VQA-Augmented Machine Translation with Cross-Modal Contrastive Learning

Zhihui Zhang[1], Shiliang Sun[1,2*], Jing Zhao[1*], Tengfei Song[3], Hao Yang[3]
[1]School of Computer Science and Technology,
East China Normal University, Shanghai 200062, China
[2]State Key Laboratory of Submarine Geoscience, School of Automation and Intelligent Sensing,
Shanghai Jiao Tong University, Shanghai 200240, China
[3]2012 Labs, Huawei Technologies CO., LTD, China
51265901049@stu.ecnu.edu.cn, shiliangsun@gmail.com, jzhao@cs.ecnu.edu.cn,
{songtengfei2, yanghao30}@huawei.com

Abstract

Multimodal machine translation (MMT) aims to enhance translation quality by integrating visual information. However, existing methods often extract visual features using pre-trained models while learning text features from scratch, leading to representation imbalance. These methods are also prone to being misled by redundant visual information, which results in suboptimal performance. To address these challenges, we propose CAMT, a novel cross-modal VQA-augmented MMT method. CAMT aligns image-source text pairs and image-question text pairs through dual-text contrastive learning, thereby improving semantic consistency across modalities. Additionally, we design an effective strategy for generating question–answer pairs to enhance fine-grained alignment and filter out irrelevant visual noise, while also addressing the scarcity of VQA annotations. Extensive experiments on multiple benchmark datasets demonstrate the effectiveness of the proposed CAMT framework, which consistently outperforms state-of-the-art MMT methods across multiple evaluation metrics.

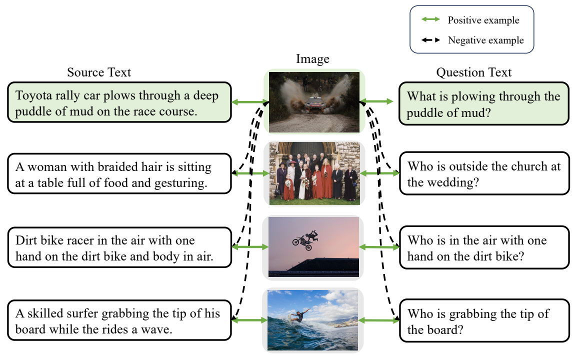

Figure 1: Illustration of positive and negative sample pairs in dual-text contrastive learning: image-source text and image-question text.

1 Introduction

Multimodal Machine Translation (MMT) enhances translation quality by incorporating visual information to address the ambiguity in traditional Neural Machine Translation (NMT). Recent studies have explored strategies to assess and improve the role of visual inputs in MMT. Although early research (Lala et al., 2018) found limited benefits from image context, later work (Li et al., 2022a) showed significant performance gains with more advanced visual encoders. These findings highlight the importance of stronger interaction between visual and textual modalities.

To mitigate these challenges, researchers have explored auxiliary tasks such as image captioning (Cheng et al., 2023), object detection, and visual question answering (VQA) (Zuo et al., 2023). In many existing MMT datasets, the limited relevance between images and their paired texts introduces substantial visual noise, complicating effective multimodal integration. Methods such as synthetic image generation (Li et al., 2022b; Yuasa et al., 2023) and dynamic image feature filtering (Ye et al., 2022a; Fang and Feng, 2022; Lu et al., 2021) have been proposed to mitigate this challenge to some extent. While previous work has constructed QA pairs for the Multi30K dataset, many MMT datasets, such as 3AM, lack such pairs, limiting the broader application of VQA. To address this, we propose a novel method to automatically generate semantically richer QA pairs for the 3AM dataset using a large language model (LLM). Unlike existing VQA methods that focus primarily on limited aspects like nouns, numbers, or colors, our approach targets deeper semantic elements essential for translation, such as prepositional phrases, action descriptions, and spatial relations. These elements are critical for resolving translation ambiguity and capturing fine-grained image-text relationships.

We further observe that most MMT systems suffer from a representation imbalance: image

[*]Corresponding authors.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 10113–10124
November 4-9, 2025

features are extracted from frozen pre-trained encoders, while text features are dynamically learned with Transformers. This imbalance introduces two key limitations. First, static visual features cannot adapt to the evolving semantics of the source text, resulting in rigid cross-modal interactions. Second, this asymmetry prevents attention mechanisms from fully capturing dynamic visual–textual relationships, thereby hindering fine-grained alignment. However, existing MMT datasets (e.g., Multi30K) contain only about 29K image–text pairs, which is orders of magnitude smaller than the datasets used to pre-train vision models. In such a data-scarce setting, retraining or fine-tuning the visual encoder is computationally prohibitive and highly prone to overfitting, making it an impractical solution.

To address these limitations, we propose a VQA-augmented contrastive learning framework that dynamically aligns vision and text. By leveraging semantically enriched question-answer pairs, our method projects visual features into a trainable latent space, where contrastive learning promotes fine-grained alignment with textual semantics. Positive pairs are constructed between images and their corresponding source and question texts, while unrelated combinations serve as negatives (Figure 1). This dual-text supervision enhances the adaptability of visual representations and bridges the semantic gap across modalities, improving translation quality.

Our main contributions are summarized as follows:

- We propose a cross-modal VQA-augmented machine translation method to address the misalignment of image and text features in current MMT systems.
- We propose a simple yet effective LLM-based QA generation strategy that mitigates the scarcity of VQA annotations and produces diverse pairs across six categories.
- Extensive experiments on benchmark datasets demonstrate that our CAMT framework significantly outperforms state-of-the-art MMT methods across multiple metrics.

