Multimodal Machine Translation with Text-Image In-depth Questioning

Yue Gao[1], Jing Zhao[1*], Shiliang Sun[2*],
Xiaosong Qiao[3], Tengfei Song[3], Hao Yang[3]
[1]School of Computer Science, East China Normal University, Shanghai, China
[2]Department of Automation, Shanghai Jiao Tong University, Shanghai, China
[3]2012 Labs, Huawei Technologies CO., LTD, China

51265901100@stu.ecnu.edu.cn, jzhao@cs.ecnu.edu.cn, shiliangsun@gmail.com,
{qiaoxiaosong, songtengfei2, yanghao30}@huawei.com

Abstract

Multimodal machine translation (MMT) integrates visual information to address ambiguity and contextual limitations in neural machine translation (NMT). Some empirical studies have revealed that many MMT models underutilize visual data during translation. They attempt to enhance cross-modal interactions to enable better exploitation of visual data. However, they only focus on simple interactions between nouns in text and corresponding entities in image, overlooking global semantic alignment, particularly for prepositional phrases and verbs in text which are more likely to be translated incorrectly. To address this, we design a Text-Image In-depth Questioning method to deepen interactions and optimize translations. Furthermore, to mitigate errors arising from contextually irrelevant image noise, we propose a Consistency Constraint strategy to improve our approach's robustness. Our approach achieves state-of-the-art results on five translation directions of Multi30K and AmbigCaps, with +2.35 BLEU on the challenging MSCOCO benchmark, validating our method's effectiveness in utilizing visual data and capturing comprehensive textual semantics.

1 Introduction

Multimodal machine translation (MMT) utilizes modalities beyond text, especially visual data, to clarify ambiguous words and supplement incomplete contexts, thereby improving translation quality (Huang et al., 2016; Yao and Wan, 2020; Li et al., 2021b; Guo et al., 2024). Some MMT studies focus on the extraction of visual information, which are devoted to extract text-related visual features for subsequent translation (Yao and Wan, 2020; Ye and Guo, 2022; Lin et al., 2020; Caglayan et al., 2021). Other studies expect to enhance cross-modal fusion and alignment to improve translation quality (Ye et al., 2022; Tayir et al., 2024; Nishihara et al., 2020; Ye et al., 2023).

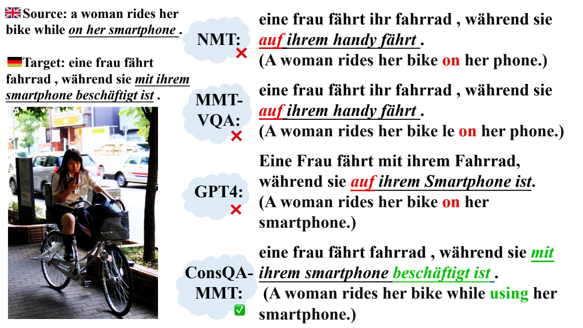

Figure 1: ***"On her smartphone"*** in English conveys ***"busy with her phone"*** in this scenario. The first three approaches directly translate the preposition ***"on"*** to ***"auf"*** in German, inaccurately shifting the meaning to ***"located above her phone."*** Our method effectively leverages image information to capture correct semantics and produce a precise translation.

Later, people found no difference in translation performance of the above models with relevant or irrelevant images and they questioned that many MMT models' utilization of images is inadequate (Wu et al., 2021; Yang et al., 2020; Barrault et al., 2018; Elliott, 2018). Zuo et al. (2023) and Li et al. (2022a) attributed this to insufficient interaction between two modalities, opening up a new research direction. For instance, Guo et al. (2023a) introduced a progressive transformer model with modality difference awareness to enhance the visual-textual interaction. Futeral et al. (2023) designed a dual-objective framework, jointly training MMT and visually conditioned masked language modeling, allowing better utilization of images in translation. Zuo et al. (2023) converted the masked source text into question-answering pairs and jointly trained MMT and Visual Question-Answering (VQA) to strengthen cross-modal interaction, enforcing their model to be sensitive to visual information.

However, these works focus on simple interactions between specific types of words (**Nouns**, **Characters**, **Colors**) and corresponding entities in the image, while accurate translation often requires

Findings of the Association for Computational Linguistics: ACL 2025, pages 9274–9287
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

attention to overall semantics and contexts. As shown in Figure 1, the phrase "on her phone" in this context means "using her phone" but NMT translates it literally into "auf ihrem Handy" which can only mean "located above her phone" in German. NMT translates incorrectly due to the incomplete context of source text. When paired with an image showing "the woman is using her phone," previous MMT methods still fail to get correct translation. According to their methods, the model can only focus on the interaction and alignment of **"woman" "bike" "smartphone"** in this example, lacking a deeper understanding of the relationships between entities. To address this, we design a Text-Image In-depth Questioning method, leveraging the extensive knowledge and understanding capabilities of Large Language Model (LLM). For example in Figure 1, the question is ***"What is the woman doing in addition to riding her bike?"*** and the answer is ***"using her phone"***. Our approach enhances the depth of text-image interaction, providing the correct translation result.

Furthermore, due to the limitations of visual information or the misalignment between text and images, the answers may not be derived from images. Forcing a strict question-answering loss will interfere with model training and negatively impact translation. To address this, we propose a Consistency Constraint strategy that relaxes the strict requirement for giving correct answers, focusing on aligning image-to-source and image-to-target semantic distances. The semantic distances are measured by the accuracy of the question-answering task.

Our main contributions can be summarized as follows:

- We design a Text-Image In-depth Questioning method to strengthen the interaction between image and text so that our model can use comprehensive image information for a high-quality translation.

- We propose a Consistency Constraint strategy which aligns the image-to-source and image-to-target semantic distances, improving overall translation accuracy by reducing the negative impact of irrelevant images.

- Experimental results show that our approach effectively probes image information and accurately captures textual semantics, achieving the new state-of-the-art performance.

2 Related Work

MMT has emerged as a rapidly growing research domain (Elliott et al., 2016). Early investigations in this field focused on three key areas: text-aware visual feature extraction techniques (Yao and Wan, 2020; Zhao et al., 2020; Lin et al., 2020; Ye and Guo, 2022; Fang and Feng, 2022), effective cross-modal representation learning (Yin et al., 2020; Caglayan et al., 2021; Fei et al., 2023; Zhao et al., 2022), and enhancing model robustness through mitigating noise propagation from irrelevant visual inputs (Ye et al., 2022; Huang et al., 2023), error accumulation in cross-modal fusion (Calixto et al., 2019; Tayir et al., 2024), and semantic drift during bilingual decoding (Nishihara et al., 2020; Ye et al., 2023; Liu et al., 2024). However, Elliott (2018) conducted adversarial experiments and revealed that replacing irrelevant images had no significant impact on translation results. Wu et al. (2021) further pointed out that the performance improvement of multimodal models might be due to regularization effects rather than the utilization of visual information. Li et al. (2022a) designed an entity mask probing task, demonstrating that image has little effect on the translation when text is complete. Long et al. (2024) suggested that visual information serves a supplementary role in MMT and can be substituted.

