Imagination and Contemplation: A Balanced Framework for Semantic-Augmented Multimodal Machine Translation

Zhuang Yu[1], Shiliang Sun[1]*, Jing Zhao[2]*, Tengfei Song[3], Hao Yang[3]
[1]State Key Laboratory of Submarine Geoscience, School of Automation and Intelligent Sensing, Shanghai Jiao Tong University, Shanghai 200240, China
[2]School of Computer Science and Technology, East China Normal University, Shanghai 200062, China
[3]2012 Labs, Huawei Technologies CO., LTD, China
yyyyzzzz@sjtu.edu.cn, shiliangsun@gmail.com, jzhao@cs.ecnu.edu.cn, {songtengfei2, yanghao30}@huawei.com

Abstract

Multimodal Machine Translation (MMT) enhances textual translation through auxiliary inputs such as images, which is particularly effective in resolving linguistic ambiguities. However, visual information often introduces redundancy or noise, potentially impairing translation quality. To address this challenge, we propose a balanced semantic-augmented framework that integrates **Imagination** and **Contemplation** in multimodal understanding. Specifically, we first generate synthetic images from the source text and align them with the authentic images via an optimal transport (OT) loss to enhance visual semantic consistency. A CLIP-based similarity gating mechanism is introduced to adaptively fuse visual features from both authentic and synthetic images during visual representation learning. To strengthen semantic grounding, a neural machine translation (NMT) branch is incorporated as a regularization signal, and a Kullback-Leibler (KL) divergence is applied between MMT and NMT outputs to mitigate modality mismatch. Furthermore, an image-text contrastive (ITC) loss aligns the final translations with image representations, reinforcing multimodal coherence. Experiments on multiple translation datasets with a diverse set of language pairs demonstrate that our framework outperforms existing baselines, particularly in cases with visually ambiguous or weakly correlated content.

1 Introduction

Multimodal machine translation (MMT) refers to methods that leverage information from various modalities to improve translation performance (Specia et al., 2016; Caglayan et al., 2019; Yao and Wan, 2020; Caglayan et al., 2021; Fei et al., 2023; Tayir et al., 2024; Tayir and Li, 2024a). A common approach integrates visual information into translation using bilingual corpora annotated with corresponding images (Specia et al., 2016; Barrault et al., 2018). As shown in Figure 1, visual context can help resolve ambiguities and improve translation accuracy (Futeral et al., 2023; Vijayan et al., 2024).

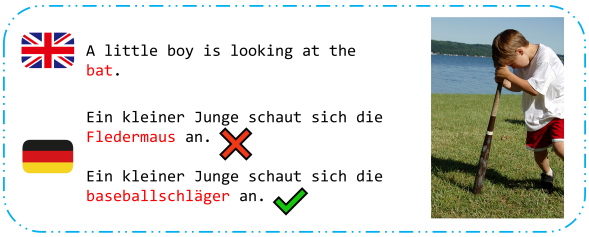


Figure 1: An example of visual information resolving the ambiguity of a word. The English word "bat" can have different meanings depending on the context, such as "a flying mammal" or "a baseball bat". When translated into German, the meanings will result in "*fledermaus*" or "*baseballschläger*".

However, despite the intuitive benefits, visual information does not always improve performance. Misalignment between text and image, visual redundancy, and inappropriate cross-modal fusion strategies can even degrade results (Grönroos et al., 2018; Lala et al., 2018; Li et al., 2022a; Vijayan et al., 2024). Effectively leveraging visual input and harmonizing semantics across modalities remains a core challenge in MMT research.

Recently, the emergence of large-scale vision-language models (VLMs), such as Flamingo (Alayrac et al., 2022), CLIP (Radford et al., 2021) and BLIP (Li et al., 2022b), have introduced a new paradigm for advancing MMT. These models learn generalized cross-modal representations from massive amounts of image-text pairs, offering a stronger semantic foundation for vision-enhanced translation. In parallel, the development of diffusion-based (Ho et al., 2020; Rombach et al., 2022) text-to-image generation models support controllable image generation, offering a promising way to enhance semantic alignment by synthesiz-

*Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 10913–10928
November 4-9, 2025

ing visual scenes grounded in textual input.

Despite these advances, current MMT systems have yet to fully integrate the capabilities of such large models. On one hand, most approaches still rely solely on raw image features, without exploring the feasibility of semantic augmentation using additional visual information. On the other hand, even when large pre-trained models are incorporated, it remains challenging to integrate visual signals in a controllable and flexible manner. Moreover, while incorporating visual enhancements, the dominant role of the source text should not be compromised, which further increases the difficulty of balancing vision and language in model design and training.

To address these issues, we propose a semantic-augmented MMT framework inspired by the dual concepts of "Imagination" and "Contemplation". In Imagination, we employ a text-to-image generation model to synthesize images that are semantically aligned with the source sentence. These synthetic images are then aligned with authentic images via optimal transport (OT) loss, and integrated using a CLIP-guided vision gated fusion (VGF) that adaptively regulates the visual contribution during fusion. In Contemplation, we introduce a parallel NMT branch and apply the KL divergence to regularize the output distribution between MMT and NMT, encouraging textual grounding. Additionally, an image-text contrastive (ITC) loss is employed to reinforce the semantic alignment between the generated translation and visual input.

Our contributions can be summarized as follows.

- We propose a novel semantic-augmented MMT framework that integrates both synthetic and authentic images to enrich visual semantics while maintaining strong textual grounding.
- We develop a controllable and semantically aligned integration mechanism, combining CLIP-guided VGF, OT loss to align visual representation, and a parallel NMT branch with KL divergence and contrastive loss to enforce visual-textual consistency.
- Experiments show that our approach significantly improves translation accuracy in MMT tasks, particularly excelling in handling complex semantics and text-image inconsistency.

2 Related Work

2.1 Multimodal Machine Translation

Since statistical machine translation, researchers have applied multimodal information to enhance machine translation systems (Afli et al., 2016; Hitschler et al., 2016). From early RNN-based encoder-decoder architectures (Zaremba et al., 2014) to transformer-based architectures (Vaswani et al., 2017), researchers have focused on leveraging visual cues to support text translation. Calixto et al. (2017) proposed a dual-attention mechanism incorporating spatial visual features, while Gated Fusion (Wu et al., 2021) introduced a gating mechanism for cross-modal fusion. Li et al. (2022a) employed selective attention to integrate images and text, and VALHALLA (Li et al., 2022c) introduced visual hallucination for training. Recent works such as Guo et al. (2023) and Chen et al. (2025) attempted to integrate synthetic images into the translation pipeline. However, few have explored how to actively enhance or regulate visual semantics, leaving a gap in handling noisy or partially aligned multimodal input.

Our work addresses this gap by introducing a semantically controllable fusion mechanism that combines synthetic images, optimal transport alignment, and CLIP-guided VGF to adaptively regulate visual contributions during translation.

2.2 Vision-Language Models in MMT

The emergence of large-scale VLMs has significantly advanced the development of multimodal understanding. These models are trained on massive image-text pairs, enabling them to learn cross-modal representations with strong generalization capabilities. Several studies (Li et al., 2022a; Zuo et al., 2023; Futeral et al., 2023; Zhu et al., 2023; Chen et al., 2025) used CLIP embeddings as visual features, while Wang et al. (2024) used BLIP for image captioning to augment text data.

