Make Imagination Clearer! Stable Diffusion-based Visual Imagination for Multimodal Machine Translation

Anonymous ACL submission

Abstract

Visual information has been introduced for enhancing machine translation (MT), and its effectiveness heavily relies on the availability of large amounts of bilingual parallel sentence pairs with manual image annotations. In this paper, we introduce a stable diffusion-based imagination network into a multimodal large language model (MLLM) to explicitly generate an image for each source sentence, thereby advancing the multimodel MT. Particularly, we build heuristic human feedback with reinforcement learning to ensure the consistency of the generated image with the source sentence without the supervision of image annotation, which breaks the bottleneck of using visual information in MT. Furthermore, the proposed method enables imaginative visual information to be integrated into large-scale text-only MT in addition to multimodal MT. Experimental results show that our model significantly outperforms existing multimodal MT and text-only MT, especially achieving an average improvement of more than 14 BLEU points on Multi30K and MSCOCO multimodal MT benchmarks.

1 Introduction

Large Language Models (LLMs) have recently demonstrated exceptional comprehension and generation abilities across a wide range of tasks, particularly in translation (Tyen et al., 2023; Liang et al., 2023; Guerreiro et al., 2023; Ranaldi et al., 2023; Zhang et al., 2024; Chen et al., 2024b,a; Chu et al., 2023). LLM-based machine translation (LLM-MT) methods generally map the source text directly to the target text (Hendy et al., 2023; Jiao et al., 2023; Le Scao et al., 2023; Iyer et al., 2023; Zeng et al., 2023; Zhao et al., 2024), while professional human translators often imagine visual information when translating source texts (Hubscher-Davidson, 2020; Bang, 1986; Long et al., 2021; Elliott and Kádár, 2017). The process of imagining involves creating scenes, relationships between objects, and commonsense details within the translation text. Therefore, generating such content is crucial for ensuring high-quality translation as it helps capture subtle nuances accurately (Yao and Wan, 2020; Lin et al., 2020; Sigurdsson et al., 2020; Song et al., 2022). Although multiple previous works in multimodal machine translation have attempted similar approaches(Long et al., 2021; Elliott and Kádár, 2017; Hitschler et al., 2016), they still face limitations such as insufficient model capacity, the requirement for image-text annotated training data, and poor quality of generated images.

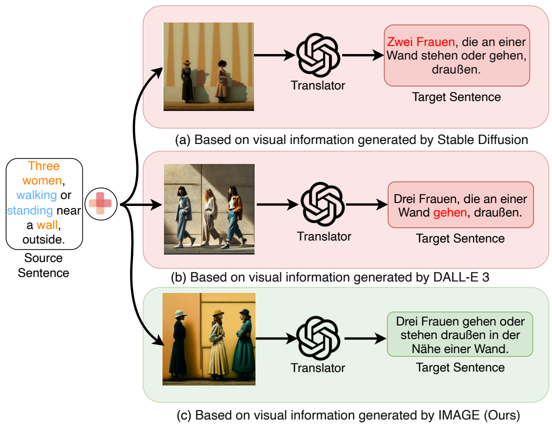

Figure 1: Illustration of the LLMs translation paradigm based on visual information. Figure a: The generated image does not include information about three women, and Figure b: The generated image lacks standing information. These issues led to the translation error.

To address these issues, we propose a framework called **IMAGE**, which stands for **I**magination-Based End-to-End **M**ultimodal L**A**rge Langua**G**e Mod**E**l Machine Translation Framework. IMAGE first generates corresponding visual information (image) from the source text, and then uses both source text and visual information to produce translation results through LLM. Current mainstream visual information generation methods (such as diffusion models (Du et al., 2023; Tang et al., 2023; Liu and Liu, 2024; Liu et al., 2024)) often struggle to generate complex scenes based on language descriptions, impacting translation performance, as shown in Figures 1(a) and (b). To ensure that the generated visual information accurately represents the source text, we heuristically build a supervisory signal based on human feedback to enhance the consistency of generated visual content with the source sentence, further improving translation performance, as illustrated in Figure 1(c).

Our framework was evaluated on the standard Multimodal Machine Translation (MMT) dataset Multi30K and the general Neural Machine Translation (NMT) dataset WMT24. Extensive experimental results confirm that the IMAGE framework based on visual imagination outperforms text-only LLM approaches. Additionally, through ablation experiments, we verified the necessity of each component in the IMAGE framework. Furthermore, analysis experiments and case studies reveal a positive correlation between the consistency of visual imagination with the text and translation performance. In summary, our contributions are as follows:

- We are the first to propose an end-to-end multimodal machine translation framework leveraging the visual imagination capabilities of LLMs. Our goal is to inspire the translation community to further integrate LLMs and multimodality into future translation research.
- Our framework uses human feedback RL during training, eliminating the need for annotated image-text data and reducing annotation costs.
- Our model demonstrates significant performance improvements on general and multimodal translation benchmarks compared to traditional multimodal translation methods and text-only LLM-MT.

2 Background

2.1 Multimodal Large Language Model

Currently, multimodal large models consist of three main components: a Large Language Model (LLM), an image encoder, and a projector. The **LLM** is responsible for modeling the joint probability distribution $p_\theta(\mathbf{w})$ of a sequence $\mathbf{w} = \{\mathbf{w}_t\}_{t=1}^{T}$, where T is the sequence length and θ represents the model parameters. The generation process of each token $\mathbf{w}_t$ in the LLM is modeled :

$$p_\theta(\mathbf{w}) = \prod_{t=1}^{T} p_\theta\left(\mathbf{w}_t \mid \mathbf{w}_{<t}\right). \quad (1)$$

For the image encoder, the input sequence contains K ordered images $\boldsymbol{I} = \{I_k\}_{k=1}^{K}$. Each image I_k is processed through a vision encoder, such as a CLIP-like encoder $\mathcal{E}_\phi(\cdot)$, which generates patch embeddings to obtain the image representation signals. These representations are then encoded by the projector $\mathcal{P}_\zeta$ (such as a linear layer), as described by Alayrac et al., 2022 into visual embeddings $\boldsymbol{V}_k = \{\mathbf{v}_\ell\}_{\ell=1}^{L}$ of length L.