To effectively leverage the auxiliary role of visual modalities in MMT, Ive et al. (2019) proposed using deliberation networks and structured visual information to generate draft translations, which are then refined based on the visual context. Yang et al. (2020) and Su et al. (2021) enhanced cross-modal interaction and visual attention through mechanisms like bidirectional and co-attention networks. Guo et al. (2023b) introduced a modality difference-aware module that progressively fuses visual features layer by layer to bridge modality gaps. Hatami et al. (2024) focused on ambiguous sentences that benefit from visual cues, enhancing both multimodal and text-only translation approaches. Futeral et al. (2023) proposed a new framework with lightweight adapters and guided self-attention, jointly training MMT and visual masked language models. Zuo et al. (2023) converted the text, previously masked in the detection task, into question-answer pairs and trained MMT with VQA. Bowen et al. (2024) proposed a MMT architecture named GRAM to train models on multimodal datasets with masked sentences, improving visual context utilization. Recently, an emerging paradigm (Long et al., 2021; Li et al., 2022b; Zhu et al., 2023) employed text-conditioned generative or retrieval models to synthesize and integrate visual features into translation process.

These studies laid the foundation for image-text alignment in MMT, while our work advances cross-modal interaction mechanisms to enhance translation accuracy and robustness.

3 Method

In this section, we first introduce the framework and the basic loss function implemented in our MMT task (§3.1). Next, we introduce our proposed ConsQA-MMT, shown in Figure 2, which features the Text-Image In-depth Questioning method (§3.2) and the Consistency Constraint strategy (§3.3). Finally, we outline our method's overall training objective (§3.4).

3.1 Framework and Losses for MMT

MMT uses visual information to improve the translation quality between two textual languages. The input includes source text $x = (x_1, x_2, \dots, x_n)$ where x_i is the i-th word, and an image v for supplementing context and eliminating ambiguity. The output is the translated sentence $\hat{\mathbf{y}} = (\hat{y_1}, \hat{y_2}, \dots, \hat{y_m})$ in the target language.

Our approach adopts an encoder-decoder architecture which is widely used in current multimodal learning. In the **encoder** phase, a traditional four-layer Transformer encoder is used to obtain text representation h_x. For the image v, a pre-trained vision model MAE (He et al., 2022) is utilized to obtain the visual representation h_v, which has been proven to achieve significant success in image encoding tasks (Zuo et al., 2023). The encoder processes are fomulated by

$$h_x = \mathbf{TextEncoder}(x), \tag{1}$$

$$h_v = \mathbf{ImageEncoder}(v). \tag{2}$$

Then, we use a single-head attention to obtain image representation $h_{\mathbf{attn}}$ that are correlated with text, where the query derives from h_x, and the key and value are both h_v:

$$h_{\mathbf{attn}} = \text{Softmax}\left(\frac{Q(h_x)K(h_v)^\top}{\sqrt{d_k}}\right) V(h_v), \tag{3}$$

where d_k is the dimension of h_x or h_v.

After that, we use a cross-modal gated fusion mechanism, to produce the text-image joint representation H. In contrast to previous approaches, we employ the hyperbolic tangent (tanh) activation function instead of the sigmoid function. The fusion mechanism is expressed by

$$H = h_x + \lambda h_{\mathbf{attn}}, \tag{4}$$

$$\lambda = \text{Tanh}(W_1 h_x + W_2 h_v), \tag{5}$$

where W_1 and W_2 are learnable variables, and λ controls the mixing ratio of visual information.

In the **decoder** phase, H is decoded to produce the target sentence $\hat{y}$. The model is trained end-to-end on parallel corpora containing textual and visual data, optimizing the loss between the predicted translation $\hat{y}$ and the ground truth y. The MMT loss is formally defined as:

$$\mathcal{L}_{\mathcal{MMT}} = -\sum\nolimits_{i=1}^{|y|} \log p(y_i|y_{<i}, x, v). \tag{6}$$

To mitigate the risk of over-reliance on visual inputs, we integrate the NMT loss into the training process and employ a KL divergence term to regularize the difference between the NMT and MMT losses, and thus obtain the following loss:

$$\mathcal{L}_{\mathcal{MT}} = \frac{\mathcal{L}_{\mathcal{MMT}} + \mathcal{L}_{\mathcal{NMT}}}{2} + \mathcal{L}_{\mathcal{KL}}, \tag{7}$$

$$\mathcal{L}_{\mathcal{KL}} = \sum_{i=1}^{|y|} \mathbf{KL}\left[p(y_i|y_{<i}, x) \,\|\, p(y_i|y_{<i}, x, v)\right]. \tag{8}$$

The KL divergence term also ensures the model maintains strong generalization and translation performance, even when visual information is unavailable.

3.2 Text-Image In-depth Questioning

To improve the utilization of visual information in translation, we integrate Visual Question Answering (VQA) as an auxiliary task, jointly training it with MMT. In our work, the VQA task is built to receive questions from textual modality, and uses visual information to predict answers. The VQA task heavily relies on visual information and requires a deeper understanding of images and complex reasoning (Amara et al., 2024; Antol et al., 2015). By leveraging a multi-task learning strategy, VQA can fill the gap in MMT's visual information processing and foster better cross-modal interaction.

A key challenge in joint training is the lack of a suitable dataset that can align VQA's deep image analysis with MMT's translation needs. MMT training data consists of (source language, image, target language) triples, while VQA training data is structured as (question, image, answer) pairs. The

