The Visual Prism: Refracting Images into Parallel Multilingual Descriptions with Structured Visual Guidance

Chengpeng Fu[1,2], **Xiaocheng Feng**[1,2*], **Yichong Huang**[1], **Wenshuai Huo**[1,2], **Baohang Li**[1], **Yang Xiang**[2], **Ting Liu**[1,2]

[1]Harbin Institute of Technology, 92 Xidazhi Street, Nangang District, Harbin, Heilongjiang Province, China
[2]Peng Cheng Laboratory, No.6001 Shahu West Road, Nanshan District, Shenzhen, Guangdong Province, China
{cpfu, xcfeng, ychuang, wshuo, baohangli, tliu}@ir.hit.edu.cn, xiangyang.hitsz@gmail.com

Abstract

Parallel corpora, as the foundation of machine translation, remain crucial even in the era of large language models (LLMs) for pre-training and fine-tuning. However, annotating parallel corpora is extremely costly, as it requires annotators to be proficient in multiple languages. To reduce this cost, prior work has explored image-pivoted corpus synthesis, generating multilingual captions for the same image as pseudo-parallel data. Unfortunately, these pseudo corpora suffer from the serious issue of **multilingual focus divergence**, *i.e.*, the model attending to distinct aspects of the image when generating captions in different languages. To address this problem, we propose a method called **PRISMS** (**P**arallel **R**efracting **I**mage**S** into **M**ultilingual descriptions with **S**tructured visual guidance), which leverages semantic graphs as structured visual guidance to unify the focus of multilingual captions. To ensure adherence to this guidance, we introduce two key techniques: supervised fine-tuning using self-generated instructional data, and reinforcement learning with a reward signal based on semantic graph consistency. Experimental results on five languages show that our PRISMS significantly improves the image-pivot parallel corpora synthesis, enabling LLMs to achieve translation performance comparable to that of models trained on manually annotated corpora.

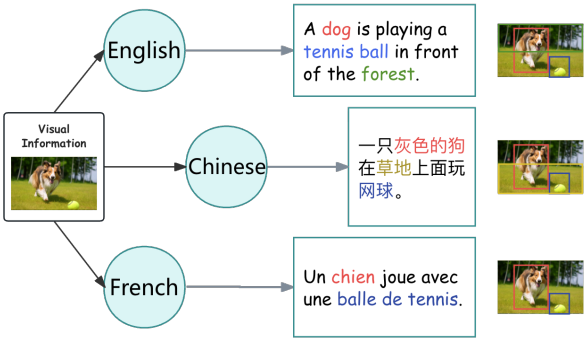

Figure 1: Visual information can act as a bridge to connect different languages. When generating descriptions in different languages using Multimodal Large Language Models, different models tend to focus on different parts of an image, leading to varying descriptions. In this figure, different colors for each language correspond to elements in the image framed with the matching colored boxes.

Introduction

Parallel corpora have long been fundamental to machine translation, underpinning both statistical approaches (Brown et al. 1990, 1993; Och and Ney 2002; Koehn, Och, and Marcu 2003; Lopez 2008), and neural approaches (Vaswani et al. 2017; Castilho et al. 2017; Stahlberg 2020; Kocmi et al. 2022). While LLMs have demonstrated impressive translation performance, they still rely heavily on parallel data—particularly in low-resource settings, where fine-tuning or even pre-training is essential for achieving competitive results (Garcia et al. 2023; Kocmi et al. 2024; Xu et al. 2024; Fu et al. 2024; Luo et al. 2025; Li et al. 2024).

The acquisition of high-quality parallel corpora remains heavily dependent on manual annotation, which poses a significant challenge for low-resource languages due to the scarcity of qualified annotators and the high cost of bilingual labor. In contrast, generating image descriptions requires only monolingual proficiency, making it a more scalable and accessible alternative. Motivated by this, prior work has explored image-pivoted corpus synthesis, where multilingual captions are independently generated for the same image using captioning models, and then used as pseudo-parallel data (Su et al. 2019; Huang et al. 2020; Yang, Fang, and Feng 2022; Fu et al. 2023).

However, this approach suffers from a fundamental limitation: **multilingual focus discrepancy**, where descriptions in different languages tend to focus on distinct aspects or regions of the same image, as illustrated in Figure 1. This inconsistency weakens the alignment between sentences and limits the effectiveness of such pseudo data for direct translation training. To mitigate this, Huang et al. (2020) proposed combining image-pivoted synthesis with iterative multimodal back-translation, and further introduced a visual-semantic embedding (VSE) loss to encourage latent alignment between language pairs.

Different from previous work, we aim to tackle the challenge of multilingual focus discrepancy, enabling better alignment between synthetic parallel sentences. To this

*Corresponding Author.

end, we propose **PRISMS** (**P**arallel **R**efracting **I**mage**S** into **M**ultilingual descriptions with **S**tructured visual guidance), which incorporates structured visual guidance (SVG) to unify the focus when generating captions across different languages. Specifically, given an image, PRISMS first extracts Visual Scene Graphs (VSGs) in a source language (e.g., English) and converts them into textual triplets. These triplets are then incorporated into the instruction prompt, guiding Multimodal Large Language Models (MLLMs) to generate source-language image descriptions grounded in the specified semantic content. Next, PRISMS generates a target-language caption by first translating the VSGs into the target language using unsupervised bilingual word mappings, and then prompting the MLLM to produce a target-language caption based on the translated triplets. Nevertheless, we observe that MLLMs often struggle to adhere to such structured guidance, frequently generating hallucinated content beyond the provided semantic information. To improve grounding fidelity, we explore two strategies: (1) supervised fine-tuning (SFT) with self-constructed instructional data, pairing high-quality captions with their Language Semantic Graphs (LSG), treated as VSG, to form (image, VSG, caption) triplets for training; and (2) reinforcement learning (RL) with a semantic graph consistency reward, where the model is optimized via Direct Preference Optimization (DPO) to favor captions that better align with the input VSG.

