LAMBDA: Large Language Model-Based Data Augmentation for Multi-Modal Machine Translation

Yusong Wang[1], **Dongyuan Li**[1], **Jialun Shen**[1], **Yicheng Xu**[1], **Mingkun Xu**[2], **Kotaro Funakoshi**[1], **Manabu Okumura**[1]
[1]Tokyo Institute of Technology, Tokyo, Japan
[2]Guangdong Institute of Intelligence Science and Technology
Hengqin, Zhuhai 519031, Guangdong, China
{wangyi, lidy, shenjl, linghan199, funakoshi, oku}@lr.pi.titech.ac.jp, xumingkun@gdiist.cn

Abstract

Multi-modal machine translation (MMT) can reduce ambiguity and semantic distortion compared with traditional machine translation (MT) by utilizing auxiliary information such as images. However, current MMT methods face two primary challenges. The first is their underperformance compared to MT methods based on pre-trained models. The second is the inadequate exploitation and integration of the image modality within the model, primarily due to a lack of triplet training data. A mainstream approach is to introduce large amounts of parallel and monolingual data to train the text model and the visual model separately. However, incorporating extensive external data can result in data imbalance, which may introduce biases during training. Additionally, the collection and cleaning of such large datasets is labor-intensive. To overcome these challenges, we introduce a novel, low-cost, large language model-based data augmentation method called LAMBDA, which can enrich the original samples and expand the dataset without requiring external images and text. We propose a fine-grained image captioning module with a noise filter to hierarchically and accurately extract unexploited information from images. Additionally, we design two specific prompts to guide the GPT-3.5 model in generating enriched texts and the corresponding translations. The enriched samples contain diverse text and strong connections between text and images, leading to significant improvements for MMT baselines, with the highest being an increase of up to 3.83 BLEU score and 3.61 METEOR score.

1 Introduction

Multi-modal machine translation (MMT) integrates additional modalities such as the visual modality (images and videos) to enhance language understanding, which can reduce ambiguity and semantic distortion to improve translation quality (Li et al., 2022a; Lan et al., 2023; Huang et al., 2023). However, some studies have found that it is difficult for MMT methods to surpass strong text-only baselines (Futeral et al., 2023). A consensus on this problem is that most MMT methods are only trained on a small text-image pair dataset called Multi30K (Elliott et al., 2016). This leads to two main issues: (i) The number of image-text pairs is small compared to the text-only corpora used for training MT methods. (ii) The description texts are short and lack diversity. Each text is simple and only describes a part of the corresponding image, as illustrated in Figure 1, which results in a large portion of visual content being neglected, limiting the opportunity for MMT models to explore the correlation between images and texts thoroughly.

Original description
A light-colored dog runs on the beach.

Augmented description (ours)
The light-colored dog runs on the beach near the ocean with footprints in the sand.

Original description
A little boy flinches as a deer kisses him on the cheek.

Augmented description (ours)
A young boy is petting a deer in the zoo, and he flinches as the deer kisses him on the cheek.

Figure 1: Examples from the Multi30K dataset indicate that original annotated text describes only a small portion of corresponding images. After being augmented by LAMBDA, the annotated text covers more aspects of the image. More examples are shown at Appendix A.

A straightforward solution is to develop an extensive multi-modal dataset from scratch, containing diverse samples with detailed annotations. However, this approach incurs substantial costs and requires considerable time for data collection, annotation, and preprocessing. Since the data uti-

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 15240–15253

lized for the MMT task is triplet data, where image data, source language data, and target language data must be aligned (Zhu et al., 2023). The high costs associated with acquiring such triplet data present a significant challenge. Thus, some MMT methods (Futeral et al., 2023; Zhu et al., 2023) attempt to leverage widely available doublet data, including image data and source language data (monolingual captions), and source language data and target language data (sentence translation pairs), from different sources to enhance the capabilities of the visual and text models, respectively. However, incorporating extensive external data from different sources may result in data imbalance, leading the model to favor predicting the more prevalent patterns, thus introducing biases during training (Yang et al., 2022). Besides, data collection and cleaning significantly increase manual resource demands. Another rising direction is using generative models like stable diffusion model (Ho et al., 2020) to generate synthetic images based on text descriptions (Yuasa et al., 2023). Nevertheless, synthetic images may exhibit different distribution patterns compared to authentic images (Guo et al., 2023).