Moreover, large language models (LLMs) such as GPT (Radford et al., 2018, 2019; Brown et al., 2020; OpenAI et al., 2024), PaLM (Chowdhery et al., 2023; Anil et al., 2023) and mT5 (Xue et al., 2021a) have demonstrated remarkable generalization and few-shot capabilities in translation tasks, especially in low-resource and zero-shot settings. These models implicitly encode rich linguistic and world knowledge, providing strong priors that significantly reduce the reliance on domain-specific training data.

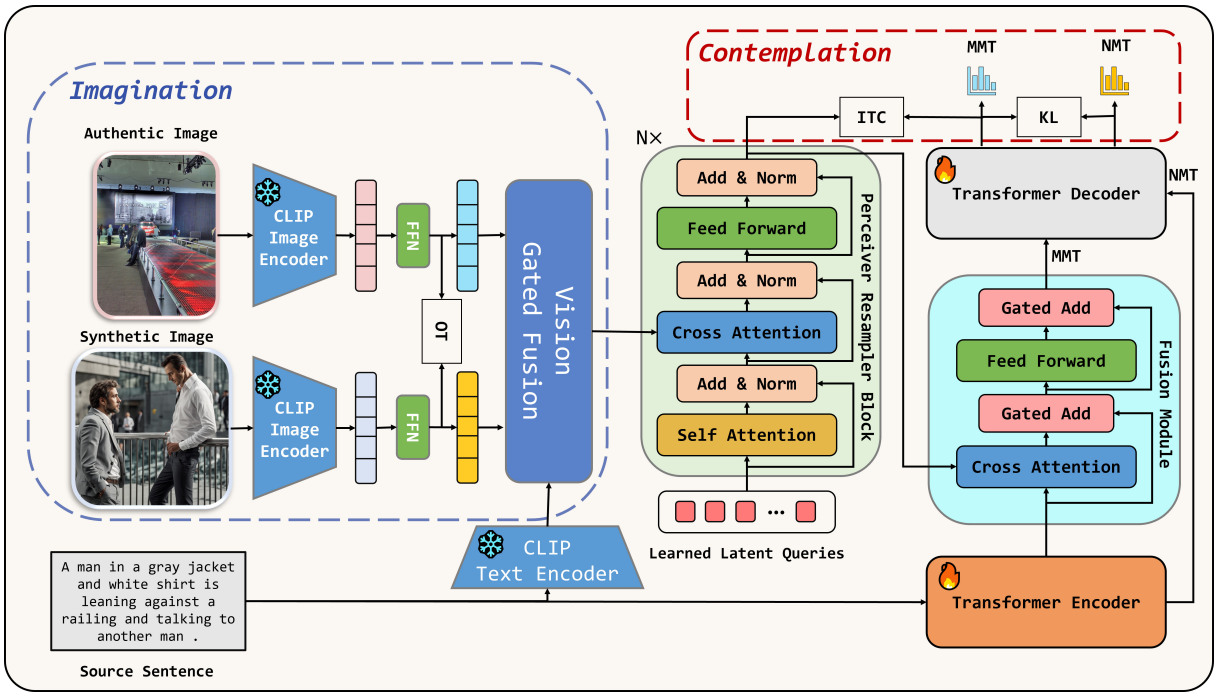


Figure 2: Overview of the framework of the proposed ImCo-MMT. The blue box represents "Imagination" while the red box represents "Contemplation". The learned latent quries are a set of learnable embeddings to extract visual representation most relevant to the text. For Transformer Encoder and Decoder, we set up two settings: training from scratch and directly using the pre-trained model.

2.3 Text-to-Image Generation

Text-to-image generation has recently witnessed remarkable progress with the development of diffusion-based models such as DALL·E (Ramesh et al., 2021, 2022) and Stable Diffusion (Rombach et al., 2022). These models enable image synthesis conditioned on textual prompts, effectively imagining visual scenes from linguistic descriptions.

In MMT tasks, text-to-image generation has been used to produce images that semantically align with the source text (Long et al., 2021; Li et al., 2022c; Guo et al., 2023; Chen et al., 2025), helping to address issues like missing authentic images or discrepancies between text and image. This approach leverages the generated visual information to increase the semantic understanding of the MMT model, thus improving the quality and robustness of translation.

3 Method

3.1 Preliminaries

Neural Machine Translation. Given a parallel corpus $D = (X, Y)$, where X and Y denote the source and target language, respectively, the translation model learns to generate Y from X (Bahdanau, 2014). The training objective is to minimize the cross-entropy loss:

$$\mathcal{L}_{\text{NMT}} = -\sum_{i=1}^{|y|} \log p(y_i|y_{<i}, x). \quad (1)$$

Multimodal Machine Translation. MMT extends NMT by incorporating an additional visual modality V^a, forming triplets $D = (X, Y, V^a)$. The model conditions on both text and image to generate the target sentence. The loss function becomes:

$$\mathcal{L}_{\text{MMT}} = -\sum_{i=1}^{|y|} \log p(y_i|y_{<i}, x, v^a). \quad (2)$$

3.2 Framework Overview

As shown in Figure 2, the framework comprises three core components: i) a dual-image encoder with authentic and synthetic images and a vision gated fusion, ii) a vision-text fusion module built upon a Perceiver Resampler (Vijayan et al., 2024), and iii) a contemplation branch that regularizes translation with text-only guidance.

Both authentic and synthetic images are encoded by CLIP image encoders, and their features are fused through a CLIP-guided vision gated fusion that dynamically adjusts the contribution of each image based on its similarity to the source text.

To capture and compress rich visual semantics from both authentic and generated images, we employ a Perceiver Resampler, inspired by Flamingo

(Alayrac et al., 2022) and BLIP-2 (Li et al., 2023). The Resampler introduces a set of learnable latent queries, which interact with the image features via repeated cross-attention. Each latent queries act as a semantic bottleneck that selectively attend to distinct regions or aspects of the visual input. This design distills high-dimensional visual features into a compact latent representation, facilitating efficient and stable fusion with the source text encoding. The latent-based visual representation is then fused with the source text features, yielding a joint representation that is semantically enriched.

Finally, to preserve the dominance of the source text, a parallel NMT branch is added. Through KL divergence and contrastive losses, this branch aligns the multimodal decoder with the text-only baseline, promoting semantic consistency while preserving the advantages of visual input.

3.3 Imagination

To enrich the visual semantics and alleviate vision-text misalignment, we introduce the Imagination module, which leverages synthetic images and enforces semantic alignment through optimal transport (OT) and vision gated fusion (VGF).

Given a source sentence x, we generate a synthetic image I_s using a text-to-image model (e.g., Stable Diffusion), and obtain visual features from both I_s and the authentic image I_a via a frozen CLIP image encoder:

$$\begin{aligned}\mathbf{V}_a &= \phi_v\left(I_a\right) = [v_a^{(1)}, \cdots, v_a^{(N)}] \in \mathbb{R}^{N\times d},\\ \mathbf{V}_s &= \phi_v\left(I_s\right) = [v_s^{(1)}, \cdots, v_s^{(N)}] \in \mathbb{R}^{N\times d},\end{aligned} \tag{3}$$

where N denotes the number of visual patches and d the feature dimension.