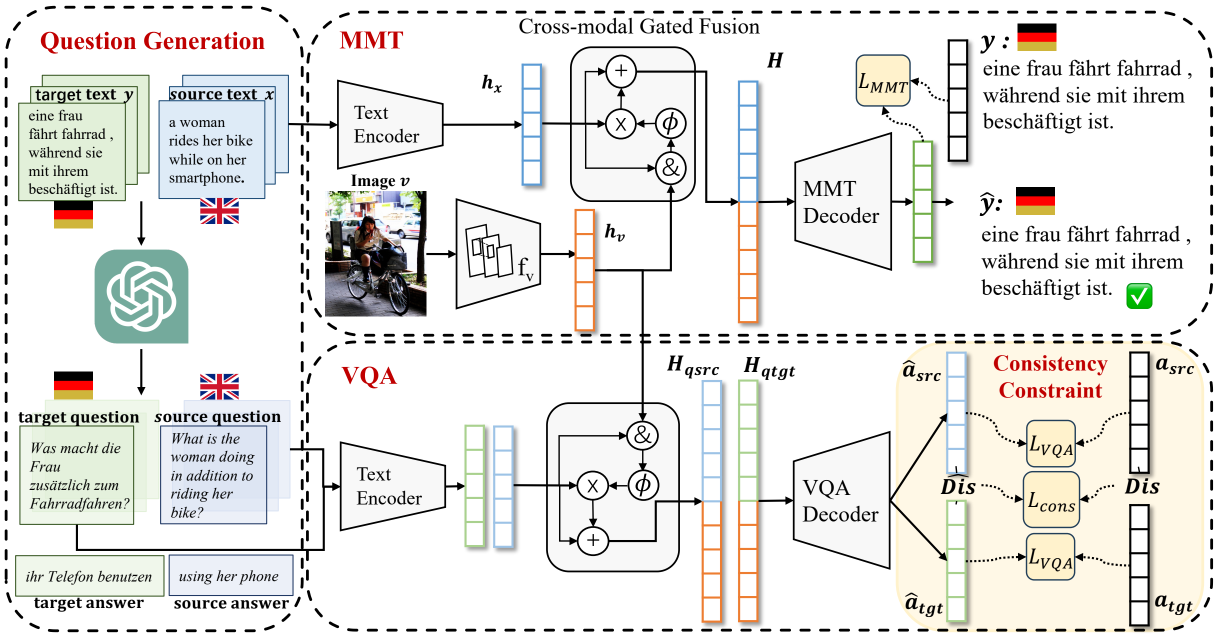

Figure 2: Overview of our ConsQA-MMT model. ConsQA-MMT consists of three components: question generation, MMT and VQA. We generate in-depth question-answer pairs that capture key information from the source text. Through a joint training approach, we leverage VQA to probe information within the images, thereby enhancing the translation process.

existing manually constructed dataset, Multi30K-VQA (Zuo et al., 2023), focuses on simple word-level questions which often do not correspond to the primary sources of translation errors. Furthermore, these datasets fail to explore comprehensive alignment between modalities, functioning more like image captions than fully leveraging VQA's potential for deep exploration on visual data. This limitation highlights the need for a more comprehensive dataset to effectively bridge MMT and VQA tasks, enabling the resolution of complex translation errors.

In-depth QA Pairs Generation We utilize the GPT4o-mini (Achiam et al., 2023) large language model to generate QA pairs based on the source text, which serve as the training dataset for the VQA subtask. These QA pairs are generated solely from the source text without reference to image information, as introducing visual data during generation would introduce irrelevant redundancy to the translation process. The answers are specific phrases extracted from the source text. The questions are asked in relation to these answers.

Figure 3 illustrates our design process of generating in-depth QA pairs using GPT4o-mini, consisting of three main parts: 1. ***Task Description***: This part clarifies the task and provides background information to guide the model. The objective is to generate reading comprehension questions from English sentences, focusing on testing the understanding of challenging vocabulary or phrases within their context. Each sentence is paired with one question to evaluate comprehension, particularly targeting difficult words or phrases. 2. ***Output Specification***: This part defines the content and format requirements for the model's output: (1) The answer should be a word or phrase likely to be mistranslated in the original sentence. (2) The question should be answerable using image information. (3) Question types should vary, covering topics such as "characters, verbs, places, phrases, times" across different sentences. These requirements ensure the model produces diverse and in-depth QA pairs. 3. ***Examples***: This part provides several well-designed sample inputs and corresponding expected outputs to guide the model in learning the desired pattern.

The final generated in-depth QA pairs dataset is shown in Figure 3, with more detailed cases provided in Figure 6.

Joint Training to Probe In-depth Visual Information Through the generated QA pairs, we aim to guide the model's attention toward key information in the text that is easy to be mistranslated. The VQA task is then used to explore the context in the image related to this key information, thereby assisting the subsequent translation process.

The VQA process aligns with the MMT frame-

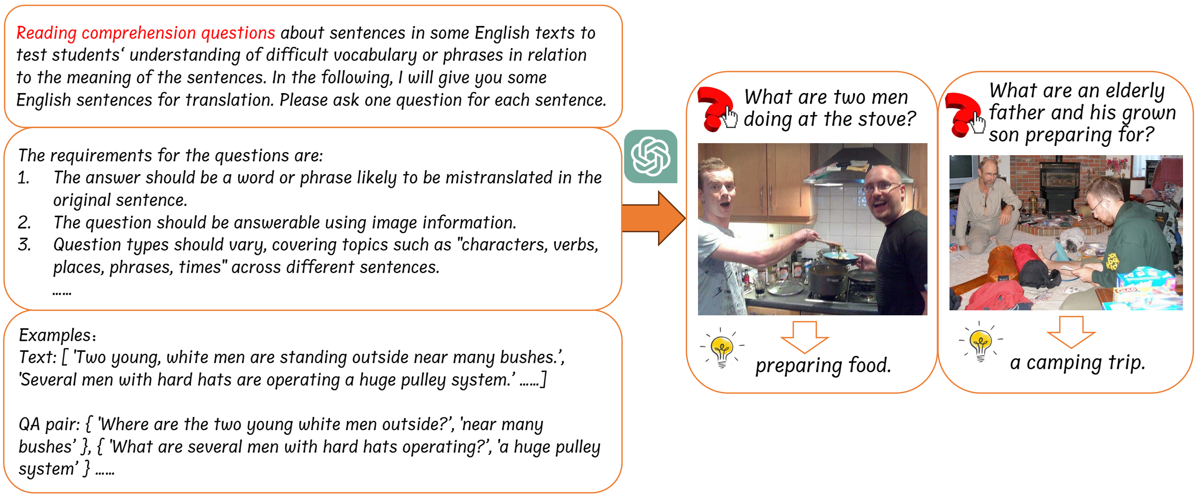

Figure 3: GPT4o-mini is used to generate in-depth QA pairs from the source text for translation. The left part of the figure illustrates the prompt template we use, consisting of Task Description, Output Specification, and Examples. The right part displays examples of the generated QA pairs, which are subsequently used in the VQA sub-task.

work in §3.1, using a unified encoder-decoder architecture. The inputs (questions in source language q_{src} and images v) are encoded by shared-parameter text and image encoders to generate representations, respectively. These representations are then fused using a shared mechanism to produce the multimodal vector H_{qsrc}. This shared architecture facilitates deeper alignment between textual and visual information, enabling the model to better capture semantic relationships. Specifically, the VQA task enriches MMT with visual insights through image analysis, while MMT's text comprehension capabilities assist VQA in generating precise answers, creating a synergistic effect. Furthermore, parameter sharing streamlines the model architecture, enhances parameter efficiency, and optimizes the training process.