2 Related Work

2.1 Multimodal Contrastive Learning

Contrastive learning (He et al., 2020) optimizes feature representations by clustering semantically similar pairs while repelling dissimilar ones. It has been widely used in scenarios such as sentence embedding learning (Yan et al., 2021), machine translation (Pan et al., 2021; Ye et al., 2022b), and text summarization (Cao and Wang, 2021). It has also been gradually extended to multimodal scenarios, including aligning image-text representations (Zhou et al., 2020), and the potential to enhance semantic robustness in adversarial training (Huang et al., 2023). Our approach focuses on semantically relevant image regions and leverages dual-text supervision to achieve fine-grained alignment between visual and textual representations, particularly for complex multimodal translation tasks.

2.2 Multimodal Machine Translation

Multimodal machine translation (MMT) aims to enhance text translation systems by integrating additional modalities, such as images and videos. Early efforts in this domain include dual attention decoders (Calixto et al., 2017), latent variable models (Calixto et al., 2019), and methods that employ visual information only in a refinement stage to address translation needs (Ive et al., 2019). Additionally, cross-lingual visual pre-training has been explored to leverage visual features across languages (Caglayan et al., 2021). However, adversarial evaluation has exposed limitations in the effective utilization of visual modalities (Elliott, 2018). Recent advances focus on improving vision-text fusion through multimodal transformers (Yao and Wan, 2020) and dynamic context-guided networks (Lin et al., 2020), optimizing visual processing via encoder-decoder calibration (Lu et al., 2021) and multi-granularity guidance (Guo et al., 2024), and leveraging pre-trained models for knowledge transfer (Gupta et al., 2023) and adversarial image generation (Guo et al., 2023a). Other notable contributions include the Soul-Mix method, which enhances multimodal translation through manifold mixing (Cheng et al., 2024), image-assisted methods that tackle ambiguity (Futeral et al., 2023), and approaches that leverage data beyond triplets to enrich multimodal machine translation (Zhu et al., 2023). More recently, ConsQA-MMT (Gao et al., 2025) enhances robustness by questioning both source and target texts with consistency constraints. In contrast, we jointly optimize VQA and contrastive learning to achieve finer semantic alignment and reduce the modality gap.

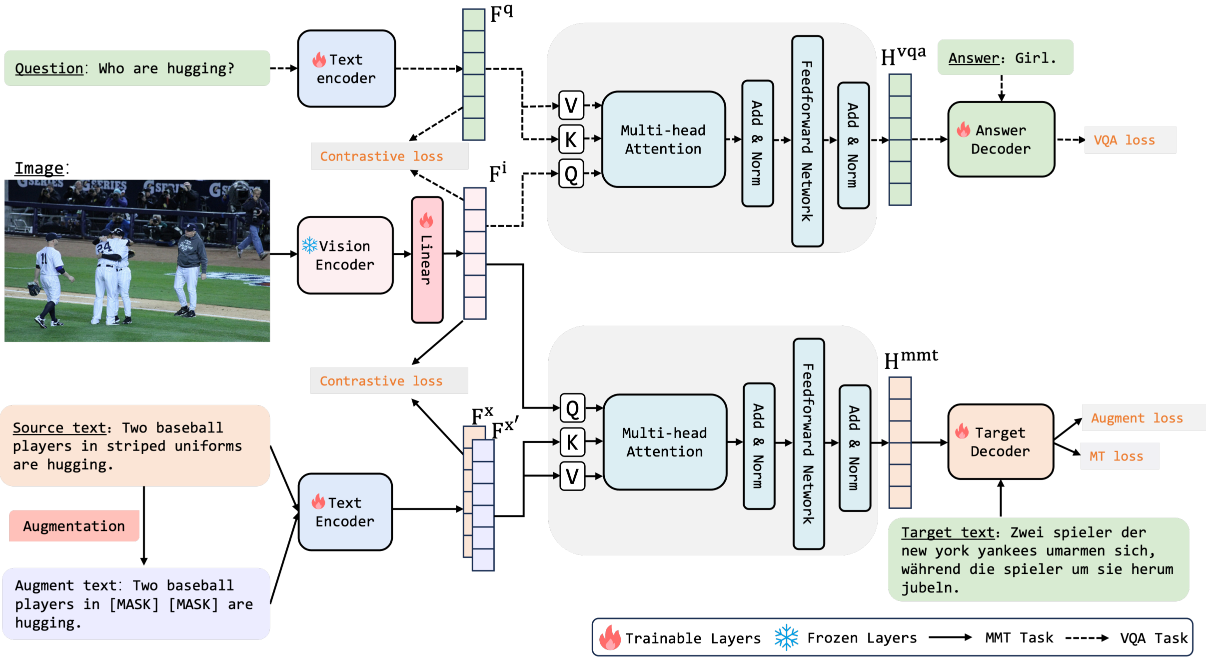

Figure 2: The overall framework of our proposed CAMT model, which includes the dual-text contrastive learning stage, VQA-augmented interaction stage, data augmentation, unified multimodal translation strategies, and the overall training objective.

3 Method

This section provides a comprehensive overview of our proposed framework, detailing its key components: the dual-text contrastive learning stage, the VQA-augmented interaction stage, data augmentation, the unified multimodal translation strategies, and the overall training objective. The overall architecture of our **C**ross-modal VQA-**A**ugmented **M**ultimodal machine **T**ranslation model (CAMT) is illustrated in Figure 2.

3.1 Dual-Text Contrastive Learning

Image-Source Text Contrastive Learning

Given a source text X and the corresponding image I, a contrastive loss function is utilized to maximize the similarity scores for correctly matched image-text pairs while minimizing the scores for mismatched pairs. The contrastive loss function is formally defined as follows:

$$\text{sim}_{i,j}(I, X) = \exp(\text{sim}(I_i, X_j)/\tau), \tag{1}$$

$$\mathcal{L}_{\text{i-s}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\text{sim}_{i,i}(I, X)}{\sum_{j=1}^{N}\text{sim}_{i,j}(I, X)}, \tag{2}$$

where $\text{sim}(\cdot,\cdot)$ denotes the similarity function (e.g., cosine similarity) between images and texts, τ is a temperature parameter, and N is the batch size.