Although both images are derived from the same text, their representations can capture different visual semantics, leading to inconsistency or even conflict in translation. To regularize the semantic consistency between $\mathbf{V}_a$ and $\mathbf{V}_s$, we adopt OT loss (Guo et al., 2023), which softly aligns the patch-level features by minimizing their total transport cost under a transport mass $\mathbf{T}$:

$$\mathcal{L}_{\mathrm{OT}}\left(\mathbf{V}_a, \mathbf{V}_s\right) = \sum_{i=1}^{N}\sum_{j=1}^{N} \mathbf{T}_{ij}\cdot \mathbf{C}_{ij}, \tag{4}$$

where $\mathbf{C}_{ij} = \mathbf{C}(v_a^i, v_s^j) = \frac{v_a^{(i)}\cdot v_s^{(j)}}{\|v_a^{(i)}\|_2 \|v_s^{(j)}\|_2}$.

We define the transport mass $\mathbf{T}$ as

$$\begin{aligned}\sum_{j=1}^{N}\mathbf{T}_{ij} &= \mu_i = \frac{\|v_a^{(i)}\|_2}{\sum_i \|v_a^{(i)}\|_2}\\ \sum_{i=1}^{N}\mathbf{T}_{ij} &= \nu_j = \frac{\|v_s^{(j)}\|_2}{\sum_j \|v_s^{(j)}\|_2}.\end{aligned} \tag{5}$$

Intuitively, OT loss finds an efficient way to "transport mass" from one visual feature distribution to another while minimizing the overall cost. By optimizing this objective, we encourage the two visual feature sets to exhibit structural and semantic alignment, leading to more coherent joint representations when fused with the source text.

Next, to balance contributions from the two visual modality, we apply a CLIP-based VGF. First, we compute the average cosine similarity between the text and each image:

$$\alpha_a = \mathrm{sim}\left(\mathbf{V}_a, \mathbf{t}\right), \quad \alpha_s = \mathrm{sim}\left(\mathbf{V}_s, \mathbf{t}\right), \tag{6}$$

where t denotes the source text x encoded by CLIP text encoder $\phi_t(\cdot)$.

Then, we compute the gate α as

$$\alpha = \sigma\left(\gamma\cdot\left(\alpha_s - \alpha_a - \varepsilon\right)\right), \tag{7}$$

where σ denotes the sigmoid function, and γ, ε are hyperparameters that control the gating coefficient and threshold.

The visual gated fusion is computed as

$$\mathbf{V}_{\text{fused}} = \alpha\cdot\mathbf{V}_s + (1-\alpha)\cdot\mathbf{V}_a. \tag{8}$$

Finally, $\mathbf{V}_{\text{fused}}$ is passed into the Perceiver Resampler to extract compact visual tokens, allowing the model to adaptively prefer the most relevant visual content during translation.

3.4 Contemplation

While the **Imagination** module encourages visual enhancement, it is equally crucial to ensure that the model remains grounded in the source text.

Concretely, we maintain a parallel NMT branch that encodes the source sentence x and produces translation logits independently. Meanwhile, the MMT branch receives the fused visual features along with the same text input. To enforce output consistency, we minimize their KL divergence that serves as a semantic regularizer:

$$\mathcal{L}_{\mathrm{KL}} = \sum_{t=1}^{T}\mathrm{KL}\left(p_{\mathrm{MMT}}\left(y_t \mid x, I\right)\|p_{\mathrm{NMT}}\left(y_t \mid x\right)\right) \tag{9}$$

where $p_{\text{MMT}}(y_t \mid x, I)$ and $p_{\text{NMT}}(y_t \mid x)$ denote the output distributions from the two branches. This constraint regularizes the multimodal decoder to stay close to the text-only distribution, reducing over-reliance on potentially noisy or redundant visual signals.

Beyond distributional consistency, we further introduce a semantic alignment loss that bridges the generated target text with the visual features. Specifically, we apply an image-text contrastive loss (ITC) between the decoder output and the fused vision representation:

$$\mathcal{L}_{\text{ITC}} = -\frac{1}{B}\sum_{i=1}^{B}\left[\log\frac{\exp(\cos(\bar{\mathbf{v}}_i, \mathbf{t}_i)/\tau)}{\sum_{j=1}^{B}\exp(\cos(\bar{\mathbf{v}}_i, \mathbf{t}_j)/\tau)} + \log\frac{\exp(\cos(\mathbf{t}_i, \bar{\mathbf{v}}_i)/\tau)}{\sum_{j=1}^{B}\exp(\cos(\mathbf{t}_i, \bar{\mathbf{v}}_j)/\tau)}\right], \quad (10)$$

where B represents the size of each Batch, $\hat{v}_i$ represents the visual features after Perceiver Resampler, t_i represents the generated text features and τ is a temperature hyperparameter. This contrastive supervision ensures that the translated sentence preserves alignment with the grounded visual semantics.

3.5 Training Objective

Our final training objective combines the standard translation loss with the imagination and contemplation modules as

$$\mathcal{L}_{\text{total}} = \frac{1}{2}(\mathcal{L}_{\text{MMT}} + \mathcal{L}_{\text{NMT}}) + \lambda_{\text{OT}}\mathcal{L}_{\text{OT}} + \lambda_{\text{KL}}\mathcal{L}_{\text{KL}} + \lambda_{\text{ITC}}\mathcal{L}_{\text{ITC}}, \quad (11)$$

where $\lambda_{\text{OT}}, \lambda_{\text{KL}}, \lambda_{\text{ITC}}$ are the hyperparameters that balance between vision-driven augmentation and language-based grounding, resulting in more accurate and robust MMT.

4 Experiments

4.1 Datasets and Training Setting

Datasets: We conduct experiments on four MMT benchmarks: Multi30k (Section 4.4) (Elliott et al., 2016), AmbigCaps (Section 4.5) (Li et al., 2021), 3AM (Section 4.6) (Ma et al., 2024) and CoMMuTE (Appendix C.1) (Futeral et al., 2023). Details of them are in Appendix A.

Training Setting: For the image encoder, we use the CLIP-ViT model [1], where we freeze all its parameters.

For the text encoder and decoder, we conduct experiments under two settings: traditional models and pre-trained models. Details of them are in Appendix B.

4.2 Comparing Systems

Similar to the training setting, we use two types of baseline methods.

(i) **Traditional MMT models**, including Transformer (Vaswani et al., 2017), ImagiT (Long et al., 2021), Selective Attention (Li et al., 2022a), VALHALLA (Li et al., 2022c), IVA-MMT (Ji et al., 2022), SAMMT (Guo et al., 2023), RG-MMT-EDC (Tayir and Li, 2024b) and VisTFC (Zhu et al., 2024). All of them have no pre-trained models and their specific configurations and tokenization rules are in Appendix B.1.

(ii) **Pre-trained MMT models**, including VGAMT (Futeral et al., 2023) and CLIPTrans (Gupta et al., 2023), which use pre-trained text encoder and decoder. There are also text-based LLMs like Qwen2.5 (Qwen et al., 2025), Llama3 (Grattafiori et al., 2024) and Alpaca-7B (Bommasani et al., 2022) and text-image-based LLMs like IMAGE (Chen et al., 2025), GPT-4o (OpenAI et al., 2024) and Qwen-VL (Bai et al., 2023).

4.3 Evaluation Metrics

We evaluate translation performance using three widely adopted metrics: BLEU [2] (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and COMET [3] (Rei et al., 2020). BLEU measures n-gram overlap between the generated and reference translations, emphasizing surface-level fluency. METEOR complements BLEU by considering synonymy, stemming, and word order, providing a more nuanced assessment of semantic accuracy. COMET is a neural metric based on pre-trained multilingual encoders and human judgments, offering a stronger correlation with translation quality.

