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 propose a stable diffusion-based imagination network integrated into a multimodal large language model (MLLM) to explicitly generate an image for each source sentence, thereby advancing multimodal MT. Particularly, we build heuristic feedback with reinforcement learning to ensure the consistency of the generated image with the source sentence without the supervision of visual information, which breaks the high-cost bottleneck of image annotation in MT. Furthermore, the proposed method enables imaginative visual information to be integrated into 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]

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., 2024a; Chen et al., 2024b,a; Chu et al., 2023; Zhu et al., 2023a). 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 enhance understanding and expression by not only relying on text but also imagining visual information (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 visual 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.

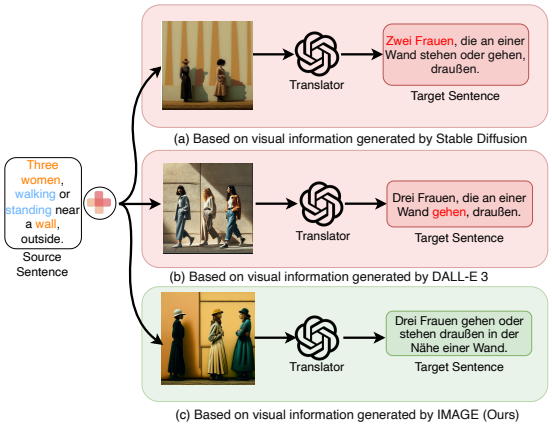

Figure 1: Illustration of the LLMs translation paradigm based on visual information. (a): The generated image does not include information about "three women," and (b): The generated image lacks "standing" information. These issues led to the translation error.

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 highly consistent visual information (image) from the source text, and then uses both source text and visual information to produce

[1] Our code will be publicly available upon acceptance.

better 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 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 framework leverages pre-trained models and reinforcement learning (RL) during training, breaking the bottleneck of high manual annotation costs and eliminating the need for annotated image-text data.
- 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

In recent studies (Bai et al., 2023; Chen et al., 2025; Liu et al., 2023), the multimodal large language models framework consist of three main components: a Large Language Model, 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 $\mathbf{w}_{<t}$ represents all the words preceding the current word. The generation process of each token $\mathbf{w}_t$ in the LLM is modeled :

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

The image encoder takes a single image I as input. This image 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 $\mathcal{P}_\zeta$ (e.g., a linear layer), as described by Alayrac et al., 2022, resulting in visual embeddings $\boldsymbol{V} = \{\mathbf{v}_\ell\}_{\ell=1}^{L}$ of length L.

Maximum likelihood estimation (MLE) aims to minimize the model's 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:

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

$$\boldsymbol{V} = \mathcal{P}_\zeta \circ \mathcal{E}_\phi\left(I\right), \quad (3)$$

wherer Θ represents the set of model parameters, which includes all the learned parameters.

2.2 Scene graph Representation

In MMT, the Scene graph consists of a linguistic semantic graph (LSG) and a visual semantic graph (VSG). LSG represents the semantic structure in text and is denoted as LSG $= (N_L, E_L)$, while VSG represents the semantic structure in visual information and is denoted as VSG $= (N_V, E_V)$. The sets N_L and N_V represent entity nodes in text sentences and visual images, respectively. These include head entities (h^l and h^v) and tail entities (t^l and t^v), where L represents textual information, while V represents visual information ($l \in L$, $v \in V$). The sets E_L and E_V represent the relationships connecting the entities in N_L and N_V, denoted as r^l and r^v, respectively.

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

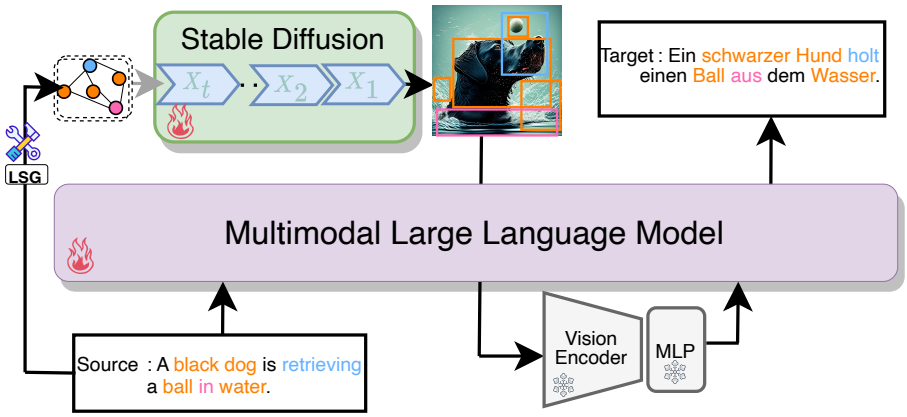

Figure 2: **Overview of our IMAGE framework.** The process involves first generating visual information of the translation input sentence using a diffusion model. Next, the translation result is obtained via LLM, informed by the generated visual information and translation of the original input sentence.