Image-Question Text Contrastive Learning

To further enhance the image's understanding of textual semantics, we introduce contrastive learning between the image I and a question Q generated from the source text X. The corresponding contrastive loss is defined as:

$$\mathcal{L}_{\text{i-q}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\text{sim}_{i,i}(I, Q)}{\sum_{j=1}^{N}\text{sim}_{i,j}(I, Q)}. \tag{3}$$

By jointly optimizing these two contrastive learning objectives, the image is guided by both source text X and question Q to capture richer and more precise semantic representations. The overall loss function is formulated as:

$$\mathcal{L}_{\text{CTR}} = \mathcal{L}_{\text{i-s}} + \mathcal{L}_{\text{i-q}}. \tag{4}$$

Through simultaneous contrastive alignment with both the source text and the question text, the image representation is progressively refined, leading to a deeper understanding of textual semantics.

3.2 VQA-Augmented Interaction

In this phase, the VQA task is executed within a multimodal encoder-decoder architecture. It demands tight integration of visual and textual elements, necessitating that the model identify pertinent visual attributes from the image and correlate them with the question's textual features.

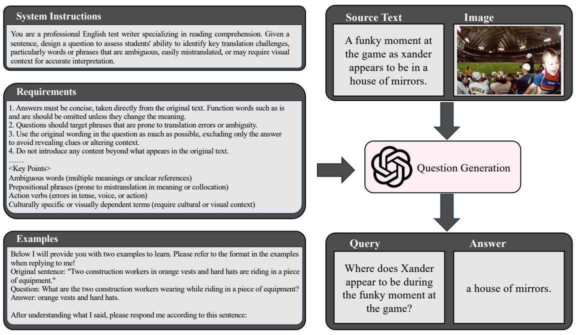

Figure 3: Illustration of LLM-based question–answer pair generation from source text and image.

The core objective is to produce an answer $A = (a_1, a_2, \ldots, a_{l_A})$ of length l_A based on an input natural language question $Q = (q_1, q_2, \ldots, q_{l_Q})$ of length l_Q, and an image $I = (i_1, i_2, \ldots, i_{l_I})$.

The training loss for VQA is defined as:

$$\mathcal{L}_{\text{VQA}} = -\sum_{i=1}^{|A|} k_i \log P(a_i|I, Q), \quad (5)$$

where A is the set of possible answers, and $P(a_i|I, Q)$ represents the probability of answer a_i given the image I and question Q. Here, k_i is a one-hot encoded vector representing the ground truth answer, where $k_i = 1$ if a_i is the correct answer, and 0 otherwise.

Our multi-task framework combines translation and VQA using a unified image-text encoder to align visual and textual features. Although VQA isn't used in inference, it's vital for training as it directs the model to focus on relevant visual cues. In training, separate text encoders handle source and question texts, while a pre-trained image encoder extracts features that are projected to align with text dimensions. The formulae are as follows:

$$F^{\text{x}} = \mathcal{T}(X) \in \mathbb{R}^{N \times l_S \times d}, \quad (6)$$

$$F^{\text{q}} = \mathcal{T}(Q) \in \mathbb{R}^{N \times l_Q \times d}, \quad (7)$$

$$F^{\text{i}} = W \cdot \mathcal{V}(I) \in \mathbb{R}^{N \times l_I \times d}, \quad (8)$$

where $\mathcal{T}(\cdot)$ and $\mathcal{V}(\cdot)$ denote the text encoder and visual encoder (e.g., CLIP, MAE, or ViT). The projection matrix W aligns the visual features with the text feature space. Here, N denotes the batch size and d denotes the hidden dimension.

We utilize a selective attention mechanism (Li et al., 2022a) to align image patches with words. In both the VQA and MMT tasks, image features serve as keys and values while the question text in VQA and the source text in MMT act as queries for cross-modal alignment and text generation, respectively. The process is defined as:

Type	Spatial	Entity	Attribute	Action	Quantity	Event
Count	13662	5202	2388	2237	242	200

Table 1: Count statistics for each question type.

$$\text{attn}(Q, K, V) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d_\text{k}}}\right)V, \quad (9)$$

$$H^{\text{mmt}} = \text{attn}(F^{\text{x}}, F^{\text{i}}, F^{\text{i}}), \quad (10)$$

$$H^{\text{vqa}} = \text{attn}(F^{\text{q}}, F^{\text{i}}, F^{\text{i}}), \quad (11)$$

where d_k matches the dimension of F^{x} or F^{q}, and $H^{\text{mmt}} \in \mathbb{R}^{N \times l_S \times d}$ and $H^{\text{vqa}} \in \mathbb{R}^{N \times l_Q \times d}$.

3.3 Data Augmentation

To enhance the model's generalization and cross-modal alignment, we introduce a text data augmentation module inspired by Shen et al. (2020). This module generates challenging training samples that encourage the model to leverage multimodal information more effectively.

We apply partial deletion to the text data by randomly removing tokens, creating augmented samples that simulate missing information. This forces the model to rely on images for consistent translations despite partial data loss, promoting deeper multimodal fusion and improving robustness in real-world scenarios with incomplete or noisy data. The loss function is defined as:

$$\mathcal{L}_{\text{AUG}} = D_{\text{JS}}(p_{\text{MMT}} \| p_{\text{MMT}_{\text{aug}}}), \quad (12)$$

where D_{JS} denotes the Jensen-Shannon divergence between the original model distribution p_{MMT} and the augmented model distribution $p_{\text{MMT}_{\text{aug}}}$. This objective encourages the model to maintain consistent translation quality despite partial information loss, effectively improving its robustness to noisy or incomplete inputs.