4.4 Results on Multi30k

Table 1 presents the main results on the Multi30k dataset across three test sets in En-De and En-Fr translation directions.We categorize the compared models into two groups based on whether they employ pre-trained encoders or decoders. Among traditional MMT models, our method consistently outperforms previous approaches in most metrics

[1] https://huggingface.co/openai/clip-vit-large-patch14

[2] https://github.com/mjpost/sacrebleu

[3] https://huggingface.co/Unbabel/wmt22-comet-da

Traditional MMT Models												
Models	*English-to-German*						*English-to-French*					
	Test2016		Test2017		MSCOCO		Test2016		Test2017		MSCOCO	
	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
Transformer†	41.02	68.22	33.36	62.05	29.88	56.64	61.80	81.08	53.46	75.62	44.52	69.43
ImagiT⋆	38.50	55.70	32.10	52.40	28.70	48.80	59.70	74.00	52.40	68.30	45.30	65.00
Selection Attention†	42.50	68.81	34.28	61.81	29.59	56.36	62.79	81.75	55.44	76.57	45.27	70.73
VALHALLA⋆	42.60	69.30	35.10	62.80	30.70	57.60	63.10	81.80	56.00	77.10	46.40	71.30
SAMMT⋆	42.50	-	36.04	-	31.95	-	63.71	-	56.17	-	46.63	-
RG-MMT-EDC⋆	42.00	60.20	33.40	53.70	30.00	49.60	62.90	77.20	55.80	72.00	45.10	64.90
VisTFC⋆	43.10	70.60	35.70	64.00	31.60	58.20	63.30	82.60	56.10	77.10	46.50	72.60
ImCo-MMT(Ours)	43.34	70.16	36.08	63.69	32.30	58.67	63.81	82.13	56.75	77.50	47.28	71.86
Pre-trained MMT Models												
Models	*English-to-German*						*English-to-French*					
	Test2016		Test2017		MSCOCO		Test2016		Test2017		MSCOCO	
	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
VGAMT⋆	43.3	69.4	38.3	65.3	35.7	54.4	67.2	**96.8**	61.6	**92.1**	51.1	81.1
CLIP-Trans⋆	43.87	-	37.22	-	34.49	-	64.55	-	57.59	-	48.83	-
LLaMA3-8B◇	30.1	69.5	24.2	66.4	21.9	62.6	50.2	77.8	40.4	72.8	34.5	70.7
Alpaca-7B◇	38.5	77.2	34.3	76.5	30.9	72.4	59.2	82.5	51.4	79.4	42.6	77.2
IMAGE◇	45.3	83.1	38.6	81.9	37.5	78.8	67.5	88.3	61.5	86.6	49.3	82.5
Qwen-VL♠	25.21	71.44	23.62	70.07	19.67	66.00	30.45	74.22	29.11	72.35	26.87	69.87
GPT-4o♠	41.80	74.82	38.88	74.90	33.52	70.12	61.64	83.21	55.76	80.42	50.11	77.89
Qwen2.5-0.5B♠	87.58	86.35	83.84	84.94	86.74	85.52	80.18	83.60	77.92	80.36	77.13	80.96
LLaMA3.2-1B♠	87.89	86.64	84.05	85.06	86.78	85.57	79.95	84.15	78.24	81.14	77.65	80.88
ImCo-MMT(Ours)	**90.58**	**88.93**	**87.04**	**87.61**	**89.62**	**88.10**	**82.68**	85.12	**80.11**	84.93	**81.08**	**83.83**

Table 1: Main translation results from the Multi30k, with BLEU, METEOR (for Traditional MMT) and COMET (for Pre-trained MMT). We use underline and **bold** to indicate the best results in Traditional MMT and Pre-trained MMT, respectively. † indicates results from Li et al. (2022a), ◇ from Chen et al. (2025), ⋆ from the corresponding papers and ♠ indicates the results are from our implementation. Among ♠, Qwen-VL and GPT-4o are no fine-tuning in Multi30k while others are fine-funing.

Models	*Turkish-to-English*		
	BLEU	METEOR	COMET
Traditional MMT Models			
Transformer♠	36.29	66.97	33.92
Gated Fusion♠	41.81	70.74	42.83
IVA-MMA⋆	39.40	70.22	-
ImCO-MMT(Ours)	42.45	71.11	44.62
Pre-trained MMT Models			
ImCO-MMT(Ours)	**98.13**	**99.06**	**94.33**

Table 2: Results on AmbigCaps of Tr-En translation. ♠ indicates the results are from our implementation, and ⋆ indicates the results come from the corresponding published papers.

Models	*English-to-Chinese*		
	BLEU	METEOR	COMET
Traditional MMT Models			
Transformer▽	11.33	31.34	-
Selective Attention▽	13.33	33.47	-
ImCO-MMT(Ours)	16.69	35.11	10.34
Pre-trained MMT Models			
BART▽	31.47	55.62	-
VL-BART▽	33.27	55.84	-
T5▽	33.09	57.26	-
VL-T5▽	34.24	**59.12**	-
ImCO-MMT(Ours)	**92.51**	50.70	**96.09**

Table 3: Results on 3AM of En-Zh translation. ▽ indicates the results are from Ma et al. (2024).

and test sets, demonstrating the effectiveness of visual fusion strategy and the benefits of incorporating textual constraints for semantic alignment.

Compared with several strong pre-trained MMT models, our pre-trained version of ImCo-MMT yields substantial and consistent gains, achieving the best BLEU on multiple benchmarks. These results highlight the scalability and robustness of our approach even when built upon large pre-trained language models. Notably, compared with strong baselines such as Qwen2.5-0.5B and LLaMA3.2-1B, ImCo-MMT shows marked improvements, especially in COMET scores, suggesting superior semantic adequacy and alignment. Overall, the results strongly validate the advantage of combining learned visual representations with controlled fusion mechanisms and textual constraints in MMT.

4.5 Results on AmbigCaps

Table 2 reports the Turkish-to-English translation results on the AmbigCaps dataset. Our proposed ImCo-MMT consistently outperforms all traditional MMT baselines, achieving the highest scores across BLEU, METEOR, and COMET. Compared to the strong Gated Fusion, ImCo-MMT shows improvements of +0.64 BLEU and +1.79 COMET, indicating better lexical and semantic

Models	Tranditional MMT			Pre-trained MMT		
	Test2016	Test2017	MSCOCO	Test2016	Test2017	MSCOCO
ImCo-MMT	43.34/70.16/68.11	36.08/63.69/63.42	32.30/58.67/59.53	**90.58/96.59/88.93**	**87.04**/91.62/**87.61**	**89.62/95.88/88.10**
w/o Imagination	43.09/69.80/66.77	35.65/63.51/62.11	31.91/57.93/58.03	89.45/95.47/87.87	85.75/**93.36**/86.69	88.45/94.72/87.18
w/o Contemplation	41.94/68.81/66.01	34.73/62.79/61.45	31.55/58.28/57.17	88.90/94.76/87.40	85.21/92.55/86.39	87.82/94.06/86.68
w/o both	41.26/68.05/65.34	34.34/62.13/60.78	31.34/57.56/56.77	88.24/94.12/86.89	84.37/91.59/85.79	87.54/93.61/86.38

Table 4: Results of the Imagination and Contemplation module on Multi30k of En-De translation, with BLEU/METEOR/COMET metrics.