Here, $K(t)$ refers to the image index used before generating the t-th word token. Maximum likelihood estimation (MLE) aims to minimize the models loss function to optimize the parameters θ, ϕ, and ζ, thereby aligning the generated sequence as closely as possible with the given data. The loss function is written as:

$$\mathcal{L}_{\mathrm{MLLM}}(\Theta, \mathbf{w}, \boldsymbol{I}) := -\mathbb{E}_t\left[\log p_\Theta\left(\mathbf{w}_t \mid \mathbf{w}_{<t}, \boldsymbol{V}_{<K(t)}\right)\right], \quad (2)$$

$$\boldsymbol{V}_{K(t)} = \mathcal{P}_\zeta \circ \mathcal{E}_\phi\left(I_{K(t)}\right). \quad (3)$$

2.2 Scene Graph Representation

In MMT, the data availability is represented as $< x, z > \in < X, Z >$, where X denotes the source-side sentences and Z represents the paired visual images. Scene Graph represents the semantic relationships between objects in text (LSG) or visual (VSG) information. We define the LSG and VSG as $LSG = (N_L, E_L)$ and $VSG = (N_V, E_V)$. The set N_L and N_V represent the entity nodes in sentences or visual images, respectively, inculding the head entity (h^l and h^v) and the tail entity (t^l and t^v), where $l \in L$ and $v \in V$. The sets E_L and E_V represent the relations (r^l and r^v) connecting these nodes in N_L and N_V.

2.3 Diffusion Models

Diffusion models (DMs) are probabilistic generative models that learn the latent structure of data $\mathbf{x} = \{\mathbf{x}_t\}_{t=1}^{T}$ through continuous-T-timestamps information diffusion. DMs gradually add Gaussian noise to an image x_0 until attaining $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. This noise injection process (the

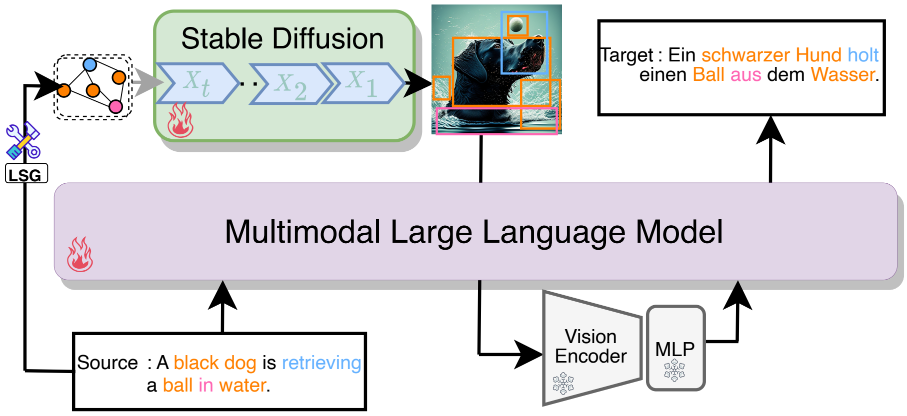

Figure 2: **Overview of our IMAGE framework.** The process involves first generating visual information of the translation input sentence using a diffusion model. Next, the translation result is obtained via LLM, informed by the generated visual information and translation of the original input sentence.

forward process) is formalized as Markov chain $q(\mathbf{x}_{1:T} \mid \mathbf{x}_0, c) = \prod_{t=1}^{T} q(\mathbf{x}_t \mid \mathbf{x}_{t-1}, c)$, where c where x_0 is the sample dataset and c is the corresponding context. The forward process is written as.

$$q(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}\left(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}\right), \quad (4)$$

where $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$. And $\mathbf{x}_t = \sqrt{\bar{\alpha}}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_t$, where $\epsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Reversing the forward process can be accomplished by training a neural network $\mu_\theta(x_t, c, t)$ with the following objective:

$$L_{DDPM}(\theta) = \mathbb{E}_{(x_0,c)\sim p, t\sim U\{0,T\}, x_t\sim q}\left[\|\tilde{\mu}(x_0,t) - \mu_\theta(x_t,c,t)\|^2\right], \quad (5)$$

where $\tilde{\mu}$ is the posterior mean of the forward process, a weighted average of x_0 and x_t. This objective is justified as maximizing a variational lower bound on the log-likelihood of the data (Ho et al., 2020).

3 Proposed Framework: IMAGE

3.1 Framework Overview

Our proposed framework, IMAGE, incorporates visual signals to enhance the performance of large models in multilingual translation tasks. Additionally, to ensure that the entity relationships within the generated visual information remain consistent with input sentences, we adopt an alignment human feedback learning approach.

Figure 2 provides an overview of IMAGE. The following subsections detail three key components: the end-to-end multimodal machine translation framework (Section 3.2), alignment human feedback learning (Section 3.3), and the model training process (Section 3.4).

3.2 End-to-End Multimodal Machine Translation Framework

IMAGE is built upon a causal decoder architecture LLM p_θ, such as Vicuna (Chiang et al., 2023). IMAGE adopts OpenAIs CLIP-Large (Radford et al., 2021) as the visual encoder $\mathcal{E}_\phi(\cdot)$, followed by a linear layer $\mathcal{P}_\zeta$ for visual embedding projection (Dong et al., 2024). To generate images, we utilize Stable Diffusion (SD) (Rombach et al., 2022) as the image decoder, with the condition projector also implemented as a linear layer. Figure 2 provides an overview of this architecture.

3.3 Alignment Human Feedback Learning

The Alignment Feedback Learning aims to enhance the quality of images generated by diffusion model through alignment between linguistic and visual information. This method comprises two core parts: **Design Reward Function** and **Alignment Optimization For Diffusion Model**.

3.3.1 Design Reward Function

To ensure consistency between the translated source sentence and the generated image, the entities and relations in the image need to match those in the source sentence as closely as possible. Based on this, we design a reward function to assess the consistency of the generated image (VSG) to the source sentence (LSG). As shown in Figure 3, the closer LSG is to VSG, the higher the consistency between the translated source sentence and the generated image. We constructed the reward function to evaluate the consistency of LSG and VSG, with a reward scoring range from 0 to 1. Since human judgments of the consistency between images and descriptive texts are also based on the analysis of entities and their relationships, this task constitutes reinforcement learning from human feedback (Ouyang et al., 2022; Christiano et al., 2017; Ziegler et al., 2019).

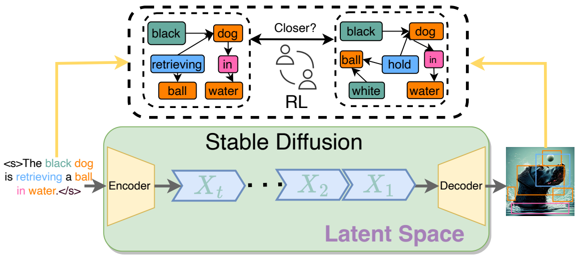

Figure 3: **RL Training Detail.** The overview of IMAGE, which leverages an alignment feedback learning framework to comprehensively enhance the visual signals performance.