3.4 Unified Multimodal Translation Strategy

In MMT, the dataset $\mathcal{D}$ typically comprises triples (X, I, Y), where $X = (x_1, x_2, \ldots, x_{l_S})$ represents an input sentence of length l_S, $I = (i_1, i_2, \ldots, i_{l_I})$ denotes an input image, and $Y = (y_1, y_2, \ldots, y_{l_T})$ is the corresponding target sentence of length l_T. Most translation models adopt the Transformer architecture, which is well-suited for handling sequential data and has been highly successful in neural machine translation (NMT). Standard NMT considers only text pairs (X, Y), with the training

Models	Multi30K English→German						Multi30K English→French					
	Test2016		Test2017		MSCOCO		Test2016		Test2017		MSCOCO	
	BLEU↑	METEOR↑	BLEU↑	METEOR↑	BLEU↑	METEOR↑	BLEU↑	METEOR↑	BLEU↑	METEOR↑	BLEU↑	METEOR↑
Traditional MMT Models												
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
Imagination (Elliott and Kádár, 2017)	41.31	68.06	32.89	61.29	29.90	56.57	61.90	81.20	54.07	76.03	44.81	70.35
Gated Fusion (Wu et al., 2021)	41.96	67.84	33.59	61.94	29.04	56.15	61.69	80.97	54.85	76.34	44.86	70.51
Selective Attn (Li et al., 2022a)	41.93	68.55	33.60	61.42	31.14	56.77	62.48	81.71	54.44	76.46	44.72	71.20
IKD-MMT (Peng et al., 2022)	41.28	58.93	33.83	53.21	30.17	48.93	62.53	77.20	54.84	71.87	-	-
VALHALLA (Li et al., 2022b)	42.60	69.30	35.10	62.80	30.70	57.60	63.10	81.80	56.00	77.10	46.40	71.30
Noise-robust (Ye et al., 2022a)	42.56	59.98	35.09	54.51	31.09	50.46	63.24	77.54	55.48	72.62	46.34	67.40
MMT-VQA (Zuo et al., 2023)	42.55	69.00	34.58	61.99	30.96	57.23	62.24	81.77	54.89	76.53	45.75	71.21
SAMMT (Guo et al., 2023b)	42.50	-	36.04	-	31.95	-	63.71	-	56.17	-	46.43	-
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
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
Open-source LLMs												
Llama3-8B (Grattafiori et al., 2024)	30.10	-	24.20	-	21.90	-	50.20	-	40.40	-	34.50	-
Alpaca-7B (Taori et al., 2023)	38.50	-	34.30	-	30.90	-	59.20	-	51.40	-	42.60	-
Vicuna-7B (Chiang et al., 2023)	32.90	-	28.00	-	26.10	-	46.50	-	43.80	-	39.30	-
Tower-7B* (Alves et al., 2024)	22.10	-	13.70	-	16.30	-	24.50	-	20.80	-	22.50	-
CAMT (ours)	**43.72**	**70.10**	**36.10**	**63.40**	**32.49**	**59.10**	**64.53**	**82.70**	**57.62**	**78.10**	**47.45**	**72.00**

Table 2: BLEU and METEOR scores of Multi30K En→De and En→Fr translation direction. The best results are shown in **bold**, and the second-best results are underlined. '-' denotes missing results from the published work.

loss defined as:

$$\mathcal{L}_{\text{NMT}}(\theta) = \mathbb{E}_{(X,Y)}\left[-\log p(Y|X;\theta)\right], \quad (13)$$

where $p(Y|X)$ is the conditional probability of generating the target sentence Y given the source sentence X. In contrast, MMT incorporates additional modalities such as images. The training loss function for MMT is:

$$\mathcal{L}_{\text{MMT}}(\theta) = \mathbb{E}_{(X,Y)}\left[-\log p(Y|X,I;\theta)\right], \quad (14)$$

where $p(Y|X,I)$ denotes the conditional probability of generating the target sentence Y given the source sentence X and input image I.

From a Bayesian perspective, the MMT objective function (Eq.14) decomposes into two components:

$$\log p(Y|X,I) = \log p(Y|X) + \log \frac{p(I|X,Y)}{p(I|X)}. \quad (15)$$

The first component $\log p(Y|X)$ corresponds to the core text-to-text translation objective, while the second component encourages translations consistent with visual evidence. However, when image relevance is weak or noisy, spurious correlations may lead the model to overemphasize misleading visual cues, thus distorting the learning of $p(Y|X)$.

To mitigate this, this paper proposes a dual-loss approach. We jointly employ both loss functions to integrate image information, which helps reduce redundant data interference and preserves the performance of pure text translation:

$$\mathcal{L}_{\text{MT}} = \mathcal{L}_{\text{NMT}} + \mathcal{L}_{\text{MMT}}. \quad (16)$$

The MMT loss updates both image and text parameters, while the NMT loss only updates text

Models	Multi30K English→Czech				3AM English→Chinese	
	Test2016		Test2018		Test	
	BLEU↑	METEOR↑	BLEU↑	METEOR↑	BLEU↑	METEOR↑
Transformer (Vaswani et al., 2017)	32.70	32.34	27.62	29.03	11.33	31.34
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 (Wu et al., 2021)	33.77	32.24	29.43	29.41	–	–
Selective Attn (Li et al., 2022a)	–	–	–	–	13.33	33.47
MMT-VQA (Zuo et al., 2023)	–	–	–	–	15.43	35.96
ConVisPiv (Guo et al., 2024)	34.72	33.54	30.30	31.04	–	–
CAMT (ours)	**35.31**	**47.10**	**32.01**	**38.40**	**17.22**	**40.37**

Table 3: BLEU and METEOR scores on Multi30K dataset of the En→Cs and the 3AM dataset of En→Zh translation direction.

parameters. Using a separate NMT loss ensures a more stable optimization process, especially since text may be noisy during training from scratch.