VGF	OT	Tranditional MMT			Pre-trained MMT		
		Test2016	Test2017	MSCOCO	Test2016	Test2017	MSCOCO
✔	✔	43.34/70.16/68.11	36.08/63.69/63.42	32.30/58.67/59.53	**90.58/96.59/88.93**	**87.04**/91.62/**87.61**	**89.62/95.88/88.10**
✖	✔	43.19/69.80/67.69	35.94/63.01/62.44	32.09/58.44/58.87	90.18/96.19/88.55	86.48/**94.09**/87.24	88.77/95.12/87.66
✔	✖	43.25/70.01/67.95	35.58/63.74/62.78	31.84/58.59/59.01	90.01/95.97/88.49	86.54/94.08/87.33	88.92/95.29/87.79

Table 5: Results of two core components within the Imagination module on Multi30k of En-De translation, with BLEU/METEOR/COMET metrics.

alignment. When leveraging pre-trained encoders and decoders, ImCo-MMT reaches near-saturation performance, significantly surpassing all baselines. These results demonstrate the strong generalization and robustness of the model in ambiguous and morphologically rich source languages such as Turkish.

4.6 Results on 3AM

As shown in Table 3, our ImCo-MMT significantly outperforms all baselines in the En-Zh translation task on the 3AM dataset, especially in the pre-trained setting. In contrast, traditional MMT models perform poorly, likely due to the challenges of Chinese tokenization. ImCo-MMT benefits from both powerful pre-trained models and cross-modal alignment, enabling robust handling distant language pairs. Notably, the relatively low scores of traditional models suggest that translating between linguistically distant pairs like English and Chinese poses a significant challenge for models trained from scratch, underscoring the importance of leveraging pre-trained knowledge for such scenarios.

5 Analysis

5.1 Effect of Imagination and Contemplation

Table 4 shows the ablation results of the Imagination and Contemplation modules on the Multi30k. Removing the Imagination module leads to a noticeable performance drop across all metrics, indicating that synthesized visual information brings meaningful semantic enrichment to the translation process. Similarly, when Contemplation is removed, the performance degradation is more significant, suggesting its crucial role in grounding and filtering visual features with textual guidance. When both are removed, the performance decreases further, confirming that the interplay between imagination and contemplation, balancing vision-driven enhancement and text-based regularization, is essential for achieving optimal MMT quality.

5.2 Impact of Image Generation

Table 5 presents the results of the ablation study on two key components within Imagination: VGF and OT constraints. Disabling VGF leads to a consistent drop across all metrics, indicating that adaptive integration of visual features through CLIP similarity improves semantic alignment and relevance. Similarly, removing the OT loss also causes performance degradation, suggesting that enforcing semantic consistency between authentic and synthetic images helps the model focus on shared visual semantics and reduce noise. When both components are active, the model achieves the best results, demonstrating their complementary roles in building a coherent and informative visual context. Additionally, we statistically analyze CLIP-based cosine similarity between authentic and synthetic images before and after applying OT on the Multi30k training set in Appendix C.6.

5.3 Impact of Textual Constraint

As shown in Table 6, the KL constraint, which is applied between MMT and NMT distributions, plays the primary role in improving overall performance. Removing the KL constraint leads to a more significant performance drop than removing ITC, indicating its central importance in enforcing consistency with strong textual priors and guiding the decoder towards more fluent and reliable linguistic outputs. In contrast, the ITC loss provides auxiliary but less substantial gains by encouraging

KL	ITC	Tranditional MMT			Pre-trained MMT		
		Test2016	Test2017	MSCOCO	Test2016	Test2017	MSCOCO
✔	✔	43.34/70.16/68.11	36.08/63.69/63.42	32.30/58.67/59.53	**90.58/96.59/88.93**	**87.04**/91.62/**87.61**	**89.62/95.88/88.10**
✖	✔	42.57/69.48/67.25	35.02/62.62/62.04	31.34/57.32/58.46	89.43/95.29/87.77	85.80/93.16/86.59	88.40/94.59/87.08
✔	✖	43.01/69.97/67.69	35.86/63.18/62.82	32.03/58.25/59.03	89.72/96.11/88.03	86.17/94.08/86.40	88.73/95.47/87.03

Table 6: Results of two core components within the Contemplation module on Multi30k of En-De translation, with BLEU/METEOR/COMET metrics.


SRC(EN): A man in a red shirt entering an establishment.
REF(DE): Ein mann in einem roten hemd betritt ein etablissement.
MMT-ONLY: Ein mann in einem roten hemd, der ein betrieb miert.
(A man in a red shirt who is painting a company.)
+Ima: Ein mann in einem roten hemd steigt in ein etablissement.
(A man in a red shirt enters a establishment.)
+Con: Ein mann in einem roten hemd betritt ein etablissement.
(A man in a red shirt is entering an establishment.)


SRC(EN): This lady has heard a funny joke and laughing.
REF(DE): Diese frau hat gerade einen lustigen witz gehört und lacht.
MMT-ONLY: Diese dame trägt ein lustiges namensschild und lacht.
(This lady is wearing a funny name tag and laughing.)
+Ima: Diese dame hat gerade eine lustige joke und lacht.
(This lady has heard a funny joke and laughs.)
+Con: Diese frau hat gerade einen lustigen witz gehört und lacht.
(This lady has heard a funny joke and is laughing.)


SRC(EN): A dog playing fetch in the water, holding an orange ball in his mouth.
REF(DE): Ein hund im wasser apportiert einen orangefarbenen ball in seinem maul.
MMT-ONLY: Ein hund spielt mit einem gelben ball im maul ein spiel im wasser.
(A dog plays a game in the water with a yellow ball in its mouth.)
+Ima: Ein hund spielt im wasser und hält einen orangefarbenen ball im maul.
(A dog plays in the water and holds an orange ball in its mouth.)
+Con: Ein hund spielt im wasser apportiert und hält einen orangefarbenen ball im seinem maul.
(A dog plays fetch in the water and holds an orange ball in his mouth.)


SRC(EN): A man with a hat rides his bike along the water.
REF(DE): Ein mann mit helm fährt am wasser entlang fahrrad.
MMT-ONLY: Ein mann mit helm fährt mit einem fahrrad auf dem wasser.
(A man in a helmet rides a bike on the water.)
+Ima: Ein mann mit hut fährt mit seinem fahrrad am wasser entlang.
(A man with a hat rides his bike along the water.)
+Con: Ein mann mit helm fährt mit dem fahrrad am wasser entlang.
(A man with a helmet rides his bike along the water.)

Figure 3: Case study on Test2016 dataset of En-De translation. MMT-ONLY refers to the output of our traditional model without Imagination and Contemplation. The red and green words denote error and correct translations, respectively.

better visual-textual alignment, especially in visually ambiguous scenarios. Together, they form a complementary mechanism, but it is the KL regularization that mainly drives the substantial improvements within the Contemplation module.

5.4 Case Study

As shown in Figure 3, the first example describes a man in a red shirt entering a store. However, the MMT-ONLY output incorrectly translates it as "painting a company", which is caused by the excessive and irrelevant visual information. Although the translation is semantically correct with Imagination, the generated words "steigt in" does not match the reference sentence "betritt". For this reason, we introduce the Contemplation module to align it with the correct text.

In contrast, the baseline of the last example produces an incorrect translation, implying the man is riding "on the water". With Imagination, the model corrects the context, but it mistakenly switches "helmet" to "hat", indicating that hallucinated visual information may introduce factual errors. Even with Contemplation, the hallucinated object persists, as textual grounding alone cannot fully override previously introduced inaccuracies.