For LSG and VSG generation, we utilize two off-the-shelf SG parsers to obtain LSG and VSG separately (as detailed in §A.2). Due to the differing number of triples in LSG and VSG, we designed a structured similarity calculation method to measure their consistency. For each triple in LSG, we calculate its similarity with each triple in VSG and select the highest score as the matching degree for that triple:

$$Score\left(\mathrm{LSG}_i, VSG\right) = \max\left(Sim\left(\mathrm{LSG}_i, \mathrm{VSG}_1\right), \cdots, Sim\left(\mathrm{LSG}_i, \mathrm{VSG}_n\right)\right), \tag{6}$$

$$\mathrm{Sim}(LSG_1, VSG_1) = \frac{SIM(h_1^l, h_1^v) + SIM(r_1^l, r_1^v) + SIM(t_1^l, t_1^v)}{3}, \tag{7}$$

where n represents number of VSG sets, h^l and h^v are the head entities, t^l and t^v are the tail entities, r^l and r^v are the relations, and SIM is off-the-shelf similarity of text model (as detailed in §A.2). Finally, the consistency reward score between sentences and images is the average score of all text triples:

$$r(x_0, c) = \frac{1}{N} \sum_{i=1}^{N} Score(LSG_i, VSG), \tag{8}$$

where c and x_0 denote the LSG of the source sentence and generated image, respectively.

3.3.2 Alignment Optimization For Diffusion Model

We assume a pre-existing diffusion model, which may be pretrained. Given a fixed sampler, the diffusion model induces a sample distribution $p_\theta(x_0|c)$. The objective of denoising diffusion reinforcement learning (RL) is to maximize a reward signal r defined on the samples and contexts:

$$\mathcal{L}_{\mathrm{IMAGERL}}(\theta) = \mathbb{E}_{c \sim p(c), x_0 \sim p_\theta(x_0|c)}[r(x_0, c)], \tag{9}$$

for a context distribution $p(c)$ of our choosing.

To improve the alignment between generated images and text, we need to optimize L_{IMAGERL}. In general, we can use the denoising loss L_{DDPM} (Equation 5), but with training data $x_0 \sim p_\theta(x_0|c)$ and an added weighting that depends on the reward $r(x_0, c)$. We refer to this general class of algorithms as Denoising Diffusion Policy Optimization (DDPO) (Black et al., 2024), framing the training of the diffusion model as a Markov Decision Process (MDP) and performing multi-step optimization for fine-tuning.

3.4 Model Training

Training of Diffusion Models with RL: The training objective is to maximize cumulative rewards, improving the alignment between images and text in Equation 9. We use policy gradient estimation to optimize the model parameters. With access to likelihoods and likelihood gradients, we can make direct Monte Carlo estimates of $\nabla_\theta \mathcal{L}_{IMAGERL}$. The process uses the score function policy gradient estimator, also known as the likelihood ratio method or REINFORCE (Williams, 1992; Mohamed et al., 2020):

$$\nabla_\theta \mathcal{L}_{IMAGERL} = \mathbb{E}\left[\sum_{t=0}^{T} \nabla_\theta \log p_\theta(x_{t-1}|x_t, c)\, r(x_0, c)\right]. \tag{10}$$

Ordered Learning Implementation: In the initial stage, each of the above learning objectives will be executed separately in a certain order to maintain a stable and effective IMAGE system. We first perform $\mathcal{L}_{\mathrm{IMAGERL}}$. After training Diffusion Models, we train LLM with the loss $\mathcal{L}$ which is the combination of $\mathcal{L}_{\mathrm{IMAGERL}}$ and $\mathcal{L}_{\mathrm{MLLM}}$:

$$\mathcal{L} = \frac{\mathcal{L}_{\mathrm{MLLM}}}{\mathcal{L}_{\mathrm{MLLM}}{}^{constant}} + \frac{\mathcal{L}_{\mathrm{IMAGERL}}}{\mathcal{L}_{\mathrm{IMAGERL}}{}^{constant}}, \tag{11}$$

where $constant$ refers to the loss value treated as a constant.

4 Experiment Setup

4.1 Data and Training Setting

Dataset: We conduct experiments on two MT benchmarks: Multi30K(Elliott et al., 2016) and

Language	English → German			English → French			Average
Testset	Test2016	Test2017	MSCOCO	Test2016	Test2017	MSCOCO	
Metric	BLEU ↑ / COMET ↑ / BLEURT ↑						
Traditional MMT							
Soul-Mix	42.5/—/—	34.5/—/—	30.9/—/—	62.4/—/—	54.8/—/—	45.7/—/—	45.1/—/—
RG-MMT-EDC	42.2/—/—	33.4/—/—	30.0/—/—	62.9/—/—	55.8/—/—	45.1/—/—	44.9/—/—
WRA-guided	39.3/—/—	32.3/—/—	28.5/—/—	61.8/—/—	54.1/—/—	43.4/—/—	43.2/—/—
ImagiT	38.6/—/—	32.1/—/—	29.7/—/—	60.8/—/—	52.8/—/—	42.5/—/—	42.7/—/—
Imagination	39.7/—/—	32.3/—/—	28.5/—/—	61.8/—/—	54.1/—/—	43.4/—/—	43.3/—/—
Open-source LLMs based on Text							
Llama3-8B	30.1/69.5/56.6	24.2/66.4/53.0	21.9/62.6/47.8	50.2/77.8/61.1	40.4/72.8/53.3	34.5/70.7/49.9	33.6/69.9/53.6
Alpaca-7B	38.5/77.2/66.2	34.3/76.5/65.9	30.9/72.4/61.5	59.2/82.5/70.2	51.4/79.4/68.3	42.6/77.2/62.9	42.8/77.5/65.8
Vicuna-7B	32.9/75.9/63.5	28.0/75.4/63.5	26.1/70.3/57.7	46.5/81.4/64.8	43.8/82.4/66.3	39.3/78.6/61.0	36.1/77.3/62.8
Tower-7B*	22.1/52.1/34.2	13.7/45.5/25.8	16.3/48.6/31.5	24.5/55.9/31.7	20.8/50.1/25.7	22.5/52.1/29.1	20.0/50.7/29.7
ALMA-7B*	23.1/66.4/59.1	18.9/66.3/57.8	13.7/62.1/55.6	21.4/67.0/52.6	17.4/65.5/50.8	17.9/65.3/52.8	18.7/65.4/54.8
ALMA-R-13B*	29.1/71.8/59.4	24.8/71.8/60.5	23.9/68.2/57.8	27.4/73.7/52.7	24.4/74.5/54.6	29.2/72.8/54.9	26.5/72.1/56.7
Open-source LLMs based on Text & Image							
DreamLLM	27.2/74.8/67.4	19.5/73.5/65.9	19.3/69.4/62.5	36.9/81.1/68.3	34.7/80.6/67.9	36.6/79.2/66.5	29.0/76.4/66.4
IMAGE	**45.3/83.1/78.1**	**38.6/81.9/76.8**	**37.5/78.8/74.6**	**67.5/88.3/81.2**	**61.5/86.6/78.8**	**49.3/82.5/72.6**	**49.9/83.5/77.0**

