30k into Chinese (Zh) and Uyghur (Uy). Our dataset **Multi30k-Distant** includes two DLPs (English-Uyghur and Chinese-Uyghur).

Image	Captions in four languages	
	Distant language pairs (En-Uy, Zh-Uy):	
	En:	A baseball player is fielding a ball.
	Uy:	كالتەك توپ تەنھەرىكەتچىسى توپ تۇتۇۋاتىدۇ.
	Zh:	一个棒球运动员正在接球。
	Close language pairs (En-De):	
	En:	A baseball player is fielding a <u>ball</u> .
	De:	Ein Baseballspieler spielt den <u>Ball</u> .

**Fig. 1:** Simple examples of distant and close language pairs.

- Words with the same color have the same meaning in different language.

Key characteristics of DLPs: Limited vocabulary overlap, divergent grammar, different writing systems and so on.

# Visual Pivoting for UMMT

## • Multimodal Fusion:

### ■ Encoder Input:

Concatenation of sentence and its corresponding image features.

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

### ■ Encoder Output:

We employ an attention-gate structure to fuse text and image features.

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

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

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

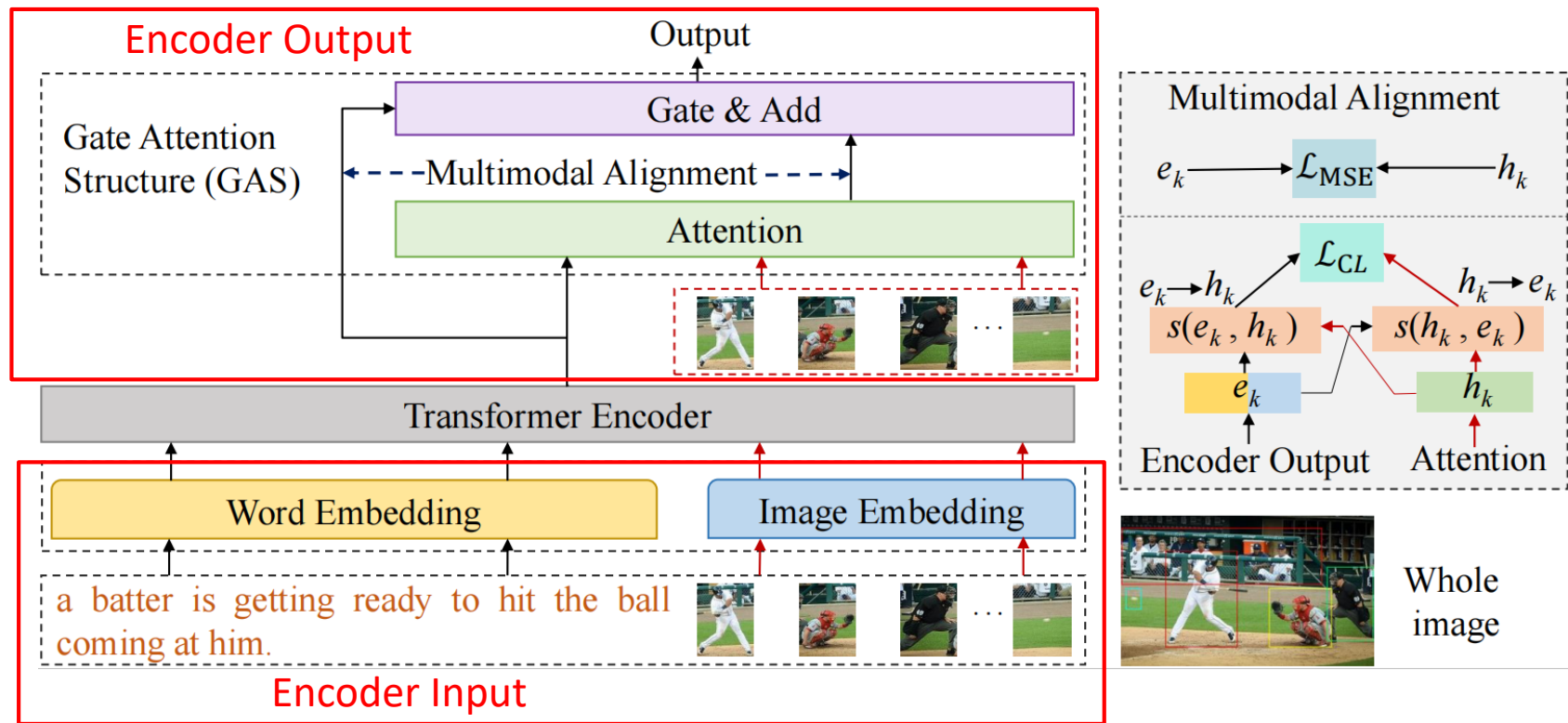


Fig. 2: The framework of our multimodal encoder.