$\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. This noise injection process (the 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 x_0 is the sample dataset, c is the corresponding context, and $\mathbf{x}_{1:T}$ represents the sequence of words from the first to the T-th word. 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:

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

Where $\tilde{\mu}$ is the posterior mean of the forward process, which is a weighted average of x_0 and x_t. Additionally, p represents the distribution of the input data, and $U\{0,T\}$ represents a uniform distribution. 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 framework, IMAGE, integrates visual signals to enhance LLM performance in multilingual translation. To maintain entity consistency in generated visuals, we apply reinforcement learning from feedback. Figure 2 outlines IMAGE, with key components detailed in: multimodal translation (§ 3.2), reinforcement learning from feedback (§ 3.3), and training process (§ 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 OpenAI's 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.

3.3 Reinforcement Learning from Feedback

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

3.3.1 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. We design a reward function to assess this consistency between VSG and LSG, as shown in Figure 3. Higher LSG-VSG similarity indicates stronger alignment. The reward function scores consistency from 0 to 1, making this task a reinforcement learning from feedback (Ouyang et al., 2022; Christiano et al., 2017).

For LSG and VSG generation, we utilize two off-the-shelf SG parsers to obtain LSG and VSG separately (as detailed in §4.3). Due to the differing number of triples in LSG and VSG, we designed a

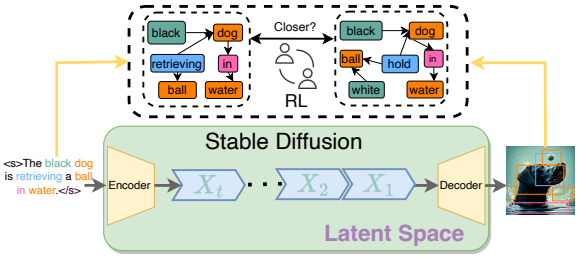

Figure 3: **RL Training Detail.** The overview of IMAGE, which leverages an alignment feedback learning framework to comprehensively enhance the visual signals performance.

structured similarity calculation method to measure their consistency. For each LSG triple, we compute its similarity with all VSG triples and take the highest score as its matching degree

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

$$\text{Sim}(\text{LSG}, \text{VSG}) = \frac{\text{SIM}(h^l,h^v)+\text{SIM}(r^l,r^v)+\text{SIM}(t^l,t^v)}{3}, \tag{7}$$

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 §4.3). Finally, the consistency reward score between sentences and images is the average score of all text triples:

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

where c denotes the LSG of the source sentence, N is the total number of the training set, and x_0 represents the generated image.

3.3.2 Feedback Optimization For Diffusion Model

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

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

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

To improve the alignment between generated images and text, we need to optimize $\mathcal{L}_{\text{IMAGERL}}$. In general, we can use the denoising loss $\mathcal{L}_{\text{DDPM}}$ (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):

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

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 $\mathcal{L}_{\text{IMAGERL}}$. After training diffusion models, we train LLM with the loss $\mathcal{L}$ which is the combination of $\mathcal{L}_{\text{IMAGERL}}$ (Equation 9) and $\mathcal{L}_{\text{MLLM}}$ (Equation 2):

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

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

4 Experiment Setup

4.1 Dataset

We experiment on Multi30K (Elliott et al., 2016) and the WMT24 test set (Kocmi et al., 2024), with dataset details in Appendix A.1.

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

Language	English → German			English → French			
Testset	Test2016	Test2017	MSCOCO	Test2016	Test2017	MSCOCO	**Average**
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: The Multi30K benchmark results include BLEU, COMET, and BLEURT scores, with bolded values indicating the highest. Each test set was evaluated five times, confirming stability and robustness through hypothesis testing ($p < 0.01$). * denotes no fine-tuning on Multi30K.

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., 2023a), 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 MLLMs.