To address the aforementioned problems and circumvent obstacles encountered by existing MMT methods, we introduce a **l**arge l**a**nguage **m**odels-**b**ased **d**ata **a**ugmentation method (LAMBDA). LAMBDA can automatically enrich the text descriptions and enhance the connections between modalities for the samples in the Multi30K dataset without relying on external data. Specifically, we divide the image into different fine-grained-level patches, allowing the less prominent regions of the entire image to be transformed into focal points within these patches. Then we use the BLIP model (Li et al., 2023b) to generate fine-grained captions from different perspectives to obtain a comprehensive image description. Considering the potential noise like incorrect recognition generated by the BLIP model, we implement the CLIP model (Radford et al., 2021) to assess the relevance between the generated captions and the corresponding image patches, discarding unrelated and incorrect captions. Subsequently, we design two specific prompts to guide the GPT-3.5 model (Brown et al., 2020) in logically enriching the original description by integrating the different fine-grained-level captions with the original description and generating corresponding, same-style translations using in-context learning (Brown et al., 2020). Finally, we combine the enriched descriptions with the original images and generated translation to create an augmented dataset, which is then trained alongside the original dataset. Our main contributions are summarized as follows:

- We introduce LAMBDA, a novel data augmentation approach that can enrich the original text description and allow for a more comprehensive exploration of the visual content as Figure 1 shown.
- LAMBDA is fully automated and exclusively augments the data within the Multi30K dataset, avoiding the use of external text and images. It prevents data imbalance caused by external sources and greatly reduces the need for manual resources.
- The integration of LAMBDA significantly improves performance for current MMT baselines and outperforms the latest data generation methods on three benchmarks, demonstrating its effectiveness.

2 Related Work

2.1 Multi-modal Machine Translation

A primary research direction in the MMT field is to design a multi-modal fusion framework to bridge the semantic gap between images and text through a well-designed cross-model attention mechanism (Wang et al., 2024b; Li et al., 2023a; Wang et al., 2023, 2024a) for aligning relevant visual information and discarding irrelevant ones. For example, Yao and Wan (2020) proposed a multi-modal transformer that can learn the representations of images based on the text to avoid encoding irrelevant information in images. Li et al. (2022a) designed a selective attention model to correlate words with related image patches. Ye et al. (2022) proposed a cross-modal interactive mask mechanism to construct a text-image relation-aware attention module in the visual transformer encoder, and only relevant visual features are extracted. Although these models achieve great success, they still cannot surpass strong MT methods (Wu et al., 2021) as they are trained on the Multi30K, a small-scale dataset comprising only 30,000 text-image pairs. The texts are mostly short sentences, causing a lack of diversity, which negatively impacts translation quality.

To address the aforementioned issues, the main direction is to introduce extra parallel data *<text src - text tgt>* and monolingual data *<text src - image>* to train the text model and visual model, separately. Compared to methods trained solely on the Multi30K dataset, these methods achieve better performance but still face challenges. For instance, Futeral et al. (2023) used 25.6 million sentences

and 15.8 million sentences for translation pairs to train the language backbone on the English-to-French (En→Fr) and English-to-German (En→De) translation tasks, respectively. Additionally, 3.3 million images aligned with English text were used to train the visual model to predict masked text using only visual information, forcing the MMT model to better utilize visual inputs. However, they directly incorporated additional parallel data from diverse sources such as Wikipedia and TED Talks, and monolingual data from Conceptual Captions (Sharma et al., 2018), which potentially increases the risk of data imbalance, leading to bias during the training stage. Zhu et al. (2023) utilized triplet data *<text src - text tgt - image>*, parallel data, and monolingual data to train both fusion-based and prompt-based models. The features from both models are combined to create an encoder representation that effectively incorporates visual information and achieves strong performance. They collected text from real-world e-commerce data crawled from TikTok Shop and Shopee, which contains a large amount of noise, such as spelling errors and incomplete sentences. In this case, data cleaning is necessary and arduous. To overcome these challenges, we introduce LAMBDA to automatically augment the Multi30K dataset without the need for additional data.