# Visual Pivoting for UMMT

## • Multimodal Fusion:

### ■ Multimodal Alignment

We employ contrastive learning in cross-modal retrieval to align inputs in shared multilingual semantic space.

$$\mathcal{L}_{\text{CL}}^{e \rightarrow h} = -\frac{1}{K} \sum_{k=1}^K \log \frac{\exp(s(e_k, h_k))}{\sum_{l=1}^K \exp(s(e_k, h_l))}$$

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

$$\mathcal{L}_{\text{CL}} = \frac{1}{2} (\mathcal{L}_{\text{CL}}^{e \rightarrow h} + \mathcal{L}_{\text{CL}}^{h \rightarrow e})$$

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

The multimodal alignment loss function is:

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

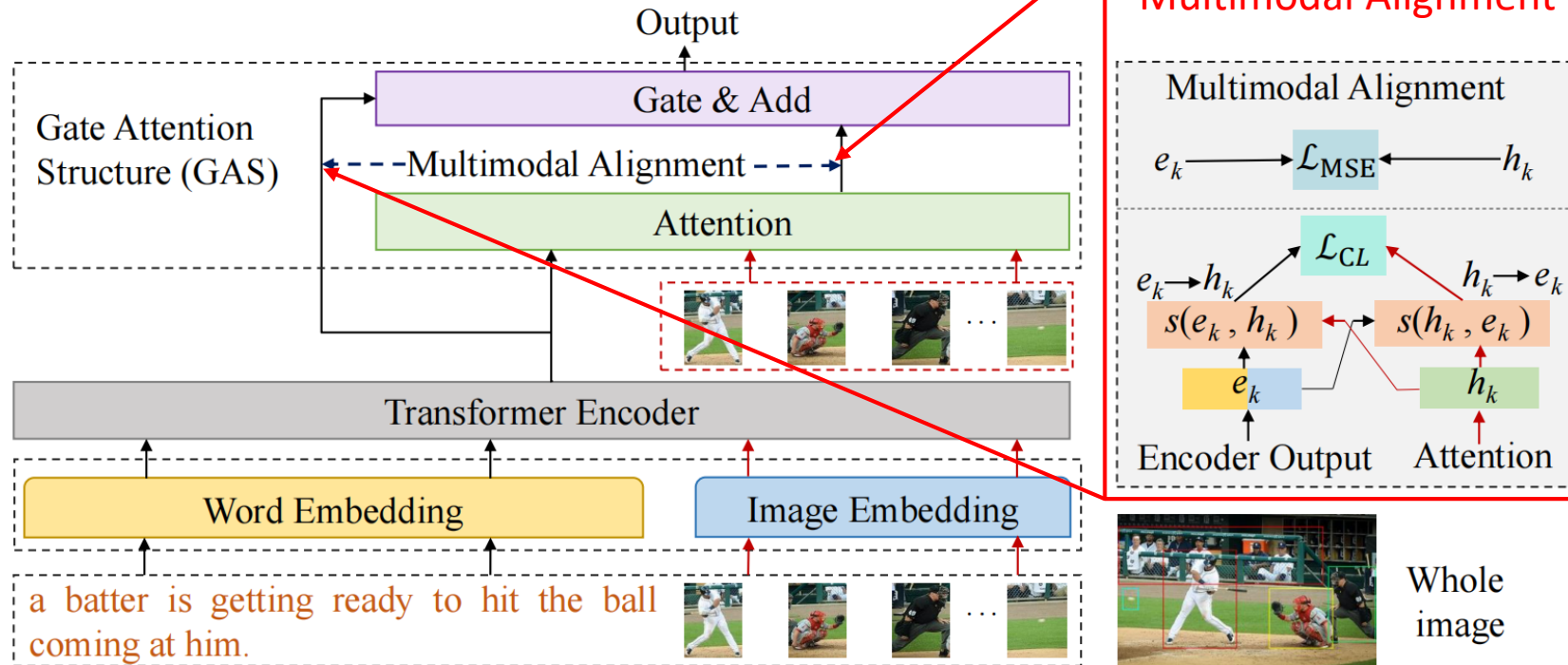
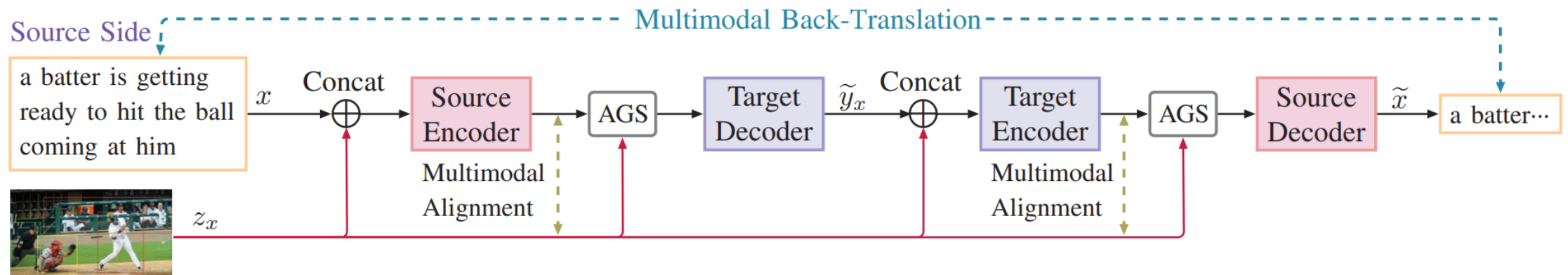