Table 1: Main translation results from the Multi30K benchmark, with BLEU, COMET, and BLEURT scores. The bolded results indicate the highest statistically significant scores (p-value < 0.01 in the paired t-test against all compared methods). * indicates that no fine-tuning was performed on the Multi30K test set.

WMT24 test set (Kocmi et al., 2024). Dataset details are in Appendix A.1.

Training Setting: Details of our training setting and off-the-shelf tools are in Appendix A.2.

4.2 Comparing Systems

We used two types of baseline methods:

(i) **Traditional Multimodal Machine Translation models (MMT)**, including Soul-Mix(Cheng et al., 2024), RG-MMT-EDC(Tayir and Li, 2024), WRA-guided(Zhao et al., 2022), Imagination(Elliott and Kádár, 2017) and ImagiT(Long et al., 2021). These MMT baselines take the source language sentence as textual input while utilizing the image as visual input. They have completed training on the Multi30k training dataset and reached convergence. The results are cited from the reported data in the paper.

(ii) **Open-source Large language models**, including Llama3-8B, Alpaca-7B, Vicuna-7B, Tower-7B, ALMA-7B, ALMA-R-13B, and DreamLLM. Among them, Llama3-8B(AI@Meta, 2024), Alpaca-7B(Bommasani et al., 2021), and Vicuna-7B(Chiang et al., 2023) are models widely used for multilingual tasks, all of which exhibit strong instruction-following capabilities. For Tower-7B(Alves et al., 2024), ALMA-7B(Xu et al., 2023), and ALMA-R-13B(Xu et al., 2024), these models were pre-trained and fine-tuned on translation datasets, outperforming ChatGPT in multiple language directions. DreamLLM(Dong et al., 2024) is a framework that unifies text and image generation in multimodal Large Language Models.

4.3 Automatic Evaluation

In evaluating our translation methodology, we initially employ COMET[1] (Rei et al., 2022) and BLEURT[2] (Sellam et al., 2020) as automatic metrics, aligning with the established standards in LLM-based translation literature (Chen et al., 2024c; He et al., 2023; Huang et al., 2024). For traditional translation evaluation, we use BLEU [3] (Papineni et al., 2002).

5 Experimental Results

5.1 Main Experiment Results on MMT task

In Table 1, we present the overall experimental results on the classic Multi30K dataset in the MMT field. First, we compare different methods fine-tuned on the same training set. Our method demonstrates significant improvement in translation performance by generating visual information, clearly outperforming text-only translation models based on the same foundational LLM in this task by average 13.7=(12.4+10.6+11.4+21.1+16.7+10)/6 BLEU score, highlighting the critical role of

[1]https://huggingface.co/Unbabel/wmt22-comet-da
[2]https://github.com/lucadiliello/bleurt-pytorch
[3]https://github.com/mjpost/sacrebleu

visual information in text translation (consistent with the conclusion in Section 5.4). Next, we compare our method with traditional multimodal machine translation (MMT) research. Traditional MMT methods, developed over years of study, can make more comprehensive use of annotated image information. However, IMAGE still surpasses these methods, showcasing the potential of multimodal large language models in MT.

5.2 Main Experiment Results on General MT

The effectiveness of IMAGE in general domain translation tasks. In the WMT24 general domain tasks, as shown in Table 2, IMAGE outperforms other methods across 4 language pairs and 3 evaluation metrics. Specifically, in the general domain, the IMAGE method outperforms Vicuna directly by +3.9 BLEU and +8.2 COMET. This indicates that the visual information enhances the translation ability of LLMs in the general MT task.

	En→Zh	En → De	En→Hi	En→Cs
	BLEU ↑ /COMET ↑ /BLEURT ↑			
Llama3-8B	11.6/56.8/33.4	12.7/54.3/36.9	1.2/39.4/31.5	3.2/47.9/25.0
Alpaca-7B	15.0/54.6/45.7	17.1/60.4/56.5	2.9/36.7/36.5	3.4/53.6/36.7
Vicuna-7B	21.8/63.9/36.4	23.3/68.2/52.1	5.6/49.4/45.0	6.7/57.9/45.2
Tower-7B*	13.5/55.5/42.8	17.2/55.7/47.2	2.0/32.1/20.2	1.4/42.9/28.9
ALMA-7B*	14.8/52.9/33.4	17.4/58.1/40.2	1.0/31.9/26.9	1.7/49.7/32.0
ALMA-R-13B*	15.2/57.4/37.2	18.3/57.2/46.8	1.3/34.1/30.9	3.5/53.2/45.5
IMAGE	26.8/77.6/57.4	23.8/73.3/60.8	6.2/51.4/47.3	16.2/69.9/53.9

Table 2: Main translation results from the WMT24 test set, with BLEU and COMET scores. The bolded results indicate the highest statistically significant scores (p-value < 0.01 in the paired t-test against all compared methods). * indicates that no fine-tuning was performed on the WMT24 test set.

The effectiveness of IMAGE in low-resource tasks. We selected 2 low-resource tasks (En→Cs, En→Hi) from WMT24. As observed in Table 2, current low-resource tasks still pose challenges to LLMs. However, compared to baseline methods, IMAGE achieved an average improvement of +14.13 COMET and +3.87 BLEU for En→Hi, and +19.03 COMET and +12.88 BLEU for En→Cs, respectively. This suggests that visual information can provide supplementary data for low-resource tasks, thereby enhancing translation performance in low-resource scenarios.

5.3 Experiment on the Correlation between Reward Scores and MT Performance

We further investigated the impact of the proposed RL training method on model translation performance. Inspired by Wu et al., 2021 and Zhu et al., 2023, we conducted a visual analysis on Multi30K (En→De), using BLEU and Reward scores (calculated as shown in Equation 8) as reference metrics.