2.2 Large Language Models

The impressive achievements of LLMs have led researchers to explore their potential as generators for different tasks. Such tasks include question generation (Yuan et al., 2023), code generation (Joshi et al., 2023), text infilling (Li et al., 2022b), dataset generation (Chia et al., 2022; Yu et al., 2023), reading content generation (Xiao et al., 2023), etc. The LLM here is utilized to logically fuse different fine-grained captions to enrich the original description and generate same-style translations, thereby providing a detailed and precise description of the corresponding images along with their translations.

3 Methodology

3.1 Task Definition

MMT aims to translate the source token sequence $\mathbf{x} = (x_1, \ldots, x_N)$ with its corresponding supplementary image $\mathbf{i}$ into the target token sequence $\mathbf{y} = (y_1, \ldots, y_M)$ in another language. The current dominant models use the seq2seq framework that comprises a multi-modal encoder and a decoder (Vaswani et al., 2017). Specifically, the encoder takes $\mathbf{x}$ and $\mathbf{i}$ as input to generate a combined intermediate representation $\mathbf{H}$. Subsequently, the decoder estimates $p(\mathbf{y}|\mathbf{H})$ to generate the target sentence by training with Cross-Entropy loss (Good, 1952):

$$\mathcal{L}_{\text{Cross-Entropy}} = -\frac{1}{M}\sum_{j=1}^{M}\log p(y_j|\mathbf{y}_{<j}, \mathbf{H}), \quad (1)$$

where $p(y_j|\mathbf{y}_{<j}, \mathbf{H})$ denotes the probability at the j-th decoding step for token y_j and $\mathbf{y}_{<j}$ is the target-side previous content for y_j. To obtain a more robust and diverse $\mathbf{H}$, we augment $\mathbf{x}$ as $\mathbf{x}' = (x'_1, \ldots, x'_L)$ and provide corresponding $\mathbf{y}' = (y'_1, \ldots, y'_R)$ by LAMBDA. The framework of LAMBDA is shown in Figure 2.

3.2 Fine-grained Image Captioning Module

Given a mini-batch $\mathcal{B} = \{(\mathbf{x}_a, \mathbf{i}_a)\}_{a=1}^{K}$ containing K samples, we divide each image into different fine-grained patches as $\{\{\mathbf{p}_{a,i}\}_{i=1}^{N_a}\}_{a=1}^{K}$ and N_a is the number of patches. In this way, less significant parts of the whole image can be turned into focal points in patches. As Figure 3 shown, each patch focuses on different elements of the image such as the ocean, the dog, and the footprint. This method ensures that subtle, often overlooked details are given attention, thereby capturing fine-grained semantics within the image to enrich the description.

We divide the image into three fine-grained levels as Figure 2 illustrated considering the following three factors: (1) The trade-offs between performance and computational cost. (2) Adding more image specifics might detract from the original meaning of the sentence. (3) Ability of the BLIP model to correctly recognize objects from overly small image patches. Fine-grained levels are macro-level (capture the overall structure and main components of the image), meso-level (smaller groups of elements and their interactions), and micro-level (capture individual elements and fine details), respectively. We use fixed boundary boxes to divide images into equal parts considering it is a straightforward and uniform way to partition images, making the process computationally simpler and less error-prone. Then, we use the pre-trained BLIP-2 model (Li et al., 2023b) to obtain descriptions of each patch. Using this method, we can obtain detailed and diverse descriptions to benefit the text model and encourage the visual model to explore more aspects of the original image.