Fig. 2: The framework of our multimodal encoder.

# Visual Pivoting for UMMT

- **UMMT Model:**

- Our UMMT model consists of multimodal denoising auto-encoding (MDA) and multimodal back-translation (MBT) model.
- MDA is extended by incorporating image features. It aims to improve the model learning ability by reconstructing noisy sentences in the same language.
- For cross-language training, we use MBT which is extended by adding image features to back-translation. It explicitly guarantees that the model has translation ability without paired sentences. MBT is carried out on the source sentence  $x$  and target sentence  $y$  respectively, and we analyze the source in detail in **Fig.3**.



**Fig.3:** Multimodal back-translation framework in the source sentence  $x$

# Experimental Results

- **Datasets:**

- For pre-training, we use the MS-COCO dataset, randomly splitting it into two sets of 64,542 images with five English descriptions each. We then use the Lingvanex translator to translate these English sentences into German, Chinese, and Uyghur, creating monolingual datasets for each language.
- We fine-tuned our model on **Multi30k** for close language pairs and on **Multi30k-Distant** for distant language pairs. We randomly split each language's training set into two non-parallel corpora, each containing 14,500 samples.

- **Metrics:**

Translation quality: **RIBES**, **BLEU**, **TER** and **METEOR**.

- **Semantic similarity:** RIBES is helpful for evaluating translations that emphasize gist or paraphrasing.
- **General quality:** BLEU and METEOR are good choices for overall quality assessment.
- **Surface-level errors:** TER is useful for identifying word order and grammatical errors.
- **Gender accuracy:** We scored the correctness of the gender pronoun by examining the gender pronoun in the translation and its reference sentence.

# Experimental Results

- Performance Comparison:

- Results on Distant Language Pairs

Table 1: Results for DLPs translation. Uyghur and Chinese are not supported by METEOR.

	En $\rightarrow$ Uy			Uy $\rightarrow$ En			Zh $\rightarrow$ Uy			Uy $\rightarrow$ Zh		
	RIBES $\uparrow$	BLEU $\uparrow$	TER $\downarrow$	RIBES $\uparrow$	BLEU $\uparrow$	TER $\downarrow$	RIBES $\uparrow$	BLEU $\uparrow$	TER $\downarrow$	RIBES $\uparrow$	BLEU $\uparrow$	TER $\downarrow$
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
<b>Ours</b>	<b>76.4</b>	<b>20.9</b>	<b>66.1</b>	<b>81.1</b>	<b>20.6</b>	<b>64.8</b>	<b>86.5</b>	<b>32.2</b>	<b>50.4</b>	<b>85.9</b>	<b>37.0</b>	<b>46.7</b>

Translation between English and Uyghur.