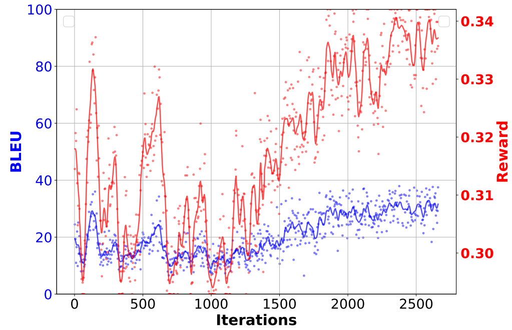

Figure 4: Analysis of the experimental setup for assessing the impact of the Iterative Refinement part on translation performance.

Figure 4 presents the results of the training phase, where the horizontal axis shows the number of training iterations, and the vertical axes show translation performance (left) and RL Reward scores (right). The results indicate that as training progresses, our method continues to optimize, with Reward scores gradually increasing and translation quality improving. Additionally, the Reward score measures the similarity between LSG and VSG. The experimental results show that as this similarity increases, the generated images align more closely with the source sentences, effectively enhancing translation performance.

5.4 Ablation Experiment on Loss

In Table 3, we quantify the contribution of each learning strategy through the ablation study. Each learning strategy has a significant impact on overall performance. The training objective aligning visual and source sentence information demonstrates a notable impact, with an average increase of 1.5 scores. Additionally, multilingual text translation showed a more significant effect, with an average increase of 7 BLEU scores. When using these two training objectives together, we observed the most significant performance improvement, with an average increase of 18.4 BLEU scores. These results confirm the long-standing findings in MMT research on the positive influence of visual information on multilingual translation tasks (Zhao et al., 2020; Fang and Feng, 2022; Elliott et al., 2016).

5.5 Ablation Experiment on Module

In Table 4, We conducted ablation studies on Multi30K to assess the role of each component in

Configuration		English → German		
L_{MLLM}	$L_{IMAGERL}$	Test2016	Test2017	MSCOCO
✗	✗	27.2 / 74.8 / 67.4	19.5 / 73.5 / 65.9	19.3 / 69.4 / 62.5
✗	✓	27.4 / 74.9 / 67.5	22.2 / 74.3 / 66.6	21.0 / 71.5 / 63.2
✓	✗	32.9 / 75.9 / 63.5	28.0 / 75.4 / 63.5	26.1 / 70.3 / 57.7
✓	✓	45.3 / 83.1 / 78.1	38.6 / 81.9 / 76.8	37.5 / 78.8 / 74.6

Table 3: Comparison of configurations with different loss functions (L_{MLLM} and $L_{IMAGERL}$). Metrics are BLEU/COMET/BLEURT.

the IMAGE. Removing the Stable Diffusion model (w/o SD) led to an average BLEU score decrease of 1.7, showing that generated visual information improves multilingual translation. Replacing SD-generated images with real images (w/ RI) caused a 1.8-point drop, indicating SD-generated images provide greater benefits (we will further discuss this phenomenon in Section 5.6). Removing vision encoder (CLIP features) (w/o VS) resulted in a significant BLEU score decline (45.43/38.6/37.5 without CLIP, compared to 39.2/35.1/33.2 with CLIP), highlighting the importance of vision encoder in aligning vision and text.

Language	English→German		
Testset	Test2016	Test2017	MSCOCO
Metrics	BLEU ↑ /COMET ↑ /BLEURT ↑		
IMAGE	45.3/83.1/78.1	38.6/81.9/76.8	37.5/78.8/74.6
- w/o SD	42.9/82.5/77.2	37.7/81.4/76.2	35.6/78.6/73.9
- w/ RI	42.6/82.3/77.0	37.9/81.3/76.1	35.5/78.7/74.1
- w/o VS	39.2/77.7/67.2	35.1/77.4/67.0	33.2/72.7/61.9

Table 4: Comparison of configurations with different modules. SD, RI and VS represent Stable Diffusion, Real Image and Vision Encoder, respectively. Metrics are BLEU, COMET, and BLEURT.

5.6 Evaluation of Generated Image Quality

To investigate the correspondence between the images generated by IMAGE and the source language sentences, we used the pretrained Stable Diffusion model and IMAGE to generate images, and then calculated the $CLIPScore$ (Hessel et al., 2021). $CLIPScore$ measures the similarity between the image and the source language sentence using the formula: $CLIPScore(c, v) = max(cos(c, v), 0)$, where c and v are the feature vectors from the text encoder and the image encoder of CLIP (Radford et al., 2021), respectively.

The evaluation results in Table 5 show that IMAGE outperforms the pretrained Stable Diffusion model across all datasets. Additionally, IMAGE-generated images exhibit higher similarity to the source language sentences than the original related images in Test2016 and Ambiguous COCO. This confirms that our method generates images that better reflect the source language, enhancing translation tasks.

Language	English → German		
Testset	Test2016	Test2017	MSCOCO
Metrics	CLIPScore ↑		
Stable Diffusion	0.72	0.72	0.71
IMAGE (SD)	**0.76**	0.76	**0.75**
Multi30K	0.75	**0.78**	0.74

Table 5: CLIPScore: Similarity between Source Language Sentences and Related Images. indicates Stable Diffusion without fine-tuning. indicates Stable Diffusion fine-tuned with RL (§3.3.1).

We also present some qualitative case study results on the Multi30K En→De test datas in Figure 5. The results indicate that, compared to Stable Diffusion and OpenAIs DALL-E 3[4], our proposed model generates more accurate images based on the source sentences, leading to higher-quality translation outcomes. A key advantage of the IMAGE model is its ability to generate visuals that correctly represent the number and relationships of object instances as defined by the source sentence, ensuring translation accuracy.