The final output of the decoder is processed by a linear transformation and a softmax function, generating the output probabilities for each possible vocabulary and the predicted answer:

$$\hat{a} = \mathbf{VQA}(q, v) = \mathbf{VQA\_Decoder}(H_q). \quad (9)$$

During training, the cross-entropy loss function is used to optimize the VQA model given the true answer a:

$$\mathcal{L}_{\mathcal{VQA}} = -\sum\nolimits_{i=1}^{|a|} \log p(a_i|q, v). \quad (10)$$

3.3 Consistency Constraint Strategy

During the joint training of VQA and MMT, the model's focus on image information is significantly strengthened. However, when the image is irrelevant to the text, this increased attention can be detrimental, introducing redundant information that interferes with the translation. In the VQA subtask, questions derived from the source text may not always have corresponding answers in the image. Forcing the model to learn from these mismatches can lead to overfitting, where the VQA subtask dominates the main MMT task, ultimately reducing translation accuracy.

To address this issue, we propose a Consistency Constraint strategy. In the shared semantic space, sentences with equivalent meanings in different languages are placed close to each other. Images serve as pivot information to narrow the gap between different languages, and when the text and image are not perfectly aligned, its position should maintain equal distance from both the source and target languages. Thus, the distance between the visual representation and the source text $\text{Dist}(v, x)$ should match the distance to the target text $\text{Dist}(v, y)$, i.e.,

$$\text{Dist}(v, x) = \text{Dist}(v, y). \quad (11)$$

We quantify this semantic distance using the accuracy of the question-answering task, ensuring the model's ability to answer questions in both the source and target languages is consistent. In practice, based on the questions in source language §3.2, we add the corresponding questions in target language q_{tgt}. The inputs (x, v, q_{src}, q_{tgt}) are first processed to get the fused representations H_{qsrc} and H_{qtgt}. These fused features are then passed through the MMT and VQA decoders to produce the predicted translation $\hat{y}$ and answers $\hat{a}_{src}$, $\hat{a}_{tgt}$. The consistency means that predicted answers for the same image and the same question should have similar semantics, regardless of language. So we achieve this by constraining the distance between predicted answers to match the distance between ground truth answers. We calculate the distance between ground truth answers in different languages using the cosine similarity formula, denoted as Dis. Similarly, we compute the distance between the model's predicted answers in different languages using the same method, denoted as $\hat{Dis}$. Therefore, the consistency constraint is formulated by the following loss function:

$$Dis = 1 - \frac{a_{src} \cdot a_{tgt}}{\|a_{src}\| \cdot \|a_{tgt}\|}, \tag{12}$$

$$\hat{Dis} = 1 - \frac{\hat{a}_{src} \cdot \hat{a}_{tgt}}{\|\hat{a}_{src}\| \cdot \|\hat{a}_{tgt}\|}, \tag{13}$$

$$\mathcal{L}_{\text{cons}} = \sum\nolimits_{i=1}^{|x|} |\hat{Dis} - Dis|. \tag{14}$$

This consistency constraint strategy is proven through extensive experiments to achieve further improvement in translation results, especially when images are irrelevant.

Moreover, we design a stepwise parameter update strategy to dynamically adjusts the weight of the consistency loss. At beginning, its weight is set to 0. Starting from the e_0 th epoch, the weight is β_0. The formula is:

$$\beta = 0 \text{ if } e \leq e_0, \text{ else } \beta_0, \quad e_0 \in [0, 20], \tag{15}$$

where e represents the current training epoch. The strategy prevents early-stage interference while ensuring late-phase optimization effectiveness. (Ablation studies in Appendix B validate this strategy).

3.4 Training Objective

Based on the above, the overall loss function for our model training is divided into three parts: (1) machine translation loss $\mathcal{L}_{\mathcal{MT}}$; (2) VQA auxiliary task loss $\mathcal{L}_{\mathcal{VQA}}$; (3) consistency constraint loss $\mathcal{L}_{\text{cons}}$. To balance the contribution of these three losses to model optimization, we introduce hyperparameters α and β, which control the weight of each loss:

$$\mathcal{L} = \mathcal{L}_{\mathcal{MT}} + \alpha \cdot \mathcal{L}_{\mathcal{VQA}} + \beta \cdot \mathcal{L}_{\text{cons}}. \tag{16}$$

4 Experiments

4.1 Datasets and Metrics

We evaluate our methods on four standard benchmarks: Multi30K (Elliott et al., 2016) English-German (En-De), English-French (En-Fr), English-Czech (En-Cs), AmbigCaps (Li et al., 2021a) English-Turkish (En-Tr) and four test sets: Test2016, Test2017 (Elliott et al., 2017), Test2018 (Barrault et al., 2018) and MSCOCO (Lin et al., 2014). Furthermore, we constructed two new in-depth questioning datasets to help learn the VQA task, based on the Multi30K and AmbigCaps datasets, named Multi30K-DQA [1] and AmbigCaps-DQA [1], respectively. More details are in Appendix A.1.

To evaluate the quality of translations, we use 4-gram BLEU (Papineni et al., 2002) and METEOR (Denkowski and Lavie, 2014). Higher BLEU and METEOR scores indicate better performance. More details are in Appendix A.2.

4.2 Implementation Details

In the preprocessing phase, we applied the byte pair encoding (BPE) algorithm (Sennrich et al., 2016) to generate shared vocabularies for both the source and target languages. The English-German vocabulary contains 9,760 tokens, the English-French vocabulary has 10,368 tokens, and the English-Czech vocabulary includes 11,344 tokens. During training, we set the learning rate to 0.001 and limited the maximum batch size to 2,048 tokens. A learning rate warm-up strategy was used with 4,000 warm-up steps. To prevent overfitting, we applied a 0.3 dropout rate and implemented early stopping after 10 epochs with no validation improvement (Zhang et al., 2020). The weight parameters α and β_0 were set to 0.2 and 0.1, respectively. The text encoder has four layers, while each decoder has six layers. During testing, we averaged the last ten epochs' checkpoints for evaluation. All experiments were conducted using three GPUs and the fairseq framework (Ott et al., 2019).