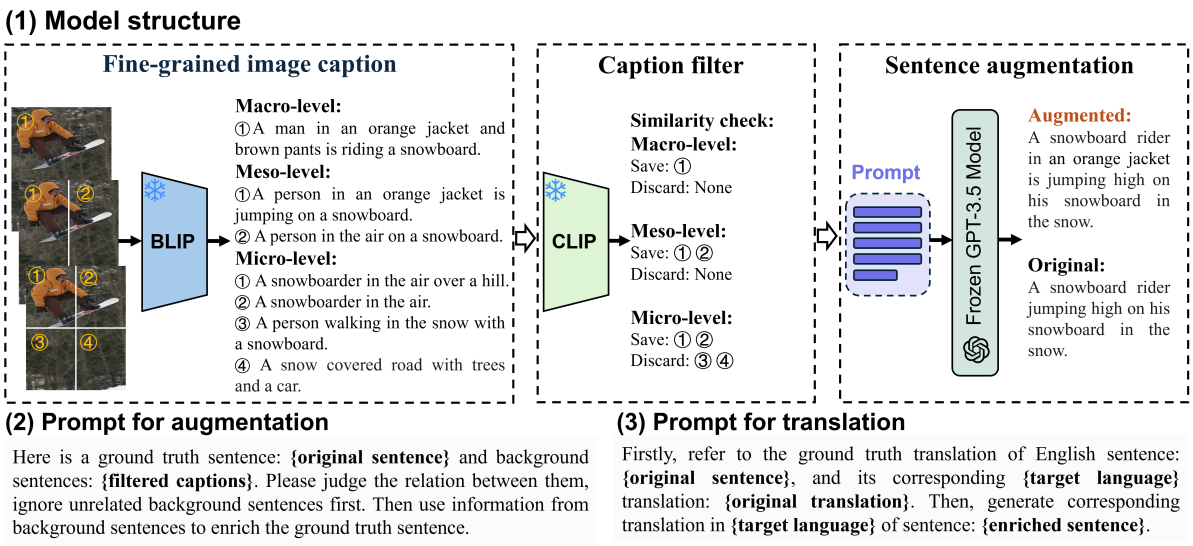

Figure 2: The structure of LAMBDA comprises four principal components: a fine-grained image captioning module, a caption filter, a sentence augmentation prompt, and an in-context learning-based translation generation prompt.

Figure 3: Compared with the left image, the image at the right has been divided into fine-grained patches, allowing for detailed recognition of elements.

3.3 Caption Filter

Considering that the BLIP-2 model may not always achieve precise recognition, it is essential to filter out inaccurate and irrelevant captions. In our approach, we utilize the CLIP model (Radford et al., 2021) to compute the relevance score between each fine-grained caption and its corresponding image patch. Samples with relevance scores below the average are discarded, ensuring only pertinent and accurate descriptions are retained. The retained captions for each sample are combined to form a background sentence list as $S = (s_1, \ldots, s_r)$ where r is the number of retained sentences.

3.4 Sentence Augmentation

In this section, we introduce designed prompts to guide the GPT-3.5 model to enrich a given original sentence and generate accurate, same-style translations. Our approach can be divided into three phases: filtration, enrichment, and translation. The designed prompt formats are presented in Figure 2.

Filtration. We present the GPT-3.5 model with an original description, alongside a series of background sentences S. The first objective of the GPT-3.5 model is to logically check the relationship between the ground truth and each background sentence using its extensive knowledge. Background sentences that do not hold direct contextual relevance to the ground truth sentence are disregarded.

Enrichment. The GPT-3.5 model leverages the contextually relevant background sentences to logically augment the content of the original description. This process enhances the original description in a manner that improves its description without distorting its intrinsic meaning.

Translation. We refer to the concept of in-context learning in LLMs. We use the original sentence and its translation as a reference example in the prompt. In this case, the GPT-3.5 model will follow the reference to translate the enriched sentence. Considering that enriched content mainly involves adding objects to the image, *i.e.*, noun words, the GPT-3.5 model can focus on translating these noun words and incorporating them into the original translation. This strategy simplifies the translation process and prevents the GPT-3.5 model from producing translations with varying patterns and potential inaccuracies, which can lead to a decline in performance.