6 Related Works

MMT Model Architecture: Multimodal Machine Translation (MMT) aims to enhance machine translation tasks through the aid of visual information (Zhang et al., 2019). Since the release of the Multi30K dataset (Elliott et al., 2016), early research has primarily focused on model architecture design (Zhou et al., 2018; Calixto and Liu, 2017; Helcl et al., 2018). Subsequent studies, such as those by Yao and Wan, 2020 and Yin et al., 2020, proposed multimodal encoders that integrate text and visual information during the encoding stage. Ive et al., 2019 and Lin et al., 2020 applied deliberation networks (Xia et al., 2017) or capsule networks (Sabour et al., 2017) in the decoder to further optimize the use of visual information. Currently, Multimodal Large Language Models (MLLMs) architectures are widely applied in multimodal tasks (Bai et al., 2023; Yue et al., 2024; Li et al., 2024; Huang and Zhang, 2024); however, their application in MMT remains underexplored. Our approach introduces MLLMs in the field of machine translation for

[4] https://openai.com/index/dall-e-3/

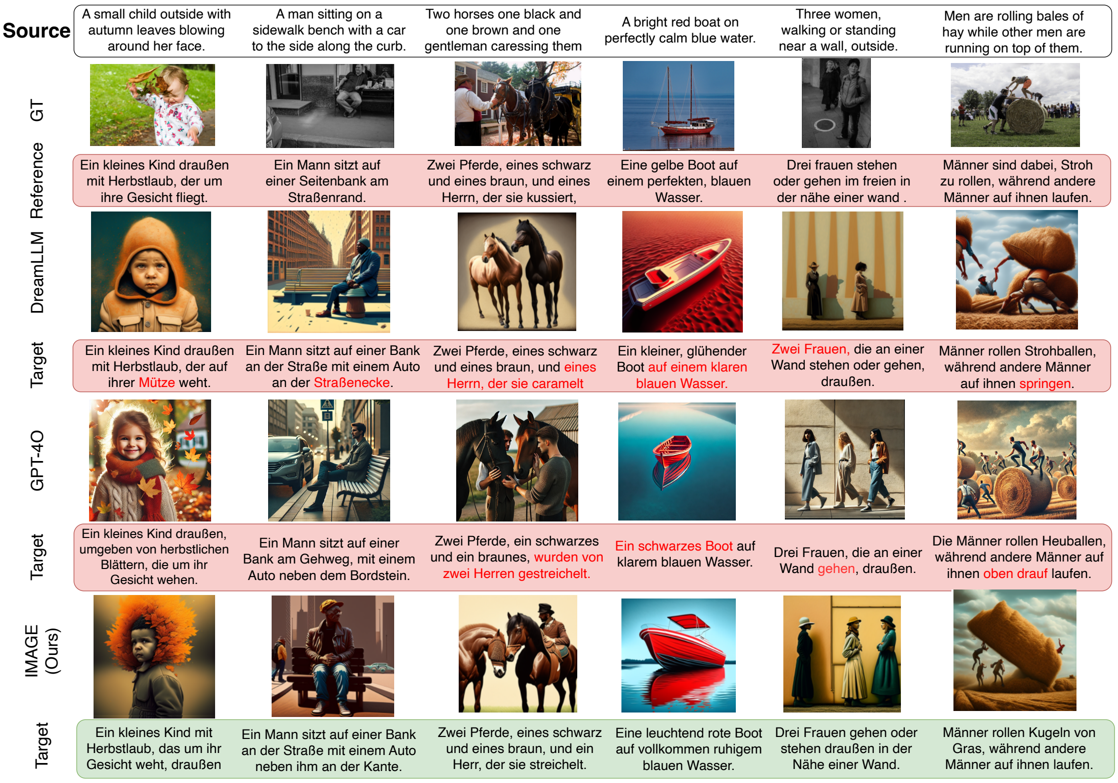

Figure 5: **Some qualitative results on the comparison of IMAGE against related work on the Multi30K En-De test set.** IMAGE, in addition to high quality image generation, correctly generates the number of given instances in the image and represents the scene more accurately overall. GPT-4O refers to using DALL-E for image generation, followed by GPT-4O model performing translation based on the source sentence and the generated image. Red words indicate the parts with translation errors.

the first time, combined with strong text-to-image models (Bolya and Hoffman, 2023; Rombach et al., 2022) to generate highly relevant, high-quality images from the source text, thereby enhancing translation performance.

Image-Free MMT: Traditional multimodal approaches require annotated images corresponding to input text, which limits their practical applicability. To overcome this limitation, Hitschler et al., 2016 proposed using target-end image retrieval to aid translation; Elliott and Kádár, 2017 designed the multi-task learning framework Imagination, which breaks down the translation task into learning both translation and visual association representations; Calixto et al., 2019 introduced latent variables to estimate the joint distribution of translations and images; Long et al., 2021 used Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) to generate visual representations for translation prediction. Additionally, Fei et al., 2023 introduced a visual scene hallucination mechanism to achieve inference-time image-free machine translation. Building on these studies, our approach further enhances translation performance in the absence of image input. The core of our approach includes: 1) eliminating the need for text and image annotation during training, significantly reducing MMT data costs; 2) using consistency training with LSG and VSG to ensure the relevance between source text and generated images, thus improving translation performance; and 3) leveraging the CLIP model to align visual and textual semantic consistency, further reducing noise interference.

7 Conclusion

Our IMAGE framework leverages imaginative generation to enhance LLM-based machine translation, providing clearer visual image that improves translation accuracy. By using graph-based supervision to refine scene and relationship clarity, IMAGE outperforms traditional text-only LLM-MT approaches, especially on complex sentences, and pioneers the integration of visual signals to boost translation performance.

8 Limitation

Our IMAGE method utilizes imaginative generation to enhance machine translation based on large language models (LLMs), delivering a clearer visual image that significantly boosts translation accuracy. However, the translation capability of our method is primarily limited by the multilingual performance of LLMs. Additionally, our method requires collaborative training of LLMs and Stable Diffusion, which demands greater computational resources.

References

AI@Meta. 2024. Llama 3 model card.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. 2022. Flamingo: a visual language model for few-shot learning. In Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022.

Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. Tower: An open multilingual large language model for translation-related tasks.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, pages 6077–6086. Computer Vision Foundation / IEEE Computer Society.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966, 1(2):3.

Gonie Bang. 1986. The imagination of the writer and of the literary translator. Babel, 32(4):198–201.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. 2024. Training diffusion models with reinforcement learning. In The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024. OpenReview.net.

Daniel Bolya and Judy Hoffman. 2023. Token merging for fast stable diffusion. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 4599–4603.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, S. Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen A. Creel, Jared Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren E. Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas F. Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, O. Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir P. Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Benjamin Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, J. F. Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Robert Reich, Hongyu Ren, Frieda Rong, Yusuf H. Roohani, Camilo Ruiz, Jack Ryan, Christopher R'e, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishna Parasuram Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei A. Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2021. On the opportunities and risks of foundation models. ArXiv.

Iacer Calixto and Qun Liu. 2017. Incorporating global visual features into attention-based neural machine translation. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017, pages 992–1003. Association for Computational Linguistics.

Iacer Calixto, Miguel Rios, and Wilker Aziz. 2019. Latent variable model for multi-modal translation. In Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, pages 6392–6405. Association for Computational Linguistics.