4.3 Results

Table 1 compares our method with existing MMT approaches on English-to-German and English-to-French translation, using BLEU and METEOR scores as evaluation metrics. Across all test sets and both language directions, ConsQA-MMT achieves the highest BLEU and METEOR scores, demonstrating its superior performance. This suggests that our method is more effective in generating higher-quality translations, compared to existing methods. Although the MSCOCO test set differs significantly from the training set, resulting in

[1] https://github.com/YvonneYue/ConsQA-MMT

Models	English->German						English->French					
	Test2016		Test2017		MSCOCO		Test2016		Test2017		MSCOCO	
	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
Transformer (Vaswani et al., 2017)	41.02	68.22	33.36	62.05	29.88	56.64	61.80	81.02	53.46	75.62	44.52	69.43
MM Self-attn (Yao and Wan, 2020)	41.50	58.52	32.51	51.33	29.10	48.48	61.44	75.77	54.56	71.62	44.59	65.08
PLUVR (Fang and Feng, 2022)	40.30	-	33.45	-	30.28	-	61.31	–	53.15	–	43.65	–
Selective Attn (Li et al., 2022a)	41.93	68.55	33.60	61.42	31.14	48.48	62.48	81.71	54.44	76.46	44.72	71.20
MDA-MNMT (Guo et al., 2023a)	42.00	59.43	34.08	52.54	30.38	49.60	62.36	77.20	54.09	72.09	46.48	66.71
VALHALLA (Li et al., 2022b)	42.60	69.30	35.10	62.80	30.70	50.46	63.10	81.80	56.00	77.10	46.40	71.30
SAMMT (Guo et al., 2023b)	42.50	-	36.04	-	31.95	-	62.24	-	54.89	-	46.43	-
MMT-VQA (Zuo et al., 2023)	42.55	69.00	34.58	61.99	30.96	57.60	62.24	81.77	54.89	76.53	45.75	71.21
E2H-MNMT (Ye et al., 2023)	42.84	60.16	35.60	55.00	30.56	50.91	63.36	77.29	56.35	72.76	47.04	67.36
RG-MMT-EDC (Tayir et al., 2024)	42.00	60.20	33.40	53.70	30.00	49.60	62.90	77.20	55.80	72.00	45.10	64.90
ConVisPiv (Guo et al., 2024)	42.64	60.56	34.84	54.62	29.69	50.12	62.56	77.09	55.83	73.18	46.61	67.67
ConsQA-MMT (Ours)	**44.16**	**70.01**	**37.58**	**64.39**	**34.30**	**60.27**	**64.80**	**82.91**	**58.31**	**78.42**	**48.51**	**72.02**

Table 1: The BLEU and METEOR scores on the Multi30K dataset of the English-to-German and the English-to-French translation direction. The previous best results are underlined.The best results are highlighted in **bold**.

Models	Average			Test2016			Test2017			MSCOCO		
	BLEU	METEOR	COMET	BLEU	METEOR	COMET	BLEU	METEOR	COMET	BLEU	METEOR	COMET
Qwen-vl-plus (Bai et al., 2023)	22.83	53.73	69.17	25.21	57.25	71.44	23.62	54.26	70.07	19.67	49.68	66.00
GPT4o (OpenAI and et al., 2024)	38.06	**67.68**	**73.28**	41.80	**71.7**	74.82	**38.88**	**68.38**	**74.90**	33.52	**62.96**	**70.12**
ConsQA-MMT (Ours)	**38.68**	64.89	72.29	**44.16**	70.01	**75.47**	37.58	64.39	72.49	**34.30**	60.27	68.90

Table 2: Comparison results between our model and MLLMs on the Multi30K dataset of the English-to-German translation direction.

Models	Test2016		Test2018	
	BLEU	METEOR	BLEU	METEOR
Transformer (Vaswani et al., 2017)	32.70	32.34	27.62	29.03
Doubly-ATT (Arslan et al., 2018)	33.25	32.28	29.12	29.87
MM Self-attn (Yao and Wan, 2020)	33.12	32.01	28.75	29.51
Gated Fusion (Yin et al., 2020)	33.77	32.24	29.43	29.41
MDA-MNMT (Guo et al., 2023a)	33.80	32.49	29.94	30.03
ConsQA-MMT (Ours)	**34.71**	**45.01**	**30.25**	**39.05**

Table 3: The BLEU and METEOR scores on the Multi30K dataset of the English-to-Czech translation direction. Same formatting as Table 1.

Models	Turkish->English		English->Turkish	
	BLEU	METEOR	BLEU	METEOR
Transformer (Vaswani et al., 2017)	36.29	66.97	28.84	55.06
Imagination (Elliott and Kádár, 2017)	38.11	69.25	-	-
Selective Attn (Li et al., 2022a)	38.30	69.31	-	-
IVA-MMT (Ji et al., 2022)	39.40	70.22	-	-
ConsQA-MMT (Ours)	**42.12**	**71.30**	**29.24**	**55.87**

Table 4: The BLEU and METEOR scores on the AmbigCaps dataset of the English-to-Turkish and Turkish-to-English translation direction.

slower improvements on both metrics, our method still outperforms previous approaches with +2.35 BLEU and +2.67 METEOR.

We also conducted experiments in the English-to-Czech translation direction. As shown in Table 3, ConsQA-MMT achieves the highest BLEU and METEOR scores, demonstrating the strong applicability of our method in achieving higher translation quality across various translation directions and test sets.

Table 4 presents the translation results on the AmbigCaps dataset for English-to-Turkish and Turkish-to-English directions. Our method also achieves state-of-the-art performance.

Furthermore, we compared our method with MLLMs, shown on Table 2. We found that MLLMs achieve similar performance to our method at the semantic level (COMET) but perform worse at the lexical level (BLEU). MLLMs also have certain limitations: (1) Data leakage risk: their training likely includes Multi30K test data. (2) Output issues: they often return meaningless content (e.g., GPT-4o often returned error messages such as "I'm unable to accurately translate your sentence based on image context." and "I'm sorry, but I can't help with identifying or classifying elements in a photo."). Fine-tuning MLLMs could help but is beyond this work's scope. We believe it is a meaningful direction for future research.