Model	*English→German*				*English→French*			
	Test2016		Test2017		Test2016		Test2017	
	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
AUG. by LAMBDA & Orig. versus Orig.								
Transformer	42.60 **+1.65**	70.43 **+2.06**	35.79 **+2.39**	65.02 **+3.21**	62.22 **+0.81**	81.67 **+0.89**	56.24 **+3.45**	77.18 **+1.72**
Gated-Fusion	41.37 **+0.70**	68.87 **+1.57**	33.08 **+1.37**	62.70 **+2.47**	59.55 -0.19	79.73 +0.00	53.16 **+1.27**	75.20 **+0.85**
Noise-Robust	42.73 **+1.15**	70.42 **+1.60**	35.64 **+2.09**	64.91 **+2.71**	62.26 **+1.13**	81.56 **+0.94**	56.41 **+2.79**	77.23 **+1.68**
SA	42.54 **+1.77**	70.41 **+2.24**	35.31 **+3.08**	64.71 **+3.61**	61.62 **+0.55**	81.39 **+0.53**	55.60 **+2.59**	76.95 **+1.65**
VGAMT	42.96 **+0.83**	57.98 **+1.17**	37.48 -0.83	52.93 -0.63	63.13 -1.24	77.00 -0.94	59.33 **+0.03**	74.22 -0.24
AUG. by AttrPrompt & Orig. versus Orig.								
Transformer	41.86 +0.91	69.02 +0.65	34.35 +0.95	62.74 +0.93	61.88 +0.47	81.23 +0.45	54.47 +0.90	76.00 +0.54
Gated-Fusion	40.32 -0.35	67.61 +0.31	32.72 +1.01	61.58 +1.35	59.90 **+0.16**	80.07 **+0.34**	52.25 +0.36	74.57 +0.22
Noise-Robust	41.94 +0.36	68.84 +0.02	34.32 +0.77	62.72 +0.52	61.64 +0.51	81.09 +0.47	54.03 +0.41	75.86 +0.31
SA	41.84 +1.07	68.78 + 0.61	33.60 +1.27	62.57 +1.47	60.20 -1.04	80.08 -0.78	52.86 -0.15	74.95 -0.35
VGAMT	41.74 -0.39	57.50 +0.69	37.43 -0.88	52.85 -0.71	61.74 -2.63	75.95 -1.99	57.77 -1.53	73.15 -1.31

Table 1: Performance of MMT models trained with the original and LAMBDA/AttrPrompt-augmented training sets, including the performance gain(+)/drop(-) compared to trained only on the original set. Underline denotes significant improvements according to the two-tailed t-test ($p < 0.05$) with Bonferroni correction. Bold indicates higher improvement between LAMBDA and AttrPrompt in performance gain.

Model	*English→German*				*English→French*			
	Test2016		Test2017		Test2016		Test2017	
	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
Transformer	40.96 +0.01	69.49 +1.12	33.45 +0.05	63.83 +2.02	57.84 -3.75	79.61 -1.17	52.12 -1.45	75.53 -0.07
Gated-Fusion	39.89 -0.78	68.34 +1.04	32.29 +0.58	62.28 +2.05	56.83 -2.91	78.47 -1.26	50.21 -1.68	74.33 -0.02
Noise-Robust	41.02 -0.56	69.53 +0.71	33.51 -0.04	63.78 +1.58	57.67 -3.46	79.65 -0.97	52.17 -1.45	75.56 +0.01
SA	40.75 +0.04	69.44 +1.30	32.99 +0.92	63.70 +2.37	57.27 -3.80	79.35 -1.20	52.00 -1.20	75.59 +0.36
VGAMT	42.19 +0.06	57.19 +0.38	35.49 -2.82	51.38 -2.17	59.52 -4.85	74.41 -3.53	55.80 -3.50	70.73 -3.73

Table 2: Performance of MMT models trained with augmented/original data (AUG. by LAMBDA versus Orig.).

Additionally, by focusing on specific nouns, the model can maintain consistency and accuracy in the enriched content.

The output of this three-phased approach is an enriched sentence with its corresponding German and French translations that encapsulate the essence of the original content while enriching it with greater detail and context. These translations are included to assess the model's ability to accurately generate and maintain detailed captions across different languages, ensuring cross-lingual consistency and performance.

4 Experiments

4.1 Experimental Settings

Datasets. Following previous MMT methods, we use the Multi30K benchmark (Elliott et al., 2016), which is composed of English sentences and images with the corresponding French and German translations. Both training and validation sets have $29,000$ and $1,014$ instances. We used three test sets to evaluate our model, including Test2016, Test2017, and Ambiguous Coco (MSCOCO) (Elliott et al., 2017). The number of instances in each test set is $1,000$, $1,000$, and 461, respectively.