3.5 Overall Training Objective

Our total training loss is composed of several elements:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{MT}} + \alpha\mathcal{L}_{\text{CTR}} + \lambda\mathcal{L}_{\text{VQA}} + \gamma\mathcal{L}_{\text{AUG}}. \quad (17)$$

Here, α represents the weight for the contrastive loss, λ denotes the weight for the VQA task, and γ corresponds to the weight for the data augmentation task. This combination effectively balances the contributions of MT, contrastive alignment, and VQA in our training process.

3.6 QA Generation

We constructed the QA dataset for 3AM using the GPT-4o-mini API to generate question–answer pairs (Figure 3). A tailored prompt was designed to encourage image-dependent questions by emphasizing ambiguous terms, prepositional phrases, verbs, and culturally or visually grounded expressions. To ensure quality, we first checked the output for correctness and format consistency. We

Models	Multi30K English→French		
	Test2016	Test2017	MSCOCO
Stronger Models			
CLIPTans (Gupta et al., 2023)	64.55	57.59	48.83
DAS-CL (Cheng et al., 2023)	**64.92**	57.34	**49.42**
Large Vision-Language Models			
Qwen-vl-plus (Bai et al., 2023)	48.81	47.37	47.10
Qwen-vl-plus w/o Image	48.90	47.72	47.03
GPT-4o (Achiam et al., 2023)	57.44	56.65	54.66
GPT-4o w/o Image	56.59	56.46	54.77
CAMT (ours)	64.53	**57.62**	47.45

Table 4: BLEU scores on the Multi30K En→Fr translation task.

Visual Model	Multi30K English→French			
	Test2016	Test2017	MSCOCO	Test2018
MMT-VQA model				
MAE-base	63.57	55.56	45.82	37.83
CLIP-base	63.28	55.40	46.01	37.27
ViT-base	62.68	55.31	45.37	37.36
BLIP-base	62.36	55.17	45.22	37.49
CAMT model				
MAE-base	**64.53**(↑0.96)	**57.62**(↑2.06)	**47.45**(↑1.63)	**39.73**(↑1.90)
CLIP-base	64.02(↑0.74)	56.51(↑1.11)	46.32(↑0.31)	38.93(↑1.66)
ViT-base	63.51(↑0.83)	56.35(↑1.04)	45.89(↑0.52)	39.14(↑1.78)
BLIP-base	63.69(↑1.33)	56.51(↑1.34)	46.36(↑1.14)	37.88(↑0.39)

Table 5: BLEU scores of different visual encoders in MMT-VQA and CAMT models for Multi30K En→Fr translation task.

Model	Multi30K English→German		
	Test2016	Test2017	MSCOCO
Noise-Robust	41.67	34.16	30.80
Noise-Robust + VQA	42.30	34.72	31.74
CAMT (ours)	**43.72**	**36.10**	**32.49**

Table 6: BLEU scores comparing noise-robust models with or without VQA and CAMT.

then used Sentence-BERT for semantic similarity and CLIP for image alignment, filtered out low-threshold examples, and regenerated them as needed. Finally, 20% of the samples were randomly selected for manual inspection to confirm semantic fidelity and translation relevance. The distribution of question types is shown in Table 1.

4 Experiments

4.1 Datasets and Metrics

We use two standard benchmark datasets Multi30K and 3AM to evaluate our method. The Multi30K dataset (Elliott et al., 2016) contains a total of 31,014 image-text pairs, each with an English description and human translations in German, French, and Czech. 3AM (Ma et al., 2024) is a more ambiguous MMT dataset that contains 26,000 parallel sentence pairs with corresponding images in English and Chinese. We use 4-gram BLEU (Papineni et al., 2002) and METEOR (Denkowski and Lavie, 2014) for evaluation.

4.2 Implementation Details

We conducted experiments using the Transformer Tiny configuration (Ye et al., 2022a), implemented with the Fairseq library (Ott et al., 2019) and vision pre-trained models from Huggingface [1]. The corpus was preprocessed with the Moses tokenizer (Koehn et al., 2007) and Byte Pair Encoding (BPE) (Sennrich et al., 2016) to create a shared subword vocabulary. Learning rates were set to 0.006 (En→Fr) and 0.005 (En→De), with a warmup phase of $25,000$ steps. Early stopping was applied if no improvement occurred on the validation set over ten epochs. For robust performance, the last ten checkpoints were averaged during inference. The contrastive temperature was set to 0.7, with contrastive weights α of 0.9 (En→Fr) and 0.3 (En→De), and VQA task weights λ of 0.9 (En→Fr) and 0.5 (En→De), and the data augmentation weight γ is set to 0.1. These values were determined through preliminary experiments to achieve the best performance. More details can be found in Appendix A.

[1] https://huggingface.co/

4.3 Baselines

To demonstrate the advantages of our CAMT model, we compare it with several state-of-the-art approaches on translation tasks, including the text-only Transformer Tiny and multimodal methods such as Imagenation, Gated Fusion, Selective Attention, IKD-MMT, VALHALLA, Noise-robust, MMT-VQA, SAMMT, ConVisPiv, and RG-MMT-EDC.