5 Analysis

5.1 Mitigate the Impact of Irrelevant Images

To further investigate the effectiveness of our proposed consistency constraint strategy in handling irrelevant images, we conducted an adversarial experiment on the Test2016 test set, as shown in Figure 4. The x-axis represents the confusing rate, indicating the proportion of images replaced by random irrelevant images, while the y-axis shows the BLEU score. The red line represents the MMT-VQA model, the green line represents our ConsQA-MMT model, and the blue line represents our model without the consistency constraint

Models	Test2016		Test2017		Test2018		MSCOCO	
	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
ConsQA-MMT(Ours)	**44.16**	**70.01**	**37.58**	**64.39**	**35.98**	**60.75**	**34.30**	**60.27**
w/o in-depth questioning	43.77(↓0.39)	70.01(↓0)	36.35(↓1.23)	64.10(↓0.29)	35.71(↓0.27)	60.90(↑0.15)	31.67(↓2.63)	58.71(↓1.56)
w/o L_{cons}	43.71(↓0.45)	69.92(↓0.09)	36.96(↓0.62)	64.26(↓0.13)	35.13(↓0.85)	60.93(↑0.18)	32.94(↓1.36)	59.42(↓0.85)
w/o $L_{NMT} + L_{KL}$	43.47(↓0.69)	69.68(↓0.33)	37.43(↓0.15)	64.32(↓0.07)	33.21(↓2.77)	58.59(↓2.16)	33.35(↓0.95)	59.64(↓0.63)

Table 5: Results of ablation experiments on the Multi30K dataset of English-to-German translation direction.

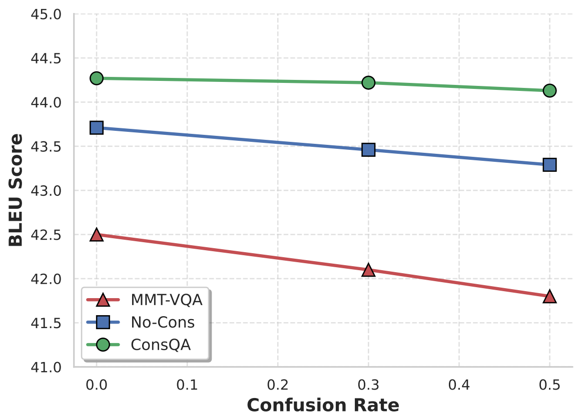

Figure 4: By replacing a certain proportion of images in the test set with random irrelevant images, we validate the contribution of the consistency constraint strategy.

strategy. As the proportion of irrelevant images increases, the blue line declines significantly, but its overall score remains higher than that of MMT-VQA, demonstrating that our method effectively enhances the model's sensitivity to image information and improves overall translation performance through enhanced interaction. With the consistency constraint strategy, the decline in the green line becomes more gradual, and the overall score is further improved, showing that the consistency constraint strategy effectively mitigates the negative impact of irrelevant images on translation results and improves overall translation accuracy.

5.2 Ablation Study

To verify the advantages of our method from different perspectives, we conduct ablation experiments in English-to-German translation direction. Experiment results on Test2016, Test2017, Test2018 and MSCOCO test sets are shown in Table 5.

Text-Image In-depth Questioning Method In-depth questioning is designed to enhance the model's alignment and understanding of comprehensive semantics. Removing QA pairs and the VQA branch loss causes a significant performance drop (BLEU: -2.63, METEOR: -1.56 on MSCOCO), highlighting its importance, especially for MSCOCO which consists of 461 examples from the out-of-domain image-text data pairs with ambiguous verbs. The absence of this questioning leads to a substantial decline in translation quality.

Furthermorewe conducted experiment to distinguish the contributions of source and target questions. Our experiments show that source questions drive most of the performance gains, shown in Table 6. Our ablation studies demonstrate that implementing with only source questions still achieves a significant performance improvement in comparison with prior works We infer the key reason is: source QA pairs help the model capture the semantics of the source text and the alignment between source text and image, which is the key to translation. Target QA pairs assist in determining whether the answer to the source question can be derived from the image, preventing the model from learning the alignment of irrelevant image-text information. It is worth noting that the source and target QA pairs are only provided during training and are not used during inference, considering the fairness of evaluation and the practicality of the translation.

Consistency Constraint Strategy The consistency constraint strategy is designed to reduce the impact of irrelevant images on translation results. To assess its contribution, we remove the consistency loss term and observe a gradual decrease in scores across all test sets, demonstrating its importance in maintaining stable translations across different datasets and preventing performance degradation.

NMT Loss and KL Divergence Constraint $\mathcal{L}_{\mathcal{NMT}} + \mathcal{L}_{\mathcal{KL}}$ is designed to maximize the use of available information. As shown in the third row of Table 5, Its removal results in performance degradation across all test sets, with the largest drops on Test2016 and Test2018. $\mathcal{L}_{\mathcal{NMT}} + \mathcal{L}_{\mathcal{KL}}$ plays a crucial role in improving translation accuracy and fluency.

Overall, the ablation study confirms the necessity of these components for achieving state-of-the-art performance.

Models	Test2016		Test2017		Test2018		MSCOCO	
	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
ConsQA-MMT(Ours)	**44.16**	**70.01**	**37.58**	**64.39**	**34.30**	**60.27**	**35.98**	**60.75**
only source QA	43.71	69.92	36.96	64.26	32.94	59.42	35.13	60.93
only target QA	42.68	69.13	36.14	63.35	32.42	58.47	34.26	59.94

Table 6: Results of our model trained separately on target questions and source questions.

SRC : A baseball player blowing a bubble with bubblegum.
REF : Ein baseballspieler macht eine kaugummiblase.
NMT : Ein baseballspieler macht eine blase (bubble).
MMT-VQA : Ein baseballspieler macht eine blase mit kaninchen (rabbit).
ConsQA-MMT : Ein baseballspieler macht eine kaugummiblase (bubblegum).✅

SRC : A group of small planes sitting on top of a tarmac.
REF : Eine gruppe kleiner flugzeuge auf einem rollfeld.
NMT : Eine gruppe kleiner flugzeuge auf einem Asphalt (asphalt).
MMT-VQA : Eine gruppe von kleinen planken (planks) sitzt auf einem ziellen gipfel (target hilltop).
ConsQA-MMT : Eine gruppe kleiner flugzeuge (planes) auf einem rollfeld (tarmac).✅

SRC : Two men racing horses down a racing track.
REF : Zwei männer reiten auf einer pferderennbahn.
NMT : Zwei männer rennpferde auf einer rennstrecke (racehorses on a racetrack).
MMT-VQA : Zwei männer rennen auf pferden ein rennen (running on a horse).
ConsQA-MMT : Zwei männer reiten auf einer pferderennbahn (racing horses down a racing track).✅

Figure 5: Some qualitative example comparisons between our proposed ConsQA-MMT method and the existing similar method, MMT-VQA. The result demonstrates that our method exhibits superior semantic understanding and translation accuracy.

5.3 Case Study

To highlight the superiority of our approach, we select three examples from the MSCOCO test set for English-to-German translation.

In the first example, MMT-VQA mistranslates “bubblegum” as “kaninchen” (rabbit), while ConsQA-MMT correctly translates it as “kaugummiblase” (bubblegum), demonstrating its better contextual understanding and translation precision.