4.3 Training Setting

Following prior work, we use Mask R-CNN (Tang et al., 2020) in the VSG generator[2] and parse sentences into dependency trees (Anderson et al., 2018) to construct LSG based on specific rules (Schuster et al., 2015)[3]. Sentence Transformers[4] (Reimers and Gurevych, 2019) measure LSG-VSG similarity. Experiments use open-source LLMs from the LLaMA2 family (Touvron et al., 2023), with DreamLLM (Dong et al., 2024) (Vicuna-7B (Chiang et al., 2023)) as the primary multimodal model. Training runs for 1.5 epochs with batch size 16 on A100 80G GPUs, using a peak learning rate of 2e-5 (3% warmup). Multi-GPU training is performed with DeepSpeed stage 2 (Rasley et al., 2020) and FP16 precision. Hyperparameter details are in the released scripts. Comparison models, including Llama3-8B and Alpaca, use the same settings. Our method achieves 15.712s/sample, matching DreamLLM's speed while delivering better performance.

4.4 Automatic Evaluation

We evaluate our translation method using COMET (Rei et al., 2022) and BLEURT (Sellam et al., 2020) for LLM-based translation, following established standards (Chen et al., 2024c; He et al., 2023; Huang et al., 2024), and BLEU (Post, 2018) for traditional evaluation.

5 Experimental Results

5.1 Main Experiment Results on MMT task

Table 1 presents the experimental results on the Multi30K dataset. Our method, which generates visual information, significantly outperforms text-only translation models based on the same foundational LLM,

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

[3]https://github.com/scofield7419/UMMT-VSH/tree/master/SG-parsing/LSG

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

achieving an average BLEU improvement of 20.28=(16.3+7.1+13.8+29.9+31.2+23.4)/6. This highlights the crucial role of visual information in translation (consistent with Section 5.4). Compared to traditional MMT approaches that leverage annotated images, our method still achieves superior performance, demonstrating the potential of multimodal large language models in machine translation.

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: The WMT24 test set results, including BLEU and COMET scores, are shown. Bolded values represent the highest scores. Each test set was evaluated five times, confirming result stability and robustness through hypothesis testing ($p < 0.01$). * indicates no fine-tuning on WMT24.

The effectiveness of IMAGE in low-resource tasks. We selected two low-resource tasks (En→Cs/Hi) from WMT24. As shown in Table 2, LLMs still struggle with these tasks. However, IMAGE outperforms baseline methods, achieving 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. This highlights the role of visual information in enhancing MT 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., 2023b, we conducted a visual analysis on Multi30K (En→De), using BLEU and Reward scores (calculated as shown in Equation 8) as reference metrics.

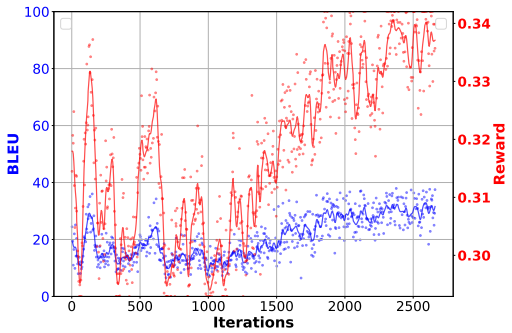

Figure 4: Correlation between Reward Scores and MT Performance

Figure 4 illustrates the training process, with iterations on the horizontal axis and translation performance (left) and RL Reward scores (right) on the vertical axes. Results show continuous optimization, with increasing Reward scores and improved translation quality. The Reward score measures the similarity between LSG and VSG, and higher similarity leads to better-aligned images, 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).

Configuration		English → German		
$\mathcal{L}_{MLLM}$	$\mathcal{L}_{IMAGERL}$	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. The ✓ and ✗ indicate whether the training includes the $\mathcal{L}_{MLLM}$ and $\mathcal{L}_{IMAGERL}$. Metrics are BLEU/COMET/BLEURT.

5.5 Ablation Experiment on Module

Table 4 presents ablation studies on Multi30K to evaluate each component of IMAGE. Removing

Stable Diffusion (w/o SD) led to a 1.7 BLEU drop, confirming the benefit of generated visual information in multilingual translation. Using real images instead of SD-generated ones (w/ RI) resulted in a 1.8-point decline, suggesting SD-generated images are more effective (discussed further in Section 5.6). Removing the vision encoder (w/o VS) significantly reduced BLEU scores (45.43/38.6/37.5 → 39.2/35.1/33.2), highlighting its crucial role in vision-text alignment.

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.