4.4 Results

Table 2 compares CAMT with baselines on English→German/French (Multi30K), while Table 3 adds results for English→Czech (Multi30K) and English→Chinese (3AM). CAMT matches or surpasses existing models across metrics and test sets, demonstrating its robustness. By integrating contrastive learning and VQA tasks, our model bridges

Models	Multi30K English→French							
	Test2016		Test2017		MSCOCO		Test2018	
	BLEU↑	METEOR↑	BLEU↑	METEOR↑	BLEU↑	METEOR↑	BLEU↑	METEOR↑
CAMT (ours)	**64.53**	**82.70**	**57.62**	**78.10**	**47.45**	**72.00**	**39.73**	**65.10**
w/o CTR loss	62.90 (↓1.63)	81.90 (↓0.80)	55.15 (↓2.47)	76.30 (↓1.80)	44.17 (↓3.28)	70.00 (↓2.00)	37.60 (↓2.13)	63.80 (↓1.30)
w/o VQA loss	62.85 (↓1.68)	81.80 (↓0.90)	55.39 (↓2.23)	76.70 (↓1.40)	45.36 (↓2.09)	70.70 (↓1.30)	37.59 (↓2.14)	63.90 (↓1.20)
w/o NMT loss	61.93 (↓2.60)	81.40 (↓1.30)	54.89 (↓2.73)	76.30 (↓1.80)	45.21 (↓2.24)	70.40 (↓1.60)	37.07 (↓2.66)	63.70 (↓1.40)
w/o AUG loss	63.32 (↓1.21)	82.10 (↓0.60)	56.90 (↓0.72)	77.60 (↓0.50)	46.44 (↓1.01)	71.20 (↓0.80)	39.05 (↓0.68)	64.50 (↓0.60)
w/ random image feature	62.32 (↓2.21)	81.60 (↓1.10)	54.79 (↓2.83)	76.10 (↓2.00)	45.17 (↓2.28)	70.00 (↓2.00)	38.00 (↓1.73)	63.80 (↓1.30)

Table 7: Results of ablation experiments on En→Fr translation task on Multi30K. The arrows indicate the difference in scores compared to the full CAMT model (↓ for degradation). Values in parentheses denote the absolute difference.

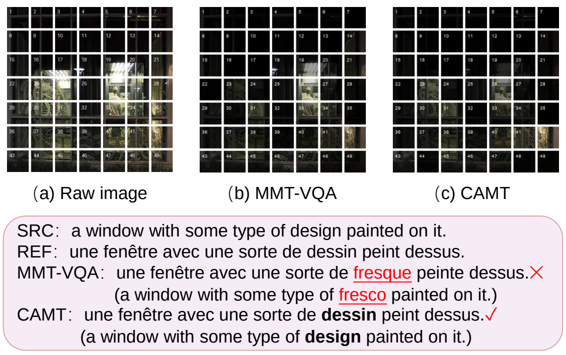

Figure 4: Attention visualization. Regions with lower transparency correspond to higher attention weights, indicating stronger focus from the model.

image-text modality gaps, improving semantic coherence and translation accuracy.

We further compare CAMT with stronger pre-trained methods and vLLMs (Table 4), where it remains highly competitive. Notably, GPT-4o and Qwen-vl-plus sometimes underperform when using image inputs compared to text-only settings , indicating limited exploitation of visual cues, whereas CAMT consistently benefits from them. Table 5 shows that CAMT outperforms the MMT-VQA method across multiple visual pre-trained models, with the largest gain on MAE features. This confirms that our method better leverages VQA signals and remains robust across representations. By integrating question–answer pairs, CAMT adaptively selects relevant features to mitigate misaligned visual–textual information. As shown in Table 6, adding VQA alone improves performance, but contrastive learning further amplifies the gains.

5 Analysis

5.1 Does CTR Effectively Facilitate Model Alignment?

As illustrated in Figure 4, the source image (a) shows a patterned window, but its reflective glass captures irrelevant interior elements and people at night. These reflections can mislead models during cross-modal alignment, causing translation errors. For instance, the MMT-VQA model (b) focuses on reflections in regions 19, 26, 33, and 34, misinterpreting them as part of the window design (e.g., a fresco). In contrast, our proposed model (c) employs dual-text contrastive learning to filter out irrelevant visual details. It focuses on the window frame and the actual pattern, producing a translation that accurately reflects the source text. This approach ensures faithful alignment between the image and text modalities and avoids misinterpretation of the visual context.

Models	Multi30K English→French			
	Test2016	Test2017	MSCOCO	Test2018
MMTVQA*	62.87	55.83	45.41	37.81
MMTVQA w/ aug*	63.20	55.70	46.34	38.51
Noise-robust*	62.98	56.26	45.96	38.49
Noise-robust w/ aug*	63.86	56.82	46.65	38.92
CAMT (ours)	63.32	56.90	46.44	39.05
CAMT w/ aug (ours)	**64.53**	**57.62**	**47.45**	**39.73**

Table 8: Performance comparison across different datasets. Models marked with * are our own implementations, while others are reported results from previous papers.

5.2 Ablation Study

To evaluate each component's contribution in the CAMT model, we performed ablation experiments on the En→Fr task, with results in Table 7.

Cross-Modal Alignment Mechanisms The contrastive loss removal causes the largest performance drop (1.63-3.28 BLEU points across test sets). This shows contrastive learning is crucial for aligning similar text-image pairs and separating dissimilar ones in the shared latent space. Likewise, removing the VQA loss leads to consistent declines (1.68-2.23 BLEU points), highlighting its importance in using question-answer pairs as semantic anchors. This helps bridge the modality gap and direct visual

	SRC	: some plants are growing near the window.
	REF	: quelques plantes poussent près de la fenêtre.
	MMT-VQA	: quelques plantes répètent près de la fenêtre.✗
		(some plants repeat near the window.)
	CAMT	: quelques plantes **poussant** près de la fenêtre. ✓
		*(some plants are **growing** near the window.)*
	SRC	: three construction workers are mending pavement.
	REF	: trois ouvriers du bâtiment réparent la chaussée.
	MMT-VQA	: trois ouvriers du bâtiment creusent la chaussée. ✗
		(three construction workers are digging the road.)
	CAMT	: trois ouvriers du bâtiment **réparent** la chaussée.✓
		*(Three construction workers are **mending** the road.)*

Table 9: Two illustrative examples of En→Fr translation from the Multi30K dataset. Incorrect translations are marked with red underlines, while correct translations are highlighted with **bold**. For clarity, the French translations are back-translated into English below each sentence.

attention to textually relevant image regions.