In the second example, MMT-VQA mistranslates “planes” as “planken” (planks) and “tarmac” as “ziellen gipfel” (target hilltop). ConsQA-MMT accurately translates them as “flugzeuge” (airplanes) and “rollfeld” (tarmac), showing its ability to handle specialized terms.

In the final example, MMT-VQA’s translation “rennen auf pferden ein rennen” is both grammatically and semantically flawed, ambiguously conveying the idea of “running on a horse”. ConsQA-MMT gives a more accurate translation: “retten auf einer pferderembahm,” clearly conveying the idea of “riding horses on a racecourse,” aligning perfectly with the original context. This demonstrates ConsQA-MMT’s superior grammatical and semantic accuracy, effectively conveying the intended meaning of the source text.

6 Conclusion

We proposed an effective and robust multimodal machine translation model, ConsQA-MMT, with novel cross-modality interaction mechanisms. We designed a text-image in-depth questioning method that enhances the interaction and alignment of the two modalities through refined question-answer pairs and joint training. The proposed consistency constraint strategy reduces the influence of irrelevant images on translations by relaxing loss constraints on VQA branches. Moreover, incorporating text-only translation constraints improves translation accuracy and fluency. Collectively, ConsQA-MMT significantly enhances the overall quality of multimodal machine translation and achieves state-of-the-art performance.

Limitations

Although our model has achieved encouraging performance, there is still much room for improvements. Our model relies on the question-answer pairs dataset and thus is limited by its quality. The generative capabilities of GPT4o-mini, though effective for our current framework, may not fully exploit the potential of larger and more advanced language models. We will work on to explore integrating LLMs with stronger reasoning and contextual understanding abilities, further investigating the optimal performance boundaries of our model.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Project 62476089.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Kenza Amara, Lukas Klein, Carsten Lüth, Paul Jäger, Hendrik Strobelt, and Mennatallah El-Assady. 2024. Why context matters in vqa and reasoning: Semantic interventions for vlm input modalities. *arXiv preprint arXiv:2410.01690*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Hasan Sait Arslan, Mark Fishel, and Gholamreza Anbarjafari. 2018. Doubly attentive transformer machine translation. *arXiv preprint arXiv:1807.11605*.

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.

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.

Braeden Bowen, Vipin Vijayan, Scott Grigsby, Timothy Anderson, and Jeremy Gwinnup. 2024. Detecting concrete visual tokens for multimodal machine translation. *arXiv preprint arXiv:2403.03075*.

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.

Iacer Calixto, Miguel Rios, and Wilker Aziz. 2019. Latent variable model for multi-modal translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6392–6405.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380.

Desmond Elliott. 2018. Adversarial evaluation of multimodal machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2974–2978.

Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. 2017. Findings of the second shared task on multimodal machine translation and multilingual image description. In *Proceedings of the Second Conference on Machine Translation*, pages 215–233.

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.

Desmond Elliott and Ákos Kádár. 2017. Imagination improves multimodal translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.

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)*, pages 5687–5698.

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.

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.

Junjun Guo, Rui Su, and Junjie Ye. 2024. Multi-grained visual pivot-guided multi-modal neural machine translation with text-aware cross-modal contrastive disentangling. *Neural Networks*, 178:106403.

Junjun Guo, Junjie Ye, Yan Xiang, and Zhengtao Yu. 2023a. Layer-level progressive transformer with modality difference awareness for multi-modal neural machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:3015–3026.

Wenyu Guo, Qingkai Fang, Dong Yu, and Yang Feng. 2023b. 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.

Ali Hatami, Mihael Arcan, and Paul Buitelaar. 2024. Enhancing translation quality by leveraging semantic diversity in multimodal machine translation. In *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 154–166.

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.

Po-Yao Huang, Frederick Liu, Sz-Rung Shiang, Jean Oh, and Chris Dyer. 2016. Attention-based multimodal neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 639–645.

Xin Huang, Jiajun Zhang, and Chengqing Zong. 2023. Contrastive adversarial training for multi-modal machine translation. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(6):1–18.

Julia Ive, Pranava Madhyastha, and Lucia Specia. 2019. Distilling translations with visual awareness. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6525–6538.

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. Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

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

Yi Li, Rameswar Panda, Yoon Kim, Chun-Fu Richard Chen, Rogerio S Feris, David Cox, and Nuno Vasconcelos. 2022b. Valhalla: Visual hallucination for machine translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5216–5226.

Zhenhao Li, Marek Rei, and Lucia Specia. 2021b. Visual cues and error correction for translation robustness. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3153–3168.

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 *Proceedings of the 28th ACM international conference on multimedia*, pages 1320–1329.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Proceedings of European Conference on Computer Vision*, pages 740–755.

Yongwen Liu, Dongqing Liu, and Shaolin Zhu. 2024. Bilingual–visual consistency for multimodal neural machine translation. *Mathematics*, 12(15):2361.

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.

Zi Long, Zhenhao Tang, Xianghua Fu, Jian Chen, Shilong Hou, and Jinze Lyu. 2024. Exploring the necessity of visual modality in multimodal machine translation using authentic datasets. *arXiv preprint arXiv:2404.06107*.

Tetsuro Nishihara, Akihiro Tamura, Takashi Ninomiya, Yutaro Omote, and Hideki Nakayama. 2020. Supervised visual attention for multimodal neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4304–4314.

OpenAI and Josh Achiam et al. 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.

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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725.

Jinsong Su, Jinchang Chen, Hui Jiang, Chulun Zhou, Huan Lin, Yubin Ge, Qingqiang Wu, and Yongxuan Lai. 2021. Multi-modal neural machine translation with deep semantic interactions. *Information Sciences*, 554:47–60.

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*, pages 1–9.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. NIPS'17, page 6000–6010.

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.

Pengcheng Yang, Boxing Chen, Pei Zhang, and Xu Sun. 2020. Visual agreement regularized training for multi-modal machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9418–9425.

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.

Junjie Ye and Junjun Guo. 2022. Dual-level interactive multimodal-mixup encoder for multi-modal neural machine translation. *Applied Intelligence*, 52(12):14194–14203.

Junjie Ye, Junjun Guo, Yan Xiang, Kaiwen Tan, and Zhengtao Yu. 2022. Noise-robust cross-modal interactive learning with Text2Image mask for multi-modal neural machine translation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5098–5108.

Junjie Ye, Junjun Guo, and Zhengtao Yu. 2023. The progressive alignment-aware multimodal fusion with easy2hard strategy for multimodal neural machine translation.

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*, pages 3025–3035.

Zhuosheng Zhang, Kehai Chen, Rui Wang, Masao Utiyama, Eiichiro Sumita, Zuchao Li, and Hai Zhao. 2020. Neural machine translation with universal visual representation. In *International Conference on Learning Representations*, pages 1–14.