The resulting pseudo-parallel descriptions are applied to tasks such as fine-tuning LLMs for translation and unsupervised machine translation. In experiments involving the fine-tuning of LLMs, using these pseudo-parallel corpora to fine-tune models like LLaMA3 demonstrates performance levels comparable to those achieved with high-quality parallel corpora. In the context of unsupervised machine translation, our method achieves state-of-the-art results in image-augmented unsupervised machine translation, highlighting the effectiveness of our approach in leveraging image data to enhance translation performance.

Method

In this section, we provide a comprehensive overview of the method. First, we introduce the framework for generating image descriptions, followed by an explanation of how Pseudo-data Supervised Fine-tuning (PFT) and Iterative Direct Preference Optimization (Iter-DPO) are employed to optimize these descriptions and how cross-linguistic alignment of descriptions is effectively controlled. The overall framework is illustrated in Figure 2.

Image Caption Generation Framework with Structured Visual Guidance

The image description generation framework employed in our method is built upon an MLLM with monolingual support. As depicted in Figure 2, for a given image i in the image dataset $I = \{i_k\}_{k=1}^{K}$, which contains K images, we first utilize an existing visual scene graph extraction tool to generate a scene graph $vsg \in \{vsg_k\}_{k=1}^{K}$. The scene graph vsg is subsequently compressed into triples T_{vsg}, which include structures such as [entity]-[relation]-[entity] and [entity]-[*optional* attribute key]-[attribute value]. All triples T_{vsg} extracted from the image are concatenated together, followed by the addition of an instruction, which is then fed into the MLLM as SVG. The MLLM are required to generate an image description based on these triples. The instruction typically requires the description to strictly adhere to the information found in the triples and the image, without adding extraneous information or omitting any details from the triples. The description is expected to cover all the information contained in the triples in a precise and comprehensive way.

Enhancement of Structured Visual Guidance Adherence

Pseudo-data Supervised Fine-tuning. Commonly used large models (especially those with smaller parameter sizes) face challenges in generating standard-compliant descriptions based on given triplet information. These models often produce verbose or irrelevant content. To further enhance the ability of large models in this regard, additional optimization is necessary. The most straightforward approach would be to fine-tune the large models using VSGs, images, and annotated descriptions. However, due to the lack of annotated data, we propose a method to construct pseudo-data for fine-tuning. We utilize the image caption dataset to construct this pseudo-data. Specifically, for each image-caption pair, we extract the Language Semantic Graph (LSG) from the caption and treat it as a pseudo-VSG. This pseudo-VSG, along with the corresponding image, is input into the MLLMs, with the original caption designated as the ground truth. This approach helps to enhance the models' ability to generate coherent and accurate descriptions.

Iterative Direct Preference Optimization. The descriptions generated by the MLLM can be converted into LSGs. Therefore, to ensure that the generated descriptions align more closely with the SVG, it suffices to maximize the similarity between the VSG and the LSG.

Based on this principle, we designed a reward model. This reward model is tasked with ensuring semantic alignment between the two graphs while guaranteeing that no information is omitted or introduced. For a VSG and an LSG, each is composed of two types of multiple triples T_{vsg} and T_{lsg}:

$$T_{vsg} = \{T_{vsg_m}^{ere}\}_{m=1}^{M} \cup \{T_{vsg_n}^{ekv}\}_{n=1}^{N}, \quad (1)$$

$$T_{lsg} = \{T_{lsg_p}^{ere}\}_{p=1}^{P} \cup \{T_{lsg_q}^{ekv}\}_{q=1}^{Q}. \quad (2)$$

We assume that a VSG consists of N relation-type triples (ERE: [entity]-[relation]-[entity]) and M attribute-type triples (EKV: [entity]-[attribute key]-[attribute value]), while an LSG consists of P relation-type triples and Q attribute-type triples.

To calculate the semantic similarity between the two graphs, we first identify, for each triple in T_{vsg}, the triple of the same type with the most semantical related triple in T_{lsg}. Then, we sum up the similarity scores between the matched