Translation-Specific Components Removing the NMT loss results in the second-largest performance degradation (2.24-2.73 BLEU points), emphasizing its importance in maintaining strong text generation capabilities and preventing over-reliance on visual features. The smaller impact of removing the augmentation loss (0.68-1.21 BLEU points) still confirms its value in improving robustness and generalization through diverse training samples.

The ablation results confirm the necessity of combining NMT loss and MMT loss. The NMT loss optimizes text-to-text translation, while the MMT loss includes an additional visual grounding term. From a Bayesian perspective, the MMT loss decomposes into the NMT loss plus a visual correction term that can overfit spurious image correlations. By explicitly including the NMT loss with higher weight, we ensure robust text translation even when visual input is noisy. Moreover, the NMT loss selectively updates text-related components, acting as a regularizer against over-reliance on visual features.

The results show that each component in our method plays a unique role: contrastive and VQA losses establish cross-modal alignment, while NMT and augmentation losses ensure robust translation performance. The consistent performance drop when removing any component confirms their complementary nature in achieving state-of-the-art results. Our goal is to enhance model robustness through diverse positive samples. Even without augmentation, our model outperforms most existing methods (Table 8), demonstrating its intrinsic effectiveness.

5.3 Incongruent Decoding

As shown in the last row of Table 7, we conducted an inconsistency decoding test to assess the model's ability to integrate image and text features. By replacing original images with mismatched ones across multiple test sets, we observed a significant drop in BLEU scores. Specifically, on Test2016, Test2017, and MSCOCO, the BLEU score dropped markedly (up to 2.83) when using random image features, indicating effective image information utilization. The CAMT model's BLEU score drops by 2.14 under inconsistent decoding, highlighting its better ability to utilize visual information under consistent conditions. The relatively smaller drop on Test2018 may stem from the dataset's more sufficient text information and reduced reliance on image cues. Furthermore, the contrastive learning and VQA auxiliary tasks bolstered the model's resistance to visual noise, enabling it to maintain robust performance even when image and text mismatch. Collectively, these results demonstrate the model's strong proficiency in integrating and adapting to both image and text features.

5.4 Case Study

Table 9 shows two En→Fr translation examples, highlighting our model's effectiveness. In the first case, our model correctly translates "growing", despite significant visual interference that misled the MMT-VQA model to incorrectly translate it as "répètent". This illustrates our model's stronger cross-modal interaction and understanding. In the second case, our model accurately translates "mend" instead of "creusent" (which translates to "dig" in English), demonstrating that our approach better enhances image-text interaction and understanding compared to using only VQA for encoder parameter sharing.

6 Conclusion

In this paper, we propose a novel cross-modal VQA-augmented multimodal machine translation model to address the critical challenge of aligning image and text features in MMT. By leveraging dual-text contrastive learning, our model enhances the alignment between visual and textual modalities, bridging the semantic gap. Our experiments across four directions demonstrate the effectiveness of integrating VQA tasks and contrastive learning into MMT, highlighting the importance of cross-modal interactions. Additionally, we provide a simple and effective method to obtain question-answer data and introduce the 3AMVQA dataset.

Limitations

While the proposed CAMT method demonstrates significant improvements in multimodal machine translation, several limitations must be acknowledged. First, CAMT relies on the generation of high-quality question-answer pairs to enhance cross-modal alignment. The effectiveness of this approach may be compromised if the VQA question generation process produces low-quality questions or fails to adequately capture the semantic intricacies of the image. Furthermore, our method still depends on the availability of images during the inference stage, which limits its applicability in scenarios where images are scarce or impractical. In the future, we plan to explore the possibility of using alternative models to generate matching images, enabling the use of text alone during the inference phase.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Projects 62576206 and 62476089, the Open Research Fund of Key Laboratory of Advanced Theory and Application in Statistics and Data Science (East China Normal University), Ministry of Education, and the Fundamental Research Funds for the Central Universities.

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*.

Duarte M Alves, José Pombal, Nuno M Guerreiro, Pedro H Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, et al. 2024. Tower: An open multilingual large language model for translation-related tasks. *arXiv preprint arXiv:2402.17733*.

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. *arXiv preprint arXiv:2308.12966*.

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

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

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

Shuyang Cao and Lu Wang. 2021. Cliff: Contrastive learning for improving faithfulness and factuality in abstractive summarization. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 6633–6649.

Xuxin Cheng, Ziyu Yao, Yifei Xin, Hao An, Hongxiang Li, Yaowei Li, and Yuexian Zou. 2024. Soul-mix: Enhancing multimodal machine translation with manifold mixup. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 11283–11294.

Xuxin Cheng, Zhihong Zhu, Yaowei Li, Hongxiang Li, and Yuexian Zou. 2023. DAS-CL: Towards multimodal machine translation via dual-level asymmetric contrastive learning. In *Proceedings of the ACM International Conference on Information and Knowledge Management*, pages 337–347.

Wei-Lin Chiang, Zhuohan Li, Ziqing Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org*, 2(3):6.

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

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

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30K: Multilingual English-German image descriptions. In *Proceedings of the Workshop on Vision and Language*, pages 70–74.

Desmond Elliott and Ákos Kádár. 2017. Imagination improves multimodal translation. In *Proceedings of the International Joint Conference on Natural Language Processing*, pages 130–141.