Baselines. To empirically verify the advantages and effectiveness of LAMBDA, we select and reproduce the following state-of-the-art MMT methods as our baselines, including Gated-Fusion (Wu et al., 2021), SA (Selective Attention + ViT-base) (Li et al., 2022a) and Noise-Robust (Ye et al., 2022), VGAMT (Futeral et al., 2023) and a text-only Transformer model (Vaswani et al., 2017) as a uni-modal baseline. Aside from comparing with training on the original data, we also select a cutting-edge augmentation baseline, AttrPrompt (Yu et al., 2023), for comparison. This method employs prompts incorporating multiple attributes to generate training text data with the GPT-3.5 model, thereby enhancing attribute diversity and mitigating biases. Following the approach described by Yuasa et al. (2023), we employ a stable diffusion model to generate corresponding images for the text data produced by AttrPrompt to create the complete augmented data.

Implementation. Following previous works, we used BLEU (Papineni et al., 2002) and METEOR

Model	*English→German* BLEU	*English→German* METEOR	*English→French* BLEU	*English→French* METEOR
AUG. by LAMBDA & Orig. versus Orig.				
Transformer	**+2.48**	**+3.45**	**+4.25**	**+3.11**
Gated-Fusion	**+0.91**	**+2.13**	**+2.76**	**+1.99**
Noise-Robust	**+1.99**	**+2.89**	**+3.83**	**+2.47**
SA	**+2.45**	**+3.25**	**+3.53**	**+3.02**
VGAMT	**+0.32**	**+0.16**	**+0.09**	-0.27
AUG. by AttrPrompt & Orig. versus Orig.				
Transformer	+0.85	+0.86	+1.32	+0.82
Gated-Fusion	+0.45	+0.72	+0.65	+0.55
Noise-Robust	-0.41	-0.19	+1.17	+0.96
SA	+0.94	+0.42	+0.11	-0.20
VGAMT	-0.69	-0.23	-1.34	-0.73
AUG. by LAMBDA versus **Orig.**				
Transformer	+0.15	+1.74	+2.41	+2.74
Gated-Fusion	+0.03	+1.81	+2.04	+2.65
Noise-Robust	-0.18	+1.70	+2.47	+2.39
SA	-0.10	+1.88	+2.62	+2.54
VGAMT	-0.37	-0.07	-1.25	-1.00

Table 3: Performance comparison between LAMBDA, AttrPrompt, and the non-augmentation on the MSCOCO dataset. Underline denotes significant improvements according to the two-tailed t-test ($p < 0.05$) with Bonferroni correction. Bold indicates higher improvement between LAMBDA and AttrPrompt in performance gain.

(Banerjee and Lavie, 2005) as the evaluation metrics for all test sets. All the results are the average of 5 runs with different random seeds. Gated-Fusion, SA, Noise-Robust, and the text-only Transformer model use the same setting with Li et al. (2022a) and Wu et al. (2021). They are trained and evaluated on the Fairseq toolkit (Ott et al., 2019). As for VGAMT, we follow the original settings in its code. Please note that the default BLEU and METEOR evaluation metrics for VGAMT are sacrebleu package [1] and Meteor 1.5 [2], respectively.

4.2 Augmenting Existing Dataset

Here, we quantitatively evaluate the effectiveness and superiority of LAMBDA. First, we combine the data augmented by LAMBDA/AttrPrompt with the original training dataset to form a unified training set. Then, we evaluate the performance of MMT baseline models trained on this combined dataset and compare it with the performance of models trained on the original Multi30K dataset. The results are presented at Table 1 (Test2016 and Test2017 test sets) and Table 3 (MSCOCO test set). Specifically, we have the following findings:

(1) Both data augmentation methods introduce performance gains on most evaluation matrices when combined with the original training set. The augmentation data can be seen as an effective complement to the original training set, which brings new insight for baseline models such as more diverse text and new visual information. Besides, the augmented dataset encourages the MMT model to further explore and utilize visual information, helping to reduce semantic distortion and improving text representation in translation tasks.

(2) LAMBDA consistently delivers better performance compared to AttrPrompt. The main reason might be that synthetic images often exhibit different distribution patterns compared to authentic images (Guo et al., 2023), and they may also depict incorrect scenes or include irrelevant information. Both of them negatively affect the training of the visual model as demonstrated in Section 4.5. Additionally, AttrPrompt generates repeated sentences with a 26.67% repetition rate[3], which limits text diversity and causing data imbalance, leading to a performance drop. In contrast, LAMBDA focuses on enriching original descriptions without using external images or generating new text. Since it works with existing non-repetitive samples, it inherently avoids introducing repetitions.