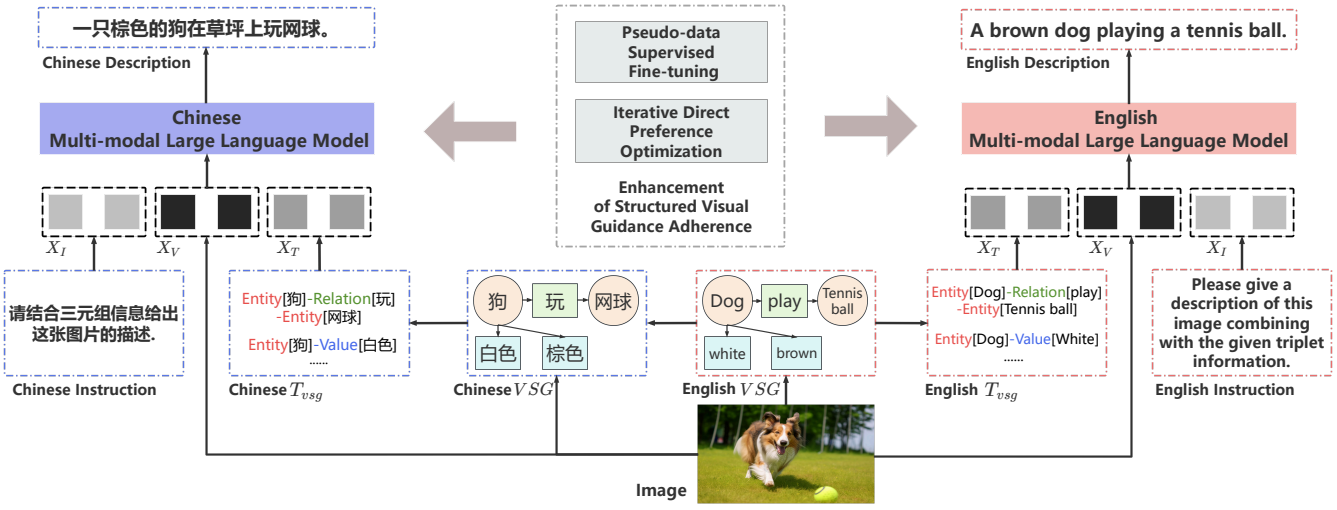

Figure 2: The figure illustrates the process of generating parallel descriptions in both Chinese and English from a given picture. By constructing a VSG, we identify the critical areas of the image that require attention. SFT and RL are employed to enhance the ability of the MLLMs to generate image descriptions that accurately encapsulate the information contained in the triplets. Furthermore, cross-linguistic vocabulary substitution is utilized to ensure consistency in the focus on image details across different languages.

triples to represent the semantic relevance:

$$\begin{aligned}\text{Score}_{\text{sem}} = &\sum_{m=1}^{M} \max_{p=1}^{P} \left(\text{Sim}(T^{\text{ere}}_{vsg_m}, T^{\text{ere}}_{lsg_p})\right) \\ &+ \sum_{n=1}^{N} \max_{q=1}^{Q} \left(\text{Sim}(T^{\text{ekv}}_{vsg_n}, T^{\text{ekv}}_{lsg_q})\right).\end{aligned} \tag{3}$$

The similarity between triples of the same type is determined using a designated similarity function, $Sim()$. To achieve this, we utilize an off-the-shelf sentence-level text similarity computation method[1], ensuring efficiency and semantic precision. Semantic relevance alone is insufficient, as the $\text{Score}_{\text{sem}}$ may still be relatively high even when the VSG and LSG are in a subgraph relationship. Specifically, M should equal P, and N should equal Q. The length constraint is defined as follows:

$$\text{Penalty}_{\text{length}} = 1 - \frac{\min(M,P)}{2 \cdot \max(M,P)} - \frac{\min(N,Q)}{2 \cdot \max(N,Q)}. \tag{4}$$

If the number of two types of triples in the two graphs is completely consistent($M = P$ and $N = Q$), the penalty term is 0. Thus, the final reward score is calculated as a combination of semantic relevance $\text{Score}_{\text{sem}}$ and the penalty term $\text{Penalty}_{\text{length}}$:

$$r(VSG, LSG) = (1 - \text{Penalty}_{\text{length}}) \cdot \text{Score}_{\text{sem}}. \tag{5}$$

We employ Iterative Direct Preference Optimization (Iter-DPO) for optimization, which involves the iterative execution of two core steps: (1) generating synthetic preferences using the given reward model, and (2) fine-tuning the language model based on these synthetic preferences.

Step 1: In each iteration $j \in \{1, 2, 3...\}$, the instruction x is input into the current MLLM checkpoint $\pi_j(y|x)$ to generate multiple candidate descriptions y. The reward model then evaluates these descriptions by assigning reward scores to each. Subsequently, the caption samples with the highest and lowest reward scores, denoted as y_w and y_l respectively, are selected to construct the synthetic preference dataset:

$$D_j = \{(x, y_w, y_l)\}. \tag{6}$$

Step 2: Subsequently, we utilize the generated preference dataset alongside the Direct Preference Optimization (DPO) method to fine-tune the large model. The loss function is defined as:

$$\begin{aligned}L_{DPO} = -\mathbb{E}_{(x,y_w,y_l)\sim D}[log\sigma(\beta log \frac{\pi_{\theta_j}(y_w|x)}{\pi_{\theta_{j-1}}(y_w|x)}) - \\ \beta log \frac{\pi_{\theta_j}(y_l|x)}{\pi_{\theta_{j-1}}(y_l|x)})].\end{aligned} \tag{7}$$

Avoiding Multilingual Focus Discrepancy

Structural Consistency of VSGs across Different Languages. The VSGs generated by visual models across different languages are often structurally inconsistent. To ensure the consistency among VSGs in different languages, we first utilize a single English scene graph generation tool to produce an English VSG and for VSGs in other languages, then perform word substitution using a bilingual lexicon obtained through unsupervised word alignment methods[2]. This approach guarantees complete consistency in graph structure. The unsupervised lexicon is trained without relying on