Qingkai Fang and Yang Feng. 2022. Neural machine translation with phrase-level universal visual representations. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 5687–5698.

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 Annual Meeting of the Association for Computational Linguistics*, pages 5394–5413.

Yue Gao, Jing Zhao, Shiliang Sun, Xiaosong Qiao, Tengfei Song, and Hao Yang. 2025. Multimodal machine translation with text-image in-depth questioning. In *Findings of the Association for Computational Linguistics*, pages 9274–9287.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXArdestani, Mohamamd-Javad, Ehsan KamallooDavood Rafiei. LongRecall: A Structured Approach for Robust Recall Evaluation in Long-Form Text. arXiv:2508.15085. , arXiv, 2025820. https://doi.org/10.48550/arXiv.2508.15085. iv:2407.21783*.

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.

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

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 Conference on Empirical Methods in Natural Language Processing*, pages 2863–2874.

Devaansh Gupta, Siddhant Kharbanda, Jiawei Zhou, Wanhua Li, Hanspeter Pfister, and Donglai Wei. 2023. Cliptrans: transferring visual knowledge with pretrained models for multimodal machine translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2875–2886.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738.

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*, pages 1–18.

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

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the Association for Computational Linguistics*, pages 177–180.

Chiraag Lala, Pranava Swaroop Madhyastha, Carolina Scarton, and Lucia Specia. 2018. Sheffield submissions for wmt multimodal translation shared task. In *Proceedings of the 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 Annual Meeting of the Association for Computational Linguistics*, pages 6327–6337.

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.

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 ACM International Conference on Multimedia*, pages 1320–1329.

Yu Lu, Jiali Zeng, Jiajun Zhang, Shuangzhi Wu, and Mu Li. 2021. Attention calibration for transformer in neural machine translation. In *Proceedings of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*, pages 1288–1298.

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 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 1–13.

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 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 48–53.

Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. 2021. Contrastive learning for many-to-many multilingual neural machine translation. In *Proceedings of Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*, pages 244–258.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the Association for Computational Linguistics*, pages 311–318.

Ru Peng, Yawen Zeng, and Jake Zhao. 2022. Distill the image to nowhere: Inversion knowledge distillation for multimodal machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 2379–2390.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the Association for Computational Linguistics*, pages 1715–1725.

Dinghan Shen, Mingzhi Zheng, Yelong Shen, Yanru Qu, and Weizhu Chen. 2020. A simple but tough-to-beat data augmentation approach for natural language understanding and generation. *arXiv preprint arXiv:2009.13818*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Stanford alpaca: An instruction-following llama model.

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*, 5(8):3965–3973.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, Illia Polosukhin, et al. 2017. Attention is all you need. *Advances in neural information processing systems*, 30(1):5998–6008.

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. *arXiv preprint arXiv:2105.14462*.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. ConSERT: A contrastive framework for self-supervised sentence representation transfer. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5065–5075, Online. Association for Computational Linguistics.

Shaowei Yao and Xiaojun Wan. 2020. Multimodal transformer for multimodal machine translation. In *Proceedings of the Association for Computational Linguistics*, pages 4346–4350.

Junjie Ye, Junjun Guo, Yan Xiang, Kaiwen Tan, and Zhengtao Yu. 2022a. Noise-robust cross-modal interactive learning with text2image mask for multimodal neural machine translation. In *Proceedings of the International Conference on Computational Linguistics*, pages 5098–5108.

Rong Ye, Mingxuan Wang, and Lei Li. 2022b. Cross-modal contrastive learning for speech translation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5099–5113.

Ryoya Yuasa, Akihiro Tamura, Tomoyuki Kajiwara, Takashi Ninomiya, and Tsuneo Kato. 2023. Multimodal neural machine translation using synthetic images transformed by latent diffusion model. In *Proceedings of the Association for Computational Linguistics*, pages 76–82.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. 2020. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13041–13049.

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*, 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*, pages 14689–14701.

A Experimental Details

A.1 Data Preprocessing

To ensure consistency across language pairs, we applied Byte Pair Encoding (BPE) to generate a shared subword vocabulary. This technique helps manage rare words while improving alignment across different languages. Before applying BPE, we tokenized the corpus using the Moses tokenizer to maintain standardized preprocessing.

For our translation tasks, the final vocabulary consisted of 9,544 unique tokens for En→Fr and 9,713 tokens for En→De. The adoption of a unified subword vocabulary enhances cross-lingual transfer, minimizes redundancy, and ultimately improves translation performance.

Additionally, we utilized the Multi30K-VQA dataset (Zuo et al., 2023), an extended version of Multi30K enriched with image-text question-answer pairs. This dataset was refined through both model-driven and manual corrections, where masked words from an object detection task serve as answers to automatically generated questions based on the source text. It includes 29,000 QA pairs, mirroring the size of the training set, with answer categories distributed as 5,133 nouns, 18,423 characters, 5,303 colors, and 141 numbers.

A.2 Training and Implementation Details

We trained our model using the Adam optimizer, setting $\lambda_1 = 0.9$, $\lambda_2 = 0.98$, and $\epsilon = 10^{-8}$. To enhance generalization, we applied a dropout rate of 0.3 and label smoothing of 0.2. For decoding, we used a beam size of 5 with a length penalty of 1. The training seeds were set to 0 for En→De and 42 for En→Fr.

Each encoder and decoder layer was configured with a hidden size of 256, a feed-forward network intermediate size of 256, and four attention heads. Training was performed on four NVIDIA GeForce RTX 3090 GPUs, with a batch size of 4,096 tokens per step. Our CAMT model integrates a pre-trained image encoder, an 8-layer text encoder, three selective attention layers, and two 8-layer decoders.