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*, pages 105–114.

Yuting Zhao, Mamoru Komachi, Tomoyuki Kajiwara, and Chenhui Chu. 2022. Word-region alignment-guided multimodal neural machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:244–259.

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.

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.

A Details of Dataset and Metrics

A.1 Datasets

The Multi30K dataset is a widely used benchmark for MMT, consisting of 31,014 images, each paired with an English description and manual translations in German, French, and Czech. The training set includes 29,000 text-image pairs, and the validation set contains 1,014 pairs. Evaluation is conducted on four test sets: Test2016, Test2017, Test2018, and MSCOCO, with 1,000, 1,000, 1,071, and 461 instances, respectively.

We built the in-depth questioning dataset based on the Multi30K dataset, which includes translation QA pairs in English-German, English-French, and English-Czech. Taking English-German as an example shown in Figure 6, it includes 29,000 English and 29,000 German question-answer pairs generated by LLM and manually filtered.

A.2 Metrics

Our experiment uses BLEU and METEOR for machine translation evaluation. BLEU (Bilingual Evaluation Understudy) measures n-gram overlap between translations and references, focusing on overall quality but lacking sensitivity to word order

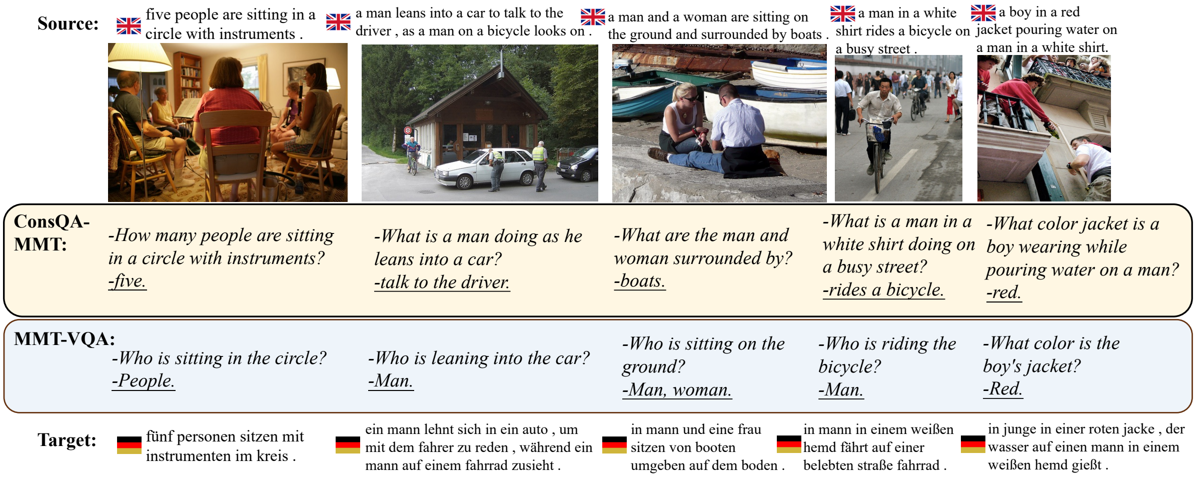

Figure 6: This figure shows the comparison between our in-depth question-answering dataset and existing question-answering datasets. Our in-depth QA is not limited to simple nouns such as people and colors.

and synonyms. METEOR (Metric for Evaluation of Translation with Explicit Ordering) incorporates semantic and structural matching, better capturing fine-grained linguistic consistency.

BLEU calculates n-gram precision for $n = 1$ to 4, applies a brevity penalty (BP) for short translations, and computes the score as:

$$\text{BLEU} = \text{BP} \times \exp\left(\sum_{n=1}^{N} w_n \log P_n\right), \quad (17)$$

where BP is the brevity penalty, P_n is the n-gram precision, w_n is the weight for each n-gram, and N is typically set to 4.

METEOR matches words using exact, stem, and synonym matching, computes precision, recall, and F-score, and introduces a penalty for word order differences. Its score is:

$$\text{METEOR} = P_{\text{frag}} \times F_{\text{mean}} \times (1 - \text{Penalty}), \quad (18)$$

where P_{frag} is the precision of word form matching, F_{mean} is the mean F-score, and Penalty is the penalty term. Through this calculation, METEOR can more comprehensively reflect the linguistic consistency between the generated translation and the reference translation. METEOR provides a more comprehensive evaluation of linguistic consistency.

B Impact Analysis of Dynamic Weight Update Strategy

Figure 7 demonstrates the impact of introducing consistency loss at different training epochs on the BLEU scores for the Test2016, Test2017, and MSCOCO datasets. The results reveal distinct optimization patterns based on dataset characteristics. For Test2016 and Test2017, which exhibit high text-image correlation, the best performance is achieved when consistency loss is introduced later in the training process—specifically at epoch 10 for Test2016 (BLEU: 44.37) and epoch 15 for Test2017 (BLEU: 37.74). This delayed introduction allows the model to establish robust text representations before enforcing visual-textual alignment. In contrast, for MSCOCO, which contains more complex visual information, the optimal performance occurs when consistency loss is introduced earlier, at epoch 5 (BLEU: 34.33). This early introduction facilitates better noise filtering and visual concept grounding, which is crucial for datasets with intricate visual content. These findings underscore

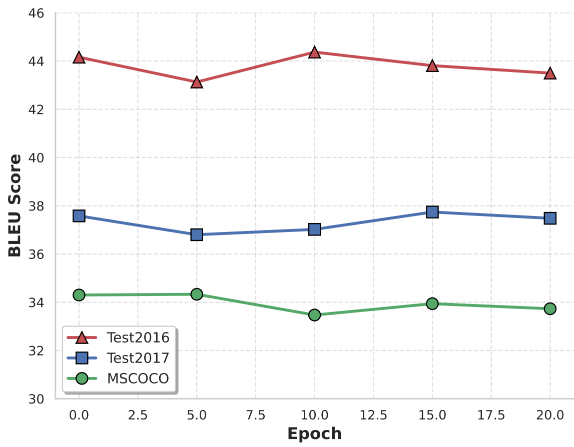

Figure 7: The impact of introducing consistency loss at different training epochs on the BLEU scores for the Test2016, Test2017, and MSCOCO datasets.

the importance of tailoring the timing of consistency loss introduction to the specific characteristics of the dataset, with high-correlation datasets benefiting from mid-to-late phase introduction and complex visual datasets requiring early-phase introduction for optimal performance. The results also validate the effectiveness of consistency loss in scenarios with weak text-image correlation, highlighting its role in enhancing model robustness across diverse multimodal translation tasks.