[1]https://github.com/UKPLab/sentence-transformers

[2]https://github.com/facebookresearch/MUSE

Testset	Flores200		Newstest2013		Newstest2012		Newstest2011		Average	
Metric	**COMET**	**BLEU**	**COMET**	**BLEU**	**COMET**	**BLEU**	**COMET**	**BLEU**	**COMET**	**BLEU**
					EN-FR					
Zero-shot	81.85	31.19	79.51	23.64	78.25	22.41	78.64	23.96	79.56	25.30
FT-GT	**85.12**	**39.14**	**82.43**	**29.31**	**81.09**	**27.88**	**81.32**	**29.91**	**82.49**	**31.56**
FT-PRISMS	84.45	37.06	81.89	28.35	80.48	26.86	80.86	28.83	81.92	30.28
Δ	0.67/2.60	2.08/5.87	0.54/2.38	0.96/4.71	0.61/2.23	1.02/4.45	0.46/2.22	1.08/4.87	0.57/2.36	1.28/4.98
					EN-DE					
Zero-shot	78.35	20.14	77.10	15.96	74.33	13.14	73.62	13.15	75.85	15.60
FT-GT	**84.17**	**27.58**	**82.04**	**21.69**	**79.96**	**18.27**	**79.60**	**18.24**	**81.44**	**21.45**
FT-PRISMS	82.50	25.42	81.12	20.72	78.85	17.64	78.49	17.84	80.24	20.41
Δ	1.67/4.15	2.16/5.28	0.92/4.02	0.97/4.76	1.11/4.52	0.63/4.50	1.11/4.87	0.40/4.69	1.20/4.39	1.04/4.81
					FR-EN					
Zero-shot	87.82	37.39	84.42	30.36	82.49	29.42	82.34	29.77	84.27	31.74
FT-GT	**88.35**	41.32	**84.75**	**32.35**	**82.90**	**31.41**	**82.64**	**31.05**	**84.66**	**34.03**
FT-PRISMS	88.28	**41.41**	84.45	32.09	82.60	30.80	82.37	30.78	84.43	33.77
Δ	0.07/0.46	-0.09/4.02	0.30/0.03	0.26/1.73	0.30/0.11	0.61/1.38	0.27/0.03	0.27/1.01	0.23/0.16	0.26/2.03
					DE-EN					
Zero-shot	87.61	36.02	84.05	27.26	82.28	25.56	81.68	23.86	83.91	28.18
FT-GT	**88.19**	**39.39**	**84.63**	**29.60**	**83.01**	**27.13**	**82.06**	**24.85**	**84.47**	**30.24**
FT-PRISMS	87.79	38.36	84.15	29.25	82.62	26.47	81.69	24.60	84.06	29.67
Δ	0.40/0.18	1.03/2.34	0.48/0.10	0.35/1.99	0.39/0.34	0.66/0.91	0.37/0.01	0.25/0.74	0.41/0.16	0.57/1.50

Table 1: Results on LLaMa 3 under zero-shot conditions, as well as after fine-tuning with authentic parallel corpora and our generated pseudo-corpora. In the delta row, the first value represents the negative difference between our method and the performance achieved using real parallel corpora. The second value represents the improvement of our method compared to the Zero-shot approach.

parallel corpora, and its performance often matches or even surpasses that of supervised lexicons. During word substitution, polysemy frequently arises. To address this, we employ image information to assist in disambiguation, leveraging MLLMs in the target language. Specifically, we concatenate multiple candidate words for the target language and issue an instruction such as: "Please select the most suitable word for this image from the list of candidates" (translated into the target language for practical application; English is used here for clarity).

Post-processing. Descriptions generated by different MLLMs based on structurally consistent VSGs can generally be regarded as parallel corpora. However, cases of omission or addition of triplet information may still occur. Therefore, a post-processing module is required after generating the descriptions. During post-processing, we convert the descriptions in both languages into triplets and then compare the number of triplets in two categories (ERE and EKV). For descriptions with inconsistent triplet counts, we exclude them from further processing.

Experiments

In this section, we present the datasets utilized, the experimental training settings for generating parallel corpora, the details of the validation experiments, and the corresponding experimental results.

Settings

Datasets. The training dataset we employed is the Multi30k dataset, which comprises a total of 29,000 images sourced from the Flickr30k dataset. Of these, 14,500 images are utilized to enhance the capability of the MLLMs to generate descriptions that accurately encapsulate the information contained in the triplets. The remaining 14,500 images are employed to generate synthetic parallel corpora for validation experiments.

LLMs. The primary MLLMs used for generating image descriptions include the Qwen-VL[3] model and the LLava1.6 model[4]. Specifically, Qwen-VL is utilized to generate English and Chinese descriptions, while LLava1.6 is used for German and French descriptions. The required LLMs only need multimodal capabilities in a single language. Although we selected a few models that support multiple languages (due to the scarcity of purely monolingual models for these languages), during the description generation process, we only leveraged their monolingual capabilities.

Metric. We employed BLEU(Papineni et al. 2002) and COMET(Rei et al. 2020) metrics to evaluate the effectiveness of our translation experiments.