5.6 Evaluation of Generated Image Quality

To examine the alignment between IMAGE-generated images and source sentences, we used a pretrained Stable Diffusion model and IMAGE to generate images and evaluated them with CLIPScore(Hessel et al., 2021). CLIPScore measures text-image similarity as $\text{CLIPScore}(c, v) = \max(\cos(c, v), 0)$, where c and v are feature vectors from CLIP's text and image encoders(Radford et al., 2021). Since this experiment evaluates the generated images, it reduces the risk of data exposure in CLIP pre-training. Results in Table 5 show that IMAGE outperforms Stable Diffusion across all datasets. Moreover, IMAGE-generated images exhibit higher similarity to source sentences than original related images in Test2016 and Ambiguous COCO, confirming its effectiveness in 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 measures the similarity between source language sentences and related images. ❄ represents Stable Diffusion without fine-tuning, while 🔥 denotes 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 and 6. The results indicate that, compared to Stable Diffusion and OpenAI's DALL-E 3[5], 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.

5.7 Human Evaluation

To verify text-image consistency and potential over-reliance on visual information, we conducted a human evaluation. We sampled 50 instances from the Multi30k En→De 2016 test set, generated images using Stable Diffusion, DALL-E, and IMAGE, and translated them with DreamLLM (Vicuna-based). Each instance was rated on Text-Image Alignment and Over-reliance on Visual Information using a 1-3 scale, where 1 indicates poor consistency or excessive reliance, and 3 indicates good consistency or no over-reliance. Two evaluators scored each instance independently, averaging the final scores (details in Appendix B).

Method	**T-Img Align. ↑**	**Over-rel. Vis. ↓**	**BLEU/COMET/BLEURT ↑**
SD+DreamLLM	1.44	2.68	26.3/76.5/68.5
DALL-E+DreamLLM	2.14	2.52	28.8/78.3/70.3
IMAGE	2.68	2.50	54.3/85.6/81.2

Table 6: Human evaluation results. T-Img Align.: Text-Image Alignment, Over-rel. Vis.: Over-reliance on Visual Information.

Table 6 shows that our method generates images with better text consistency than other approaches. Additionally, excessive reliance on visual information is significantly influenced by the degree of text-image alignment. Lower consistency increases dependence on visual information, thereby affecting machine translation performance.

6 Related Works

MMT Model Architecture: Multimodal Machine Translation (MMT) enhances translation by incorporating visual information (Zhang et al., 2019). Since the introduction of the Multi30K dataset (Elliott et al., 2016), early research focused on model architectures (Zhou et al., 2018; Calixto and Liu, 2017; Helcl et al., 2018). Later studies explored multimodal encoders integrating text and visuals (Yao and Wan, 2020; Yin et al., 2020), as well as deliberation and capsule networks in decoders (Ive et al., 2019; Lin et al., 2020). Some works propose pretrained encoder-decoder

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

Figure 5: **Qualitative comparison of IMAGE on Multi30K En-De test set.** IMAGE not only generates high-quality images but also accurately reflects object counts and scene details. GPT-4o uses DALL-E for image generation, followed by translation with GPT-4o. Red words highlight translation errors.

frameworks for MT (Shan et al., 2022; Vijayan et al., 2024). While Multimodal Large Language Models (MLLMs) are widely used in multimodal tasks (Bai et al., 2023; Yue et al., 2024; Li et al., 2024; Huang and Zhang, 2024), their role in MMT remains underexplored. We introduce MLLMs for MMT, leveraging strong text-to-image models (Bolya and Hoffman, 2023; Rombach et al., 2022) to generate high-quality, contextually relevant images, enhancing MT performance.

Image-Free MMT: Traditional multimodal translation methods rely on annotated images, limiting practical applicability. To address this, prior works explored alternative strategies: Zhang et al., 2020 used target-end image retrieval; Elliott and Kádár, 2017 proposed the "Imagination" multi-task framework; Calixto et al., 2019 introduced latent variables for joint translation-image modeling; Long et al., 2021 employed GANs (Goodfellow et al., 2014) to generate visual features; Fei et al., 2023 designed a visual scene hallucination mechanism for image-free translation; and Yuasa et al., 2023 incorporated diffusion models for image generation, though still requiring annotated data. Our approach enhances translation without image input by removing text-image annotations, improving source-text relevance via LSG-VSG consistency, and leveraging CLIP for better visual-text alignment.