These cases show that while images provide valuable contextual information, irrelevant or redundant details can lead to mistranslations that are difficult to correct even with textual constraints, underscoring the necessity for effective regulation and precise alignment between text and image.

6 Conclusion

In this work, we propose ImCo-MMT, a novel MMT framework that balances imaginative visual synthesis and grounded textual contemplation. By introducing two complementary modules: Imagination and Contemplation, we achieve substantial improvements across multiple benchmarks. Our results demonstrate that well-structured multimodal

integration, rather than sheer reliance on large scale backbones, plays a crucial role in advancing translation quality, providing a promising direction for advancing MMT. Future work will explore end-to-end integration of generation and translation.

Limitations

Despite the promising results achieved by ImCoMMT, several limitations remain. First, although the imagination module introduces visually enhanced information through generated images, the quality and semantic alignment of these synthetic images are not always guaranteed, especially for linguistically ambiguous or abstract source sentences. Second, while our contemplation module incorporates cross-modal alignment via ITC and KL losses, it does not explicitly guide the image generation process itself, resulting in a gap between image-text alignment and the actual visual content being synthesized. Third, due to the modular nature of our framework and the reliance on external pre-trained components (e.g., CLIP, diffusion models, large language decoders), training and inference can be computationally demanding, and integration into a fully end-to-end pipeline remains a challenge. In future work, we plan to address these limitations by jointly optimizing image generation and translation, and by exploring lighter-weight yet effective model alternatives.

Acknowledgements

This work is supported by the National Natural Science Foundation of China under Projects 62576206 and 62476089.

References

Haithem Afli, Loïc Barrault, and Holger Schwenk. 2016. Building and using multimodal comparable corpora for machine translation. *Natural Language Engineering*, 22(4):603–625.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, and 1 others. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, and 109 others. 2023. Palm 2 technical report. *Preprint*, arXiv:2305.10403.

Dzmitry Bahdanau. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

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. *Preprint*, arXiv:2308.12966.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. 2022. Beit: Bert pre-training of image transformers. *Preprint*, arXiv:2106.08254.

Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. 2018. Findings of the third shared task on multimodal machine translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 304–323, Belgium, Brussels. Association for Computational Linguistics.

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, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, and 95 others. 2022. On the opportunities and risks of foundation models. *Preprint*, arXiv:2108.07258.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Ozan Caglayan, Menekse Kuyu, Mustafa Sercan Amac, Pranava Madhyastha, Erkut Erdem, Aykut Erdem, and Lucia Specia. 2021. Cross-lingual visual pre-training for multimodal machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1317–1324, Online. Association for Computational Linguistics.

Ozan Caglayan, Pranava Madhyastha, Lucia Specia, and Loïc Barrault. 2019. Probing the need for visual

context in multimodal machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4159–4170, Minneapolis, Minnesota. Association for Computational Linguistics.

Iacer Calixto, Qun Liu, and Nick Campbell. 2017. Doubly-attentive decoder for multi-modal neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1913–1924, Vancouver, Canada. Association for Computational Linguistics.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer.

Andong Chen, Yuchen Song, Kehai Chen, Xuefeng Bai, Muyun Yang, Liqiang Nie, Jie Liu, Tiejun Zhao, and Min Zhang. 2025. Make imagination clearer! stable diffusion-based visual imagination for multimodal machine translation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26567–26583, Vienna, Austria. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, and 1 others. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. *Preprint*, arXiv:2010.11929.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30K: Multilingual English-German image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74, Berlin, Germany. 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)*, pages 5980–5994, Toronto, Canada. 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.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Stig-Arne Grönroos, Benoit Huet, Mikko Kurimo, Jorma Laaksonen, Bernard Merialdo, Phu Pham, Mats Sjöberg, Umut Sulubacak, Jörg Tiedemann, Raphael Troncy, and Raúl Vázquez. 2018. The MeMAD submission to the WMT18 multimodal translation task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 603–611, Belgium, Brussels. Association for Computational Linguistics.

Wenyu Guo, Qingkai Fang, Dong Yu, and Yang Feng. 2023. Bridging the gap between synthetic and authentic images for multimodal machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2863–2874, Singapore. Association for Computational Linguistics.

Devaansh Gupta, Siddhant Kharbanda, Jiawei Zhou, Wanhua Li, Hanspeter Pfister, and Donglai Wei. 2023. Cliptrans: transferring visual knowledge with pre-trained models for multimodal machine translation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2875–2886.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

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 (Volume 1: Long Papers)*, pages 2399–2409, Berlin, Germany. Association for Computational Linguistics.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.

Baijun Ji, Tong Zhang, Yicheng Zou, Bojie Hu, and Si Shen. 2022. Increasing visual awareness in multimodal neural machine translation from an information theoretic perspective. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6755–6764, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization. *Preprint*, arXiv:1412.6980.

Chiraag Lala, Pranava Swaroop Madhyastha, Carolina Scarton, and Lucia Specia. 2018. Sheffield submissions for wmt18 multimodal translation shared task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 624–631.

Bei Li, Chuanhao Lv, Zefan Zhou, Tao Zhou, Tong Xiao, Anxiang Ma, and JingBo Zhu. 2022a. On vision features in multimodal machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6327–6337, Dublin, Ireland. Association for Computational Linguistics.

Jiaoda Li, Duygu Ataman, and Rico Sennrich. 2021. Vision matters when it should: Sanity checking multimodal machine translation models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8556–8562, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022b. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Yi Li, Rameswar Panda, Yoon Kim, Chun-Fu Richard Chen, Rogerio S Feris, David Cox, and Nuno Vasconcelos. 2022c. Valhalla: Visual hallucination for machine translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5216–5226.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022.

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*, pages 5738–5748, Online. Association for Computational Linguistics.

Xinyu Ma, Xuebo Liu, Derek F. Wong, Jun Rao, Bei Li, Liang Ding, Lidia S. Chao, Dacheng Tao, and Min Zhang. 2024. 3AM: An ambiguity-aware multimodal machine translation dataset. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1–13, Torino, Italia. ELRA and ICCL.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *Preprint*, arXiv:2307.01952.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, and 1 others. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *Preprint*, arXiv:2204.06125.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695.

Lucia Specia, Stella Frank, Khalil Sima'an, and Desmond Elliott. 2016. A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 543–553, Berlin, Germany. Association for Computational Linguistics.

Turghun Tayir and Lin Li. 2024a. Unsupervised multimodal machine translation for low-resource distant language pairs. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 23(4):1–22.

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

Turghun Tayir, Lin Li, Bei Li, Jianquan Liu, and Kong Aik Lee. 2024. Encoder-decoder calibration for multimodal machine translation. *IEEE Transactions on Artificial Intelligence*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Vipin Vijayan, Braeden Bowen, Scott Grigsby, Timothy Anderson, and Jeremy Gwinnup. 2024. Adding multimodal capabilities to a text-only translation model. In *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 14–28, Chicago, USA. Association for Machine Translation in the Americas.

Yusong Wang, Dongyuan Li, Jialun Shen, Yicheng Xu, Mingkun Xu, Kotaro Funakoshi, and Manabu Okumura. 2024. LAMBDA: Large language model-based data augmentation for multi-modal machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15240–15253, Miami, Florida, USA. Association for Computational Linguistics.