Translation between Chinese and Uyghur.

- Although XLM is initialized with a pre-trained model trained on 322,710 monolingual sentences, it fails to translate complete sentences.
- Compared to the baseline model, our model benefits from the image introduction and multimodal alignment approach.

# Experimental Results

- Performance Comparison:

- Results on Close Language Pairs

Translation between English and German

- Compared to this, XLM on CLP provides quite satisfactory experimental results.
  - Among the baseline models, VUMMT yields the best results, while our model benefits from the outstanding image introduction method.

Table 2: 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

# Experimental Results

- Human Evaluation:

- 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.

- Multimodal Inputs and Alignment:

- 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.

Table 3: 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

Table 4: Experimental results (BLEU) of images on different branch models. VPLM: visual pre-training language modeling, MDA: multimodal denoising auto-encoding model and MBT: multimodal back-translation model.

			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

# Experimental Results

- Image Features with Different Granularity:

- Reg. : region features, using Faster R-CNN to extract
- Gri. : grid features, using Resnet101 to extract
- DLPs are significantly improved in both features, while CLP is in grid features.

- Supervised Case:

- Our supervised method shows the best than other baselines.

Table 5: Experimental results (BLEU) of image features with different granularity.

	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	<b>30.7</b>	<b>34.4</b>
Reg&Gri.	<b>20.9</b>	<b>20.6</b>	<b>32.2</b>	<b>37.0</b>	29.0	32.3

Table 6: 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
<b>Ours</b>	<b>44.8</b>	<b>39.8</b>	<b>65.3</b>	<b>64.9</b>

# Experimental Results

## • Bucketed Analysis:

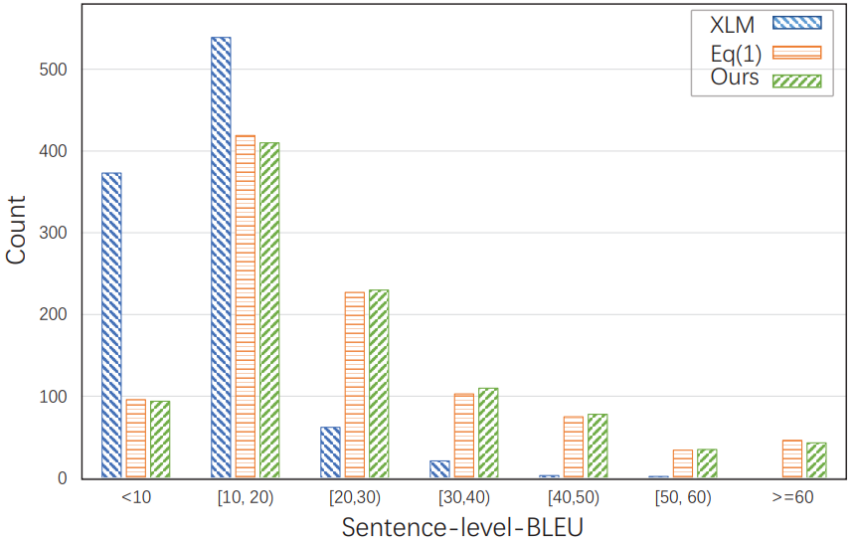
- It can be found that when the BLEU value is less than 20, the relationship between the number of sentences from small to large is inversely related to the output quality of our model for the whole test set.
- XLM has no more than 60 BLEU sentences in translation.

## • Case Study:

- Our model 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.



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 səriq 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.



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

# Discussion

- **Contributions:**

- 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.
- We find that visual content is more qualified to improve the alignment of DLPs latent space.
- The experimental results show that in unsupervised MT between gender and gender-neutral language, images contribute to improving gender accuracy.

- **Limitations:**

- As can be seen from Fig. 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.

# Conclusions

- We 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.