RL Fine-tuning Diffusion Models: Fan and Lee, 2023 proposed a policy gradient-based training method to improve the sampling efficiency of diffusion models. Following this direction, Black et al., 2024; Fan et al., 2023 utilized policy gradient algorithms to optimize text-to-image diffusion models for better alignment with human preferences, incorporating a single-image-based reward function (Xu et al., 2023b; Zhang et al., 2024b). Miao et al., 2024 introduced a diversity reward mechanism based on image sets to efficiently measure the discrepancy between the generated and reference distributions. Our approach pioneers LSG-VSG feedback for better image-text alignment, eliminating annotated data reliance, and reducing training costs.

7 Conclusion

Our IMAGE framework enhances LLM-based translation by generating clear visual representations, refining scene and relationship clarity through graph-based supervision. It outperforms text-only LLM-MT, especially on complex sentences, pioneering visual signal integration for improved translation.

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.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. 2025. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*.

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.

Ying Fan and Kangwook Lee. 2023. Optimizing ddpm sampling with shortcut fine-tuning. In *International Conference on Machine Learning*.

Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, P. Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. 2023. Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. *ArXiv*, abs/2305.16381.

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.

Matthieu Futeral, Cordelia Schmid, Ivan Laptev, Benoît Sagot, and Rachel Bawden. 2023. Tackling ambiguity with images: Improved multimodal machine translation and contrastive evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5394–5413, Toronto, Canada. 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.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning.

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.

Zichen Miao, Jiang Wang, Ze Wang, Zhengyuan Yang, Lijuan Wang, Qiang Qiu, and Zicheng Liu. 2024. Training diffusion models towards diverse image generation with reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10844–10853.

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.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 186–191. Association for Computational Linguistics.

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.

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*.

Bin Shan, Yaqian Han, Weichong Yin, Shuohuan Wang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2022. Ernie-unix2: A unified cross-lingual cross-modal framework for understanding and generation. *ArXiv*, abs/2211.04861.

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*.

Vipin Vijayan, Braeden Bowen, Scott Grigsby, Timothy Anderson, and Jeremy Gwinnup. 2024. Adding multimodal capabilities to a text-only translation model. *ArXiv*, abs/2403.03045.

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.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2023a. 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.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. 2023b. Imagereward: learning and evaluating human preferences for text-to-image generation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 15903–15935.

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.

Ryoya Yuasa, Akihiro Tamura, Tomoyuki Kajiwara, Takashi Ninomiya, and Tsuneo Kato. 2023. Multimodal neural machine translation using synthetic images transformed by latent diffusion model. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 76–82, Toronto, Canada. 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. 2024a. 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*.

Yinan Zhang, Eric Tzeng, Yilun Du, and Dmitry Kislyuk. 2024b. Large-scale reinforcement learning for diffusion models. In *European Conference on Computer Vision*.

Zhuosheng Zhang, Kehai Chen, Rui Wang, Masao Utiyama, Eiichiro Sumita, Z. Li, and Hai Zhao. 2020. Neural machine translation with universal visual representation. In *International Conference on Learning Representations*.

Tiejun Zhao, Muven 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.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Lingpeng Kong, Jiajun Chen, Lei Li, and Shujian Huang. 2023a. Multilingual machine translation with large language models: Empirical results and analysis. *arXiv preprint arXiv:2304.04675*.

Yaoming Zhu, Zewei Sun, Shanbo Cheng, Luyang Huang, Liwei Wu, and Mingxuan Wang. 2023b. 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.

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.

B Human Evaluation Details

To ensure the quality and reliability of the evaluation, we recruited 2 annotators with a strong background in both machine translation and image processing, each with at least two years of experience in these fields. Before the evaluation began, the annotators participated in calibration sessions, where they reviewed the scoring criteria and discussed examples of different degrees of alignment and over-reliance to ensure consistency across evaluations. The two metrics were defined as follows:

1. **Text-Image Alignment:** This metric assessed how well the generated image reflected the content and context of the translated text. Scores were assigned based on the degree of congruence between the image and the text, with 1 indicating significant misalignment and

3 indicating a strong match between the image and the textual content.

2. **Over-reliance on Visual Information:** This metric evaluated the extent to which the translation leaned on the visual content rather than relying on the semantic richness of the text. A score of 1 was given if the translation showed clear dependence on the visual elements, compromising the accuracy or richness of the text. A score of 3 was given if the translation was well-grounded in the textual content, with no excessive influence from the generated image.

Each instance was independently scored by 2 evaluators. To maintain objectivity and minimize potential biases, annotators worked independently and were blinded to the identity of other evaluators' scores. After individual evaluations, the final score for each aspect was determined based on the average of the two evaluators' scores. In cases where a significant disparity occurred (i.e., a difference of more than 1 point), the annotators discussed the instance in question to reach a consensus.