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 (Volume 1: Long Papers)*, pages 6153–6166, Online. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021a. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021b. mt5: A massively multilingual pre-trained text-to-text transformer. *Preprint*, arXiv:2010.11934.

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*, pages 4346–4350, Online. Association for Computational Linguistics.

Zhuang Yu, Shiliang Sun, Jing Zhao, Tengfei Song, and Hao Yang. 2025. Memory reviving, continuing learning and beyond: Evaluation of pre-trained encoders and decoders for multimodal machine translation. *Preprint*, arXiv:2504.18012.

Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. 2014. Recurrent neural network regularization. *CoRR*, abs/1409.2329.

Shaolin Zhu, Shangjie Li, and Deyi Xiong. 2024. Vistfc: Vision-guided target-side future context learning for neural machine translation. *Expert Systems with Applications*, 249:123411.

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*, pages 2679–2697, Toronto, Canada. Association for Computational Linguistics.

Yuxin Zuo, Bei Li, Chuanhao Lv, Tong Zheng, Tong Xiao, and JingBo Zhu. 2023. Incorporating probing signals into multimodal machine translation via visual question-answering pairs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14689–14701, Singapore. Association for Computational Linguistics.

A Datasets Details

Multi30k (Elliott et al., 2016) mainly contains two translation directions: English-German and English-French, including 31014 images with English captions and German and French translations. The training set and validation set contain 29,000 and 1,014 instances, respectively. We reported the results on the Test2016, Test2017 and MSCOCO test sets, which contain 1000, 1000 and 461 instances, respectively.

AmbigCaps (Li et al., 2021) is an English-Turkish dataset, in which the number of samples in the training set, validation set, and test set are 89601, 1000, and 1000 respectively.

3AM (Ma et al., 2024) is an English-Chinese MMT data set that contains a wide range of visual concepts and fuzzy data. Its samples come from some existing image + text data sets, and finally form a training set containing 23,931 samples, a validation set of 1,000 samples, and a test set of 1,000 samples.

CoMMuTE (Futeral et al., 2023) is a contrastive multilingual multimodal translation evaluation dataset, which covers several translation directions: English-French, English-German, English-Czech, etc. From above, we test test our model in three translation directions: English-German, English-French, and English-Chinese. It consists of 155 lexically ambiguous English sentences, each with two translations and two possible meanings, and two pictures to determine which translation is correct.

B Training Setting

For the Perceiver Resampler, we use six blocks, that is, $N = 6$ and set 32 queries where each query has a dimension of 768. For the training objective, we set $\lambda_{\text{OT}} = 0.1, \lambda_{\text{KL}} = 0.1, \lambda_{\text{ITC}} = 0.1$.

B.1 Traditional MMT Models

We adopt the standard Transformer architecture as a baseline, which follow Wu et al. (2021) to conduct experiments with Transformer-Tiny configuration. This setup allows us to examine whether the proposed multimodal components can effectively learn from scratch without the support of external knowledge. The model consists of 4 encoder and decoder layers. The hidden size is 128 and the filter size of FFN is 256. There are 4 heads in the multi-head self-attention mechanism. We set the dropout as 0.3 and the label smoothing as 0.1. Besides, we learn a joint BPE code for 10,000 merging operations for both the source and target languages, resulting in vocabularies of 9,716 and 9,548 entries for the En-De and En-Fr tasks, which is the same as the baseline tradition MMT models in Section 4.2.

Our experiments were implemented using the open-source Fairseq (Ott et al., 2019) framework. During training, we set the dropout rate to 0.3 and applied label smoothing with a factor of 0.1. Each training batch consists of 2048 tokens, and the gradient update frequency is set to 5. After experimental analysis on the validation set, the hyper-parameters are set as $\alpha = 0.1, \beta = 3, \gamma = 0.4$. During decoding, the beam size is set to 5. For optimization, we used the Adam optimizer (Kingma and Ba, 2017) with $\beta_1 = 0.9, \beta_2 = 0.98$ and $\epsilon = 10^{-8}$. In terms of learning rate scheduling, we employed a warmup period of 2000 steps, starting from $1e^{-7}$ and gradually increasing to $5e^{-3}$. All models are trained and evaluated using two 4090 GPUs.

B.2 Pre-trained MMT Models

Yu et al. (2025) have revealed a modality-dependent asymmetry: pre-trained decoders offer stable improvements in generation quality, while the benefit of pre-trained encoders strongly depends on the degree of visual-textual alignment. Thus, we incorporate several widely used pre-trained models as encoders or decoders of our framework, such as mBART (Liu et al., 2020), mT5 (Xue et al., 2021b), Qwen (Qwen et al., 2025) and LLaMA (Grattafiori et al., 2024).

We select mBART and mT5 as our encoders, and Qwen and LLaMA as our decoders. The specific experimental results are Appendix C.3. In fact, we found that different pre-trained encoders and decoders have little impact on the final translation

Models	*English-to-German*				*English-to-French*				*English-to-Chinese*			
	Multi30k(train)								3AM(train)			
	BLEU/METEOR/COMET/Accuracy											
Pre-trained MMT Models												
VGAMT	29.3	43.0	18.4	**59.0**	32.2	48.5	36.2	**67.1**	-	-	-	-
ImCo-MMT	**77.72**	**83.14**	**77.96**	50.3	**80.87**	**85.11**	**79.26**	51.2	**97.06**	**44.68**	**96.52**	**50.1**

Table 7: Results on CoMMuTE. Since this dataset does not have a training set, we have to train it on other datasets and then test on it.

performance, so the model shown in the text is a combination of mBART+Qwen. The model is trained on two 4090 GPUs for 5 epochs with a batch size of 8, a peak learning rate of 1e-5 with 0.1 warmup ratio.

B.3 Text2Image Generation Models

For the text-to-image generation, we choose Stable Diffusion XL (SDXL) (Podell et al., 2023) to synthesize images on two 4090 GPUs, which introduces a refiner module to improve the quality of generated images. According to official recommendations, the size of the image we generated is 1024 × 1024. During the generation process, we set the number of inferences to 100, the proportion of the high-frequency part of the noise to 0.5, and the guidance scale to 7.5.

C Additional Results

C.1 Results on CoMMuTE

To evaluate the effectiveness of the framework in resolving translation ambiguities, we conduct experiments within the CoMMuTE (Futeral et al., 2023) benchmark, which requires high-quality annotated images to resolve ambiguities in translation.

The Accuracy metric is a custom metric in the CoMMuTE dataset. For each example, there are correct translation results and incorrect translation results. Its calculation is the ratio of the number of samples where the perplexity of the correct translation result is greater than that of the incorrect result to the total number of samples. However, in fact, when we tested it, we found that the fluctuation range of this metric was very large, so it is only of reference significance.

Table 7 presents results on the CoMMuTE test set, using models trained on Multi30k and 3AM. Our ImCo-MMT achieves substantial improvements across all language pairs, with particularly large margins in COMET and BLEU scores. This indicates strong generalization to unseen multimodal contexts. While VGAMT shows higher accuracy, its BLEU and COMET are much lower, suggesting limited semantic alignment. ImCo-MMT benefits from cross-modal contrastive learning and powerful pre-trained components, enabling better transfer performance despite domain shift.