[3]https://github.com/QwenLM/Qwen-VL

[4]https://github.com/haotian-liu/LLaVA

Testset	Flores200		Newstest2012	
Metric	**COMET**	**BLEU**	**COMET**	**BLEU**
FT-ORI	74.09	22.97	71.80	17.21
FT-SVG	35.71	2.34	37.64	3.11
FT-SVG-QE	70.88	24.24	69.59	19.76
FT-PRISMS	**84.45**	**37.06**	**80.48**	**26.86**

Table 2: The EN-FR results on Flores200 and Newstest2012 compared to original generated descriptions and selected descriptions.

VSG and LSG Extractors. For the generation of VSG, we adopted existing methods (Tang et al. 2019; Tang 2020; Tang et al. 2020)[5]. It is important to emphasize that we do not prioritize the complete identification of entity relations and attributes during VSG generation. The obtained VSG typically lacks attributes such as color information; therefore, we randomly incorporated such attributes using image caption data to enhance its richness. Our primary concern is ensuring that the information emphasized in both languages is consistent. It is entirely acceptable if both languages focus on only a subset of the information present in the image. As for the generation of LSG, we utilized the SNG-GRAPH[6], which performs well in English. For other languages, we leveraged LLMs through few-shot learning or simple fine-tuning to achieve satisfactory results.

Main Results

Compared to Real Parallel Corpus. We first fine-tuned the LLM using the generated pseudo-parallel corpus (noted "**FT-PRISMS**") and then evaluated its performance on several translation test sets. The following methods were used for comparison:

- Zero-shot performance: Directly prompting the large language model to perform translation without any fine-tuning.
- Fine-tuning with human-annotated parallel corpora noted "**FT-GT**": Using manually labeled data for fine-tuning, followed by testing the translation performance.

The LLM we selected is LLaMA3[7], and the language pairs evaluated are English-French and English-German. The test sets used for evaluation are WMT newstest and Flores200.

The results of fine-tuning LLaMA3-8B with the generated parallel corpus for English-French and English-German translations are shown in Table 1. The results of our method presented in the table are based on fine-tuning with pseudo-data. Results derived from Iter-DPO are comparable with those obtained through PFT, and the relevant discussion is provided in the analysis experiments section. We observed that fine-tuning with both real and synthetic parallel corpora leads to substantial improvements over directly using the base model. The improvement of our method over the base model is still significant, with English-to-French translation on the Flores200 test set achieving an increase of over 5 BLEU points, and an average improvement of over 2 BLEU points and 1.5 COMET points across all tested languages. This demonstrates that the quality of the parallel corpus produced by our method is comparable to that of real data and that it effectively supports the fine-tuning of large models for machine translation.

[5]https://github.com/KaihuaTang/Scene-Graph-Benchmark.pytorch

[6]https://github.com/vacancy/SceneGraphParser/tree/master

[7]https://github.com/meta-llama/llama3

Testset	F2016	F2017	C2017	F2018
UNMT	32.55	-	-	-
PVP	40.95	-	-	-
XLM	42.61	37.43	32.17	34.40
UNMT-CS	43.95	37.79	32.84	35.06
UNMT-WVR	44.14	38.57	33.42	35.89
UNMT-PRISMS	**45.62**	**41.58**	**35.07**	**37.02**
Δ	1.48	3.01	1.65	1.13

Table 3: EN-FR, FR-EN, DE-EN and EN-DE's average BLEU results on unsupervised machine translation task on Flickr2016, Flickr2017, COCO2017 and Flickr2018 test set.

When comparing our method to the approach using real parallel corpora, we found that our method generally performs slightly lower. However, the performance is still quite close, and in some language directions (e.g., French-to-English translation on the Flores200 test set), our method even slightly surpasses the method using real parallel corpora. This indicates that the quality of the parallel corpus generated by our method is acceptable. Additionally, considering the high cost of obtaining real parallel corpora, our method demonstrates a significant advantage in terms of cost-efficiency since it only requires the mapping between a single language and images. This further highlights the low-cost nature of our approach. We also observed that the improvement of our method in EN-to-X directions is more pronounced compared to X-to-EN. This might be because the LLaMA3 model is inherently more proficient in English than in other languages, resulting in better performance when generating English. Therefore, while fine-tuning with parallel corpora improves the model for X-to-EN translation, the improvement is not as substantial as that for EN-to-X.

Furthermore, We conducted tests on pairs of non-English language and low-resource language, using the Flores200 dataset as a test set. The results are shown in Figure 3. We found that the conclusions remain consistent: the pseudo-parallel corpora we constructed achieve improvements compared to Zero-shot methods and approach the performance of using real parallel corpora.

Comparison with Original Image-Pivot Parallel Corpus Synthesis. We compared our method against the original image-pivot parallel corpus synthesis approach, referred to as **ORI**. Additionally, we reported results using a vanilla MLLM to generate captions based solely on SVGs, without any SVG adherence enhancement. This variant is referred to as **SVG**. Descriptions generated via the SVG-only approach