C Experiment on Ambiguity Resolution

The purpose of this experiment is to evaluate the effectiveness of the IMAGE framework in resolving translation ambiguities. Specifically, we aim to test the performance of our method on datasets designed to challenge models with ambiguous language, such as CoMMuTE (Futeral et al., 2023). We conducted experiments within the CoMMuTE benchmark, which requires high-quality annotated images to resolve ambiguities in translation. We compared our method against three strong baselines: Graph-MMT, Gated Fusion, VTLM + MMT, and VGAMT. Additionally, we included the DreamLLM model for comparison. The results of the CoMMuTE task are presented in the table below:

Method	En-Fr (ACC)	En-De (ACC)
Graph-MMT*	50.2	49.1
Gated Fusion*	50.0	49.7
VTLM + MM*	50.1	50.0
VGAMT*	67.1	59.0
DreamLLM	50.0	50.0
IMAGE	50.1	50.5

Table 7: Performance on CoMMuTE task (accuracy in %). * denotes results reported from paper Futeral et al., 2023.

Although the IMAGE framework was not specifically designed for disambiguation tasks, it still performs comparably to three strong baselines in the CoMMuTE task. This demonstrates the robustness and versatility of our approach. However, the results around 50.0 for this metric indicate that while the visual content does not significantly enhance the translation, it also does not detract from it. This suggests that the metric may not be a good fit for our method in this context.

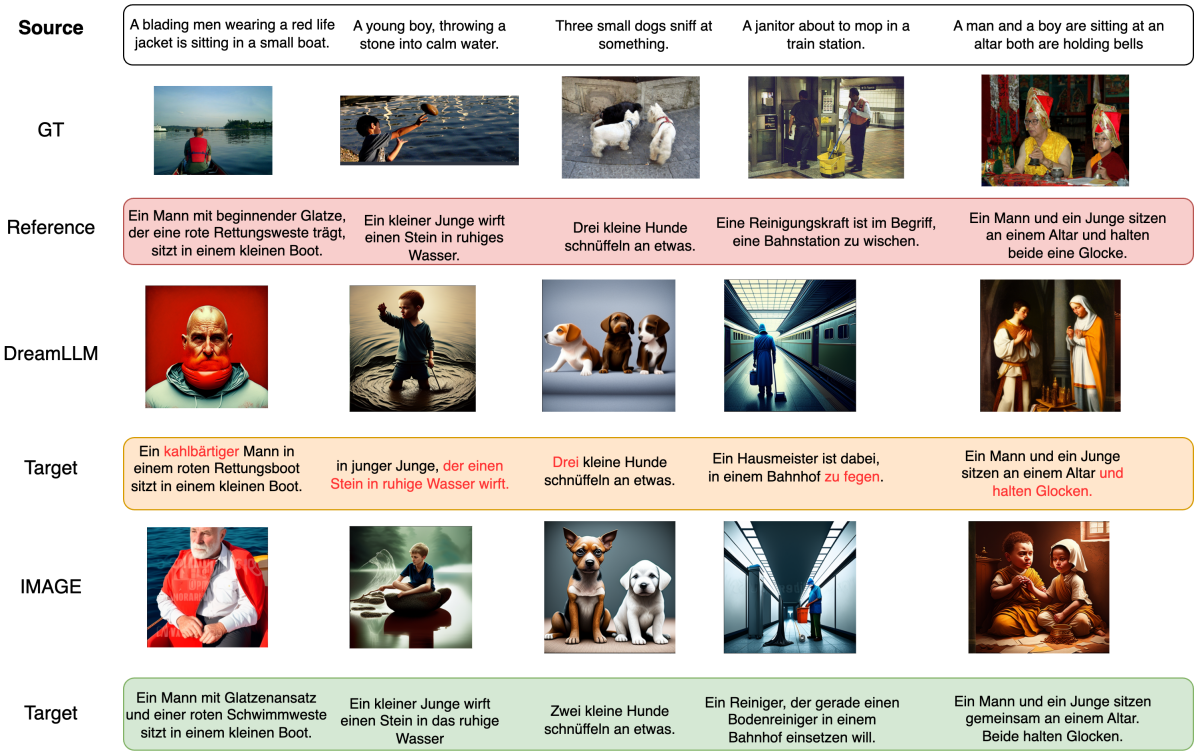

Figure 6: **Qualitative comparison of IMAGE on Multi30K En-De devlopment set.** IMAGE not only generates high-quality images but also accurately reflects object counts and scene details. Red words highlight translation errors.