Andong Chen, Kehai Chen, Yang Xiang, Xuefeng Bai, Muyun Yang, Tiejun Zhao, and Min Zhang. 2024a. Llm-based translation inference with iterative bilingual understanding. CoRR, abs/2410.12543.

Andong Chen, Lianzhang Lou, Kehai Chen, Xuefeng Bai, Yang Xiang, Muyun Yang, Tiejun Zhao, and Min Zhang. 2024b. Benchmarking llms for translating classical chinese poetry: Evaluating adequacy, fluency, and elegance. arXiv preprint arXiv:2408.09945.

Andong Chen, Lianzhang Lou, Kehai Chen, Xuefeng Bai, Yang Xiang, Muyun Yang, Tiejun Zhao, and Min Zhang. 2024c. DUAL-REFLECT: Enhancing large language models for reflective translation through dual learning feedback mechanisms. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 693–704, Bangkok, Thailand. Association for Computational Linguistics.

Xuxin Cheng, Ziyu Yao, Yifei Xin, Hao An, Hongxiang Li, Yaowei Li, and Yuexian Zou. 2024. Soul-mix: Enhancing multimodal machine translation with manifold mixup. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 11283–11294, Bangkok, Thailand. Association for Computational Linguistics.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, pages 4299–4307.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. CoRR, abs/2311.07919.

Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, Xiangwen Kong, Xiangyu Zhang, Kaisheng Ma, and Li Yi. 2024. Dreamllm: Synergistic multimodal comprehension and creation. In The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024. OpenReview.net.

Chengbin Du, Yanxi Li, Zhongwei Qiu, and Chang Xu. 2023. Stable diffusion is unstable. In Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023.

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, hosted by the 54th Annual Meeting of the Association for Computational Linguistics, VL@ACL 2016, August 12, Berlin, Germany. The Association for Computer Linguistics.

Desmond Elliott and Ákos Kádár. 2017. Imagination improves multimodal translation. In Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017 - Volume 1: Long Papers, pages 130–141. Asian Federation of Natural Language Processing.

Qingkai Fang and Yang Feng. 2022. Neural machine translation with phrase-level universal visual representations. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022, pages 5687–5698. Association for Computational Linguistics.

Hao Fei, Qian Liu, Meishan Zhang, Min Zhang, and Tat-Seng Chua. 2023. Scene graph as pivoting: Inference-time image-free unsupervised multimodal machine translation with visual scene hallucination. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023, pages 5980–5994. Association for Computational Linguistics.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada, pages 2672–2680.

Nuno Miguel Guerreiro, Duarte M. Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André F. T. Martins. 2023. Hallucinations in large multilingual translation models. CoRR, abs/2303.16104.

Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2023. Exploring human-like translation strategy with large language models. ArXiv, abs/2305.04118.

Jindrich Helcl, Jindrich Libovický, and Dusan Varis. 2018. CUNI system for the WMT18 multimodal translation task. In Proceedings of the Third Conference on Machine Translation: Shared Task

Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018, pages 616–623. Association for Computational Linguistics.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. arXiv preprint arXiv:2302.09210.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021, pages 7514–7528. Association for Computational Linguistics.

Julian Hitschler, Shigehiko Schamoni, and Stefan Riezler. 2016. Multimodal pivots for image caption translation. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers. The Association for Computer Linguistics.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual.

Jiaxing Huang and Jingyi Zhang. 2024. A survey on evaluation of multimodal large language models. arXiv preprint arXiv:2408.15769.

Yichong Huang, Xiaocheng Feng, Baohang Li, Chengpeng Fu, Wenshuai Huo, Ting Liu, and Bing Qin. 2024. Aligning translation-specific understanding to general understanding in large language models. arXiv preprint arXiv:2401.05072.

Séverine Hubscher-Davidson. 2020. Translation and the double bind of imaginative resistance. Translation Studies, 13(3):251–270.

Julia Ive, Pranava Madhyastha, and Lucia Specia. 2019. Distilling translations with visual awareness. In Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, pages 6525–6538. Association for Computational Linguistics.

Vivek Iyer, Pinzhen Chen, and Alexandra Birch. 2023. Towards effective disambiguation for machine translation with large language models. In Proceedings of the Eighth Conference on Machine Translation, WMT 2023, Singapore, December 6-7, 2023, pages 482–495. Association for Computational Linguistics.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is chatgpt a good translator? a preliminary study. arXiv preprint arXiv:2301.08745, 1(10).

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinthór Steingrímsson, and Vilém Zouhar. 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, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024. Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2023. Encouraging divergent thinking in large language models through multi-agent debate. arXiv preprint arXiv:2305.19118.

Huan Lin, Fandong Meng, Jinsong Su, Yongjing Yin, Zhengyuan Yang, Yubin Ge, Jie Zhou, and Jiebo Luo. 2020. Dynamic context-guided capsule network for multimodal machine translation. In MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020, pages 1320–1329. ACM.

Bingyan Liu, Chengyu Wang, Tingfeng Cao, Kui Jia, and Jun Huang. 2024. Towards understanding cross and self-attention in stable diffusion for text-guided image editing. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024, pages 7817–7826. IEEE.

Jinxiu Liu and Qi Liu. 2024. R3CD: scene graph to image generation with relation-aware compositional contrastive control diffusion. In Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada, pages 3657–3665. AAAI Press.

Quanyu Long, Mingxuan Wang, and Lei Li. 2021. Generative imagination elevates machine translation.

In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021, pages 5738–5748. Association for Computational Linguistics.

Shakir Mohamed, Mihaela Rosca, Michael Figurnov, and Andriy Mnih. 2020. Monte carlo gradient estimation in machine learning. J. Mach. Learn. Res., 21:132:1–132:62.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35:27730–27744.

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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event, volume 139 of Proceedings of Machine Learning Research, pages 8748–8763. PMLR.

Leonardo Ranaldi, Giulia Pucci, and André Freitas. 2023. Empowering cross-lingual abilities of instruction-tuned large language models by translation-following demonstrations. CoRR, abs/2308.14186.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020, pages 3505–3506. ACM.

Ricardo Rei, José GC De Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André FT Martins. 2022. Comet-22: Unbabel-ist 2022 submission for the metrics shared task. In Proceedings of the Seventh Conference on Machine Translation (WMT), pages 578–585.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019, pages 3980–3990. Association for Computational Linguistics.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022, pages 10674–10685. IEEE.

Sara Sabour, Nicholas Frosst, and Geoffrey E. Hinton. 2017. Dynamic routing between capsules. In Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, pages 3856–3866.