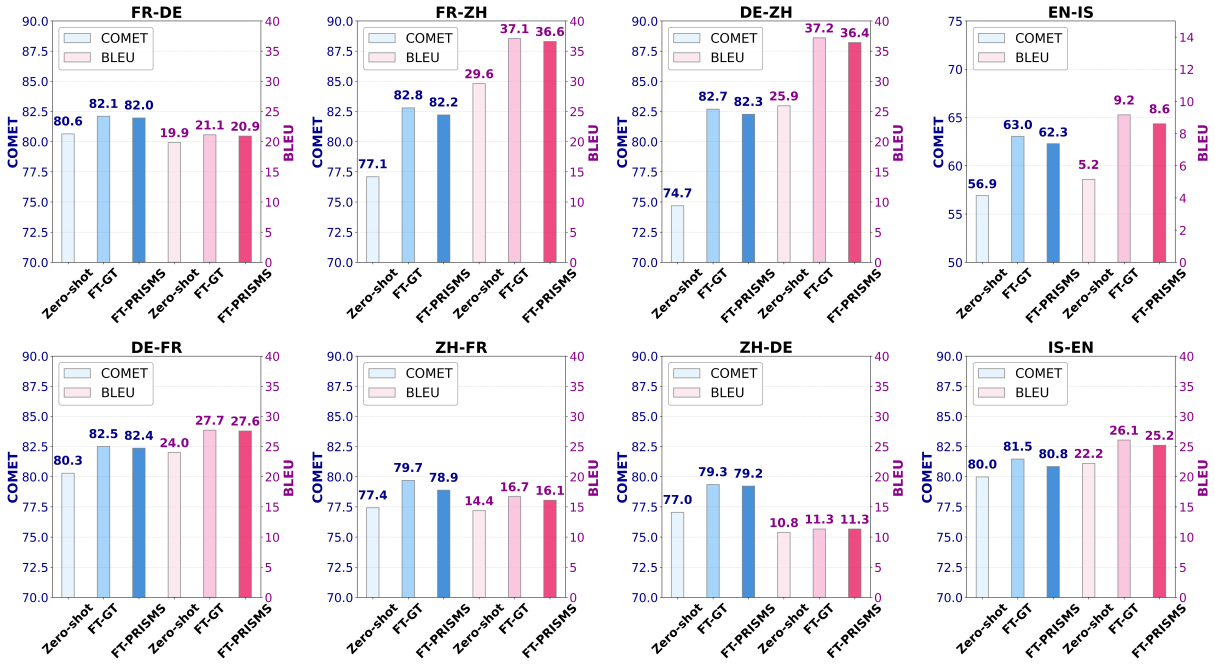

Figure 3: The results on Flores200 for non-English language pairs and low-resource language directions.

and subsequently filtered using QE[8] are denoted as **SVG-QE**. All resulting descriptions are treated as parallel corpora and used to fine-tune LLaMA3.

EN-FR translation results on the Flores200 and Newstest2012 datasets are summarized in Table 2. As shown, directly using MLLMs to generate descriptions is insufficient for constructing effective parallel corpora. Fine-tuning large models on such data leads to a substantial drop in translation quality compared to the zero-shot baseline.

Moreover, directly injecting SVG information into the prompt leads to a sharp performance drop, as the base models initially lack the ability to interpret SVG content, resulting in low-quality descriptions. However, after additional PFT or Iter-DPO training, MLLMs acquire the capability to process SVGs effectively. The resulting captions in different languages focus on the same visual content, allowing them to serve as valid parallel corpora and produce results comparable to those achieved with real parallel data.

Analysis

Unsupervised Machine Translation

Our method of using images to generate parallel corpora can be viewed as a form of unsupervised machine translation (UNMT). Based on this, we further utilized the pseudo-parallel corpora to train an unsupervised machine translation model based on the XLM framework. Specifically, we first fine-tuned the XLM model using the generated pseudo-parallel corpora. Then, we applied iterative back-translation and denoising autoencoder training on the fine-tuned model, which are standard steps in the UNMT training process.

[8]https://github.com/Unbabel/COMET

We compared our method with the following existing UNMT approaches:

- **XLM** (Conneau and Lample 2019): using a masked language model to train a cross-lingual language model for initialization.
- **UMMT** (Su et al. 2019): incorporating visual features for denoising autoencoder and back-translation.
- **PVP** (Huang et al. 2020): employing multimodal back-translation and introduces pseudo visual pivoting.
- **UNMT-CS** (Jones et al. 2023): utilizing a code-switching method, where words in the source sentence are replaced with their corresponding translations from bilingual dictionaries, to address lexical confusion.
- **UNMT-WVR** (Fu et al. 2023): leveraging word-level images added as visual matrices during training to enhance the UNMT process.

We evaluated our method and the above approaches on EN-FR and EN-DE on the Multi30k test set, which includes Flickr2016, Flickr2017, COCO2017, and Flickr2018. Averaged results are presented in Table 3 and detailed results can be found in the Supplementary Materials.

From the results in Table 3, we can observe that our method achieves impressive results on all test datasets for both EN-FR and EN-DE directions, outperforming previous UNMT approaches that leverage images. In some directions, our method achieves an improvement of over 3 BLEU points, indicating that the quality of our generated pseudo-parallel corpora is relatively high and provides a strong initialization for UNMT systems.

Language	EN-ZH		ZH-EN	
Metric	**COMET**	**BLEU**	**COMET**	**BLEU**
Zero-shot	69.69	17.69	84.17	21.11
FT-PFT	**84.46**	**41.80**	84.51	23.42
FT-IterDPO	84.34	41.71	**84.73**	**23.90**

Table 4: Comparison of PFT and Iter-DPO on Flores200.

Pseudo-data Supervised Fine-tuning vs. Iterative Direct Preference Optimization

In the main experiments, we primarily report the outcomes of PFT. Here, we further analyze the performance of Iter-DPO on the 7B model. As presented in Table 4, we observed that, based on the medium-scale parameter MLLM, Iter-DPO achieved results comparable to PFT, with some languages showing improvements while others exhibited only marginal gains.