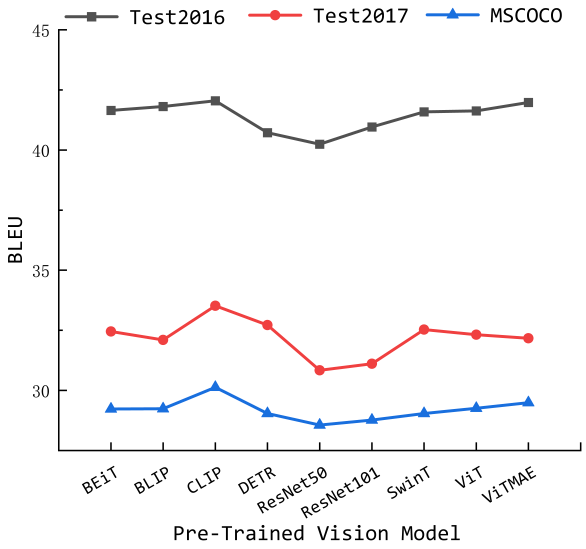


Figure 4: BLEU values under different pre-trained vision models.

C.2 Choice of Vision Model

Using stronger vision models as image encoders can often extract better visual feature information (Zuo et al., 2023). Li et al. (2022a) has demonstrated that more robust pre-trained visual models can significantly enhance MMT models. In order to investigate which pre-trained vision models has the greatest improvement effect on our model, we conduct experiments with several widely used pre-trained visual models, including BEiT (Bao et al., 2022), BLIP (Li et al., 2022b), BLIP2 (Li et al., 2023), DETR (Carion et al., 2020), ResNet (He et al., 2016), SwinT (Liu et al., 2021), ViT (Dosovitskiy et al., 2021), ViT-MAE (He et al., 2022). Figure 4 shows the BLEU values of different models on the three En-De test sets of Multi30k. We can see that CLIP exhibits significant performance

Models	*English-to-German*								
	Test2016			Test2017			MSCOCO		
	BLEU/METEOR/COMET								
mBART-Large + Qwen2.5-0.5B	88.24	94.12	86.89	84.37	91.59	85.79	87.54	93.61	86.38
mT5-Base + Qwen2.5-0.5B	89.43	95.29	87.77	85.80	93.16	86.59	88.40	94.59	87.08
mBART-Large + Llama3.2-1B	**89.72**	**96.11**	88.03	**86.17**	**94.08**	86.40	**88.73**	**95.47**	87.03
mT5-Base + Llama3.2-1B	89.54	95.52	**88.06**	85.96	93.52	**86.97**	88.36	94.72	**87.11**

Table 8: Main translation results from the Multi30k benchmark with different pre-trained models. We use mT5-Base with 580M parameters, mBART-Large with 610M parameters, Qwen2.5-0.5B and Llama-3.2-1B.

AUT IMG	SYN IMG	*English-to-German*								
		Test2016			Test2017			MSCOCO		
		BLEU/METEOR/COMET								
✔	✖	40.81	67.23	65.55	33.63	60.46	59.24	31.08	55.93	57.25
✖	✔	40.15	66.96	65.68	32.87	60.76	59.46	30.75	56.28	57.10
✔	✔	**41.88**	**68.75**	**65.95**	**34.20**	**61.81**	**60.75**	**31.76**	**57.34**	**57.67**

Table 9: Results of different visual modalities on translation. AUT IMG represents authentic images while SYN IMG represents synthetic images.

improvement due to its powerful cross-modal modeling knowledge, which is specifically optimized for cross-modal understanding to maximize the benefits in MMT tasks.

C.3 Impact of Different Pre-trained Models

Table 8 explores the impact of various combinations of pre-trained encoder-decoder models on the English-to-German translation task.

While there are slight performance variations across different model pairs, the overall differences remain marginal, suggesting that the proposed framework is relatively robust to the choice of backbone models. This further suggests that the gains in performance are primarily attributed to our multimodal design and training strategies rather than the specific pre-trained components employed. Moreover, given the relatively small size of the Multi30k dataset and its predominantly simple sentence structures, fine-tuning large pre-trained models on this dataset can easily lead to performance saturation. For consistency and efficiency, we adopt mBART-Large + Qwen2.5-0.5B as our default baseline in all experiments.

C.4 Impact of Different Text-to-Image Models

As shown in table 10, we have tested the image generation quality of the Stable Diffusion series, including SD1.4, SD1.5, and SDXL. Qualitatively, SDXL generated higher-quality images than other versions. And despite the visual improvements in SDXL, performance differences across models remained minor. Thus, we believe that fine-tuning T2I models for this specific MMT task, rather than merely switching to stronger generic generators, is a more promising direction.

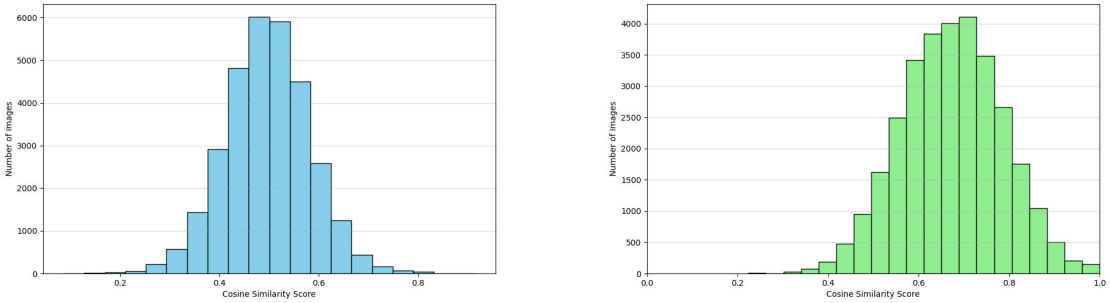
Models	*English-to-German*		
	Test2016	Test2017	MSCOCO
	BLEU/METEOR		
SD1.4	43.20/69.64	35.58/63.74	**32.39**/58.31
SD1.5	**43.39**/69.80	35.81/63.30	32.03/58.66
SDXL	43.34/**70.16**	**36.08/63.69**	32.30/**58.67**

Table 10: Main translation results from Multi30k benchmark with the Stable Diffusion series.

C.5 How Visual Modality Affect Translation?

To further investigate this aspect, we carried out additional experiments comparing different image modalities in the En-De direction, namely (1) using only authentic images, (2) using only synthetic images, and (3) using both visual modalities. Results are reported on table 9.

These experiments were conducted without incorporating the Contemplation module. Interestingly, we observed that combining both visual modalities yields better performance compared to using either one alone. However, when considered alongside the main results presented in the paper, we found that introducing the KL constraint (Contemplation) tends to attenuate this effect.



(a) Before OT constraint.

(b) After OT constraint.

Figure 5: Cosine similarity between authentic images and synthetic images.

C.6 Visual Semantic Consistency

As shown in Figure 5, to assess the effect of OT constraints on aligning visual semantics between authentic and synthetic images, we analyze the distribution of their CLIP-based cosine similarities on the Multi30k training set. Without OT, the similarity scores follow a Gaussian-like distribution centered around 0.5, indicating frequent semantic drift between the two modalities. After applying the OT constraint, the distribution shifts toward a higher mean of approximately 0.7, with reduced variance. This suggests that OT effectively pulls semantically corresponding image pairs closer in the feature space, promoting better cross-modal alignment. In practice, this enforces semantic consistency while filtering out specific modality noise, particularly from visually irrelevant or overly stylized content in the synthetic images. These findings validate the role of OT as a regularizer that refines multimodal fusion by anchoring synthetic features to authentic semantics.