Sebastian Schuster, Ranjay Krishna, Angel X. Chang, Li Fei-Fei, and Christopher D. Manning. 2015. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In Proceedings of the Fourth Workshop on Vision and Language, VL@EMNLP 2015, Lisbon, Portugal, September 18, 2015, pages 70–80. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. arXiv preprint arXiv:2004.04696.

Gunnar A. Sigurdsson, Jean-Baptiste Alayrac, Aida Nematzadeh, Lucas Smaira, Mateusz Malinowski, João Carreira, Phil Blunsom, and Andrew Zisserman. 2020. Visual grounding in video for unsupervised word translation. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020, pages 10847–10856. Computer Vision Foundation / IEEE.

Yuqing Song, Shizhe Chen, Qin Jin, Wei Luo, Jun Xie, and Fei Huang. 2022. Enhancing neural machine translation with dual-side multimodal awareness. IEEE Trans. Multim., 24:3013–3024.

Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. 2020. Unbiased scene graph generation from biased training. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020, pages 3713–3722. Computer Vision Foundation / IEEE.

Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. 2023. What the DAAM: interpreting stable diffusion using cross attention. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023, pages 5644–5659. Association for Computational Linguistics.

Turghun Tayir and Lin Li. 2024. Unsupervised multimodal machine translation for low-resource distant language pairs. ACM Trans. Asian Low Resour. Lang. Inf. Process., 23(4):55.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. CoRR, abs/2307.09288.

Gladys Tyen, Hassan Mansoor, Peter Chen, Tony Mak, and Victor Cărbune. 2023. Llms cannot find reasoning errors, but can correct them! arXiv preprint arXiv:2311.08516.

Ronald J. Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. Mach. Learn., 8:229–256.

Zhiyong Wu, Lingpeng Kong, Wei Bi, Xiang Li, and Ben Kao. 2021. Good for misconceived reasons: An empirical revisiting on the need for visual context in multimodal machine translation. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021, pages 6153–6166. Association for Computational Linguistics.

Yingce Xia, Fei Tian, Lijun Wu, Jianxin Lin, Tao Qin, Nenghai Yu, and Tie-Yan Liu. 2017. Deliberation networks: Sequence generation beyond one-pass decoding. In Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, pages 1784–1794.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2023. A paradigm shift in machine translation: Boosting translation performance of large language models.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation.

Shaowei Yao and Xiaojun Wan. 2020. Multimodal transformer for multimodal machine translation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, pages 4346–4350. Association for Computational Linguistics.

Yongjing Yin, Fandong Meng, Jinsong Su, Chulun Zhou, Zhengyuan Yang, Jie Zhou, and Jiebo Luo. 2020. A novel graph-based multi-modal fusion encoder for neural machine translation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, pages 3025–3035. Association for Computational Linguistics.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9556–9567.

Jiali Zeng, Fandong Meng, Yongjing Yin, and Jie Zhou. 2023. Improving machine translation with large language models: A preliminary study with cooperative decoding. CoRR, abs/2311.02851.

Hongbin Zhang, Kehai Chen, Xuefeng Bai, Yang Xiang, and Min Zhang. 2024. Paying more attention to source context: Mitigating unfaithful translations from large language model. In Findings of the Association for Computational Linguistics ACL 2024, pages 13816–13836, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Wen Zhang, Yang Feng, Fandong Meng, Di You, and Qun Liu. 2019. Bridging the gap between training and inference for neural machine translation. arXiv preprint arXiv:1906.02448.

Tiejun Zhao, Muyun Xu, and Antony Chen. 2024. A review of natural language processing research. Journal of Xinjiang Normal University (Philosophy and Social Sciences), pages 1–23.

Yuting Zhao, Mamoru Komachi, Tomoyuki Kajiwara, and Chenhui Chu. 2020. Double attention-based multimodal neural machine translation with semantic image regions. In Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, EAMT 2020, Lisboa, Portugal, November 3-5, 2020, pages 105–114. European Association for Machine Translation.

Yuting Zhao, Mamoru Komachi, Tomoyuki Kajiwara, and Chenhui Chu. 2022. Word-region alignment-guided multimodal neural machine translation. IEEE ACM Trans. Audio Speech Lang. Process., 30:244–259.

Mingyang Zhou, Runxiang Cheng, Yong Jae Lee, and Zhou Yu. 2018. A visual attention grounding neural model for multimodal machine translation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018, pages 3643–3653. Association for Computational Linguistics.

Yaoming Zhu, Zewei Sun, Shanbo Cheng, Luyang Huang, Liwei Wu, and Mingxuan Wang. 2023. Beyond triplet: Leveraging the most data for multimodal machine translation. In Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023, pages 2679–2697. Association for Computational Linguistics.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul F. Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. CoRR, abs/1909.08593.

A Data and Training Setting

A.1 Dataset Detail

Multi30K(Elliott et al., 2016) We evaluate our methods on two standard benchmarks: Multi30K English→German (En→De) and English→French (En→Fr). Multi30K is a widely used MMT dataset, containing 31,014 images with one English description and the manual translation in German and French. The training and validation sets consist of 29,000 and 1,014 instances, respectively. We reported the results on the Test2016, Test2017, Test2018 and MSCOCO test sets, which includes 1, 000, 1,000, 1071 and 461 instances, respectively.

WMT24 test set (Kocmi et al., 2024) To further validate the effectiveness of our framework in general translation, we also conducted tests on the WMT24 English→German (En→De), English→Chinese (En→Zh), English→Czech (En→Cs), and Eglish→Hindi (En→Hi) test sets. Among them, En→De and En→Zh are high-resource MT tasks, while En-Cs and En-Hi are low-resource tasks.

A.2 Training Setting

Following prior research, we use Mask R-CNN (Tang et al., 2020) as part of a VSG generator[5]. For LSG generation, we parse sentences into dependency trees (Anderson et al., 2018) and transform them into scene graphs based on specific rules (Schuster et al., 2015). The SIM tool for calculating the similarity between LSG and VSG uses Sentence Transformers[6] (Reimers and Gurevych, 2019).

The main experiments is conducted on open-source LLMs from the LLaMA2 family(Touvron et al., 2023). Specifically, we select Dream-LLM(Dong et al., 2024) as our multimodal large language model, which is based on Vicuna-7B (Chiang et al., 2023). The model is trained for 1.5 epochs with a batch size of 16, a peak learning rate of 2e-5 with 3% warmup ratio. We use Deepspeed stage 2(Rasley et al., 2020) to conduct multi-GPU distributed training, with training precision FP16 enabled. For more specific hyperparameters, please refer to our released scripts. For other models used for comparison, such as Llama3-8B and Alpaca , the settings are also the same.

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

[6] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2