This phenomenon might be attributed to the fact that Iter-DPO generates samples using the model itself, but the medium-parameter model's capacity to produce accurate descriptions guided by SVG remains relatively limited. Moreover, it is worth noting that the PFT method relies on image description datasets for constructing pseudo-data. In scenarios where such datasets are unavailable, Iter-DPO serves as a viable alternative optimization method.

Language Semantic Graph Generation Based on Large Language Models

For LSG generation, we used the SNG-GRAPH method for English, while for other languages, LLMs were guided via prompt learning to produce LSG. Five manually annotated LSG examples were provided as references for few-shot learning. To evaluate the LLM's performance, 100 text samples from the MULTI30k 2016 test set were manually annotated for LSG in both English and Chinese, with the latter derived from Google-translated[9] English results. We conducted tests using Qwen-3[10], and the results are presented in Figure 4. For the annotated results, we calculated the precision values, recall values, and F1-scores. During this process, we introduced a similarity threshold value. If the similarity between the generated triples and the annotated triples exceeded this threshold, we considered them as matched, deeming the model-generated results acceptable. The similarity was computed using Sentence-BERT. From the table, we observed that the LLM's annotation results closely align with the annotations for both English and Chinese, indicating that the LLM performs effectively in generating LSG.

Ablation on Post-processing Module

We also conducted ablation experiments on the post-processing module. Specifically, we fine-tuned the LLaMA3 model using parallel corpora without post-processing. The

[9]https://translate.google.com

[10]https://github.com/QwenLM/Qwen3

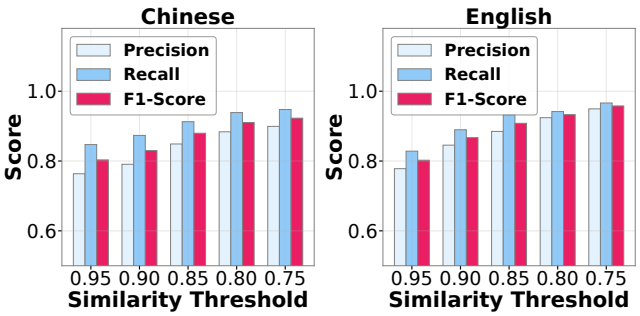

Figure 4: The Precision-Score, Recall-Score, and F1-scores for both Chinese and English when generating LSGs using LLMs.

Metric	COMET	BLEU	COMET	BLEU
Testset	**Flores200**		**Newstest2013**	
FT-PRISMS	**84.45**	**37.06**	**81.89**	**28.35**
/w.o. post	83.32	35.01	81.45	27.78
Testset	**Newstest2012**		**Newstest2011**	
FT-PRISMS	**80.48**	**26.86**	**80.86**	**28.83**
/w.o. post	80.05	26.12	80.19	28.22

Table 5: Fine-tuning the results of LLaMA3 using parallel corpora obtained after removing the post-processing module on EN-FR direction.

results on four test sets across two language directions (en-fr, en-de) are shown in Table 5. We found that in all test sets, the results without post-processing were consistently lower than those with post-processing, with a maximum drop of nearly two BLEU points observed on the Flores test set for en-fr. This indicates that the post-processing method plays a significant role in filtering out misaligned parallel corpora. However, we also observed that, despite the performance drop, the results without post-processing still outperformed the Zero-shot results in Table 1. This demonstrates that our pseudo-parallel corpus generation method, even without post-processing, achieves reasonably good alignment.

Conclusion and Future Work

In this work, we propose PRISMS, a vision-grounded framework for synthesizing multilingual parallel corpora by leveraging structured image semantics. Specifically, we use the sense graph as the SVG to ensure that descriptions generated in different languages focus on the same elements in the image, thereby avoiding multilingual focus discrepancy. PFT and Iter-DPO are then applied to optimize MLLMs, compelling them to produce semantically aligned descriptions across languages. In the future, we plan to further explore the integration between the multimodal and multilingual capabilities of LLMs, including research on the mechanisms of mutual enhancement between these abilities.

Acknowledgments

Xiaocheng Feng is the corresponding author of this work. We thank the anonymous reviewers for their insightful comments. This work was supported by the National Natural Science Foundation of China (NSFC) (grant 6252200908, 62276078, U22B2059), the Fundamental Research Funds for the Central Universities (XNJKKGYDJ2024013), the State Key Laboratory of Micro-Spacecraft Rapid Design and Intelligent Cluster (MS01240122), the Major Key Project of PCL (Grant No. PCL2025A12, PCL2025A03), and the National Science and Technology Major Program (Grant No. 2024ZD01NL00101).

References

Brown, P. F.; Cocke, J.; Della Pietra, S. A.; Della Pietra, V. J.; Jelinek, F.; Lafferty, J.; Mercer, R. L.; and Roossin, P. S. 1990. A statistical approach to machine translation. *Computational linguistics*, 16(2): 79–85.

Brown, P. F.; Della Pietra, S. A.; Della Pietra, V. J.; and Mercer, R. L. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2): 263–311.

Castilho, S.; Moorkens, J.; Gaspari, F.; Calixto, I.; Tinsley, J.; and Way, A. 2017. Is neural machine translation the new state of the art? *The Prague Bulletin of Mathematical Linguistics*, (108).

Conneau, A.; and Lample, G. 2019. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32.

Fu, C.; Feng, X.; Huang, Y.; Huo, W.; Li, B.; Wang, H.; Qin, B.; and Liu, T. 2024. Relay Decoding: Concatenating Large Language Models for Machine Translation. *arXiv preprint arXiv:2405.02933*.

Fu, C.; Feng, X.; Huang, Y.; Huo, W.; Wang, H.; Qin, B.; and Liu, T. 2023. Enabling Unsupervised Neural Machine Translation with Word-level Visual Representations. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 12608–12618. Singapore: Association for Computational Linguistics.

Garcia, X.; Bansal, Y.; Cherry, C.; Foster, G.; Krikun, M.; Johnson, M.; and Firat, O. 2023. The unreasonable effectiveness of few-shot learning for machine translation. In *International Conference on Machine Learning*, 10867–10878. PMLR.

Huang, P.-Y.; Hu, J.; Chang, X.; and Hauptmann, A. 2020. Unsupervised Multimodal Neural Machine Translation with Pseudo Visual Pivoting. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8226–8237. Online: Association for Computational Linguistics.

Jones, A.; Caswell, I.; Saxena, I.; and Firat, O. 2023. Bilex Rx: Lexical Data Augmentation for Massively Multilingual Machine Translation. *CoRR*, abs/2303.15265.

Kocmi, T.; Avramidis, E.; Bawden, R.; Bojar, O.; Dvorkovich, A.; Federmann, C.; Fishel, M.; Freitag, M.; Gowda, T.; Grundkiewicz, R.; et al. 2024. Findings of the WMT24 general machine translation shared task: the LLM era is here but mt is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, 1–46.

Kocmi, T.; Bawden, R.; Bojar, O.; Dvorkovich, A.; Federmann, C.; Fishel, M.; Gowda, T.; Graham, Y.; Grundkiewicz, R.; Haddow, B.; Knowles, R.; Koehn, P.; Monz, C.; Morishita, M.; Nagata, M.; Nakazawa, T.; Novák, M.; Popel, M.; and Popović, M. 2022. Findings of the 2022 Conference on Machine Translation (WMT22). In Koehn, P.; Barrault, L.; Bojar, O.; Bougares, F.; Chatterjee, R.; Costa-jussà, M. R.; Federmann, C.; Fishel, M.; Fraser, A.; Freitag, M.; Graham, Y.; Grundkiewicz, R.; Guzman, P.; Haddow, B.; Huck, M.; Jimeno Yepes, A.; Kocmi, T.; Martins, A.; Morishita, M.; Monz, C.; Nagata, M.; Nakazawa, T.; Negri, M.; Névéol, A.; Neves, M.; Popel, M.; Turchi, M.; and Zampieri, M., eds., *Proceedings of the Seventh Conference on Machine Translation (WMT)*, 1–45. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics.

Koehn, P.; Och, F. J.; and Marcu, D. 2003. Statistical phrase-based translation. In *2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Langauge Technology (HLT-NAACL 2003)*, 48–54. Association for Computational Linguistics.

Li, J.; Zhou, H.; Huang, S.; Cheng, S.; and Chen, J. 2024. Eliciting the Translation Ability of Large Language Models via Multilintual Finetuning with Translation Instructions. *Transactions of the Association for Computational Linguistics*, 12.

Lopez, A. 2008. Statistical machine translation. *ACM Computing Surveys (CSUR)*, 40(3): 1–49.

Luo, Y.; Zheng, T.; Mu, Y.; Li, B.; Zhang, Q.; Gao, Y.; Xu, Z.; Feng, P.; Liu, X.; Xiao, T.; et al. 2025. Beyond Decoder-only: Large Language Models Can be Good Encoders for Machine Translation. *arXiv preprint arXiv:2503.06594*.

Och, F. J.; and Ney, H. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the 40th Annual meeting of the Association for Computational Linguistics*, 295–302.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Rei, R.; Stewart, C.; Farinha, A. C.; and Lavie, A. 2020. COMET: A Neural Framework for MT Evaluation. In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2685–2702. Online: Association for Computational Linguistics.

Stahlberg, F. 2020. Neural machine translation: A review. *Journal of Artificial Intelligence Research*, 69: 343–418.

Su, Y.; Fan, K.; Bach, N.; Kuo, C.-C. J.; and Huang, F. 2019. Unsupervised multi-modal neural machine translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10482–10491.

Tang, K. 2020. A Scene Graph Generation Codebase in PyTorch. https://github.com/KaihuaTang/Scene-Graph-Benchmark.pytorch.

Tang, K.; Niu, Y.; Huang, J.; Shi, J.; and Zhang, H. 2020. Unbiased Scene Graph Generation from Biased Training. In *Conference on Computer Vision and Pattern Recognition*.

Tang, K.; Zhang, H.; Wu, B.; Luo, W.; and Liu, W. 2019. Learning to Compose Dynamic Tree Structures for Visual Contexts. In *Conference on Computer Vision and Pattern Recognition*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Xu, H.; Sharaf, A.; Chen, Y.; Tan, W.; Shen, L.; Van Durme, B.; Murray, K.; and Kim, Y. J. 2024. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. *arXiv preprint arXiv:2401.08417*.

Yang, Z.; Fang, Q.; and Feng, Y. 2022. Low-resource Neural Machine Translation with Cross-modal Alignment. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 10134–10146. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.