(3) LAMBDA shows improvement in the En→De translation task performance on the pre-trained language model-based MMT method (VGAMT) but shows a performance drop in the En→Fr translation task. Standard Transformer models trained from scratch often benefit from additional and diverse text data, enhancing their ability to learn from a broader linguistic base. On the other hand, pre-trained language models like mBART, already incorporate extensive prior linguistic knowledge and exhibit less noticeable improvements. In the situation where LLMs like the GPT-3.5 model are used to generate translations for dataset augmentation, these pairs might introduce noise, as the quality and style of generated translations may not fully match the pre-training data. Additionally, the quality of these generated translations might be inferior to genuine human translations, especially when the target language involves complex expressions like French. Pre-trained language models are adept at dealing with various linguistic styles and contexts, this lower quality or stylistic inconsistency might

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

[2] https://www.cs.cmu.edu/~alavie/METEOR

[3] repetition rate = 1 - unique sentences / total sentences

negatively impact the model's performance. Although MMT models can benefit from additional visual information, their performance still largely depends on the quality and consistency of the text.

4.3 Training with Augmented Data

We assess the quality of the generated datasets quantitatively by evaluating the test performance of the baseline models trained on both the augmented data and the original data. From Table 2 and Table 3, we can draw the following conclusions: Firstly, the augmented data with generated translations in the En→De nearly matches the quality of the original data, as the enriched content provides an additional improvement in performance. Secondly, in the En→Fr, deviations often occur in the corresponding generated translations as mentioned in Section 4.2, leading to a performance drop. Compared to training on the unified dataset in Section 4.2, combining original human translations with augmented data helps mitigate biases introduced by LLM-generated data (Meng et al., 2022). The underperformance indicates potential areas for future improvement. Overall, we argue that LLMs have the potential to generate high-quality training data to augment translation datasets using carefully crafted prompts and selected reference samples.

Dataset	Vocabulary size	Objects	Length
Original	9,700	3.88	13.11
Augmented	10,200	5.27	17.72

Table 4: Comparison of original and rewritten datasets in terms of vocabulary size, average number of object count, and average sentence length.

4.4 Diversity Analysis

To quantify the diversity of the augmented training data, we use vocabulary size to evaluate its lexical diversity compared to the original data (Yu et al., 2023). We also present data on the average sentence length and the average number of objects per sentence as supplementary metrics. The statistical results are presented in Table 4 and we can find that the augmented dataset has higher lexical diversity, a greater number of objects, and longer sentences than the original dataset. This indicates its potential to enrich the training process for more nuanced and effective MMT models. However, it is important to recognize that although a greater number of objects may help MMT models to better connect text with corresponding images and explore the various parts of the image, it also increases the difficulty for these models to precisely align text tokens with the corresponding image patches.

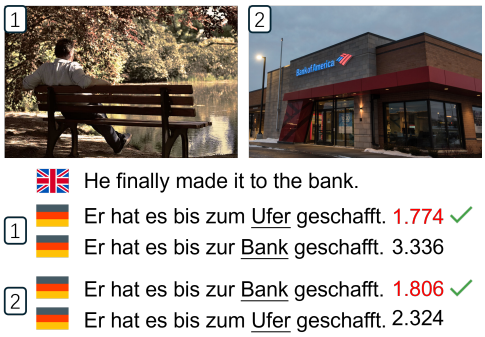

Figure 4: Perplexity scores on a CoMMuTE example. This sample shows that the visual model can extract key information from images to help the text model understand the ambiguous English word "bank".

4.5 Visual Feature Exploration

In this section, we assess the effectiveness of LAMBDA in improving the ability of the visual model. CoMMuTE dataset is a Contrastive Multilingual Multi-modal Translation Evaluation dataset proposed by Futeral et al. (2023) to evaluate the performance of the visual model. This dataset includes 155 English sentences that are lexically ambiguous. For each sentence, there are two translations that capture distinct meanings of the sentence, alongside two images clarifying which translation is accurate. Specifically, MMT models are asked to rank the pairs of translations based on the perplexity score. It covers En→Fr and En→De tasks. An example of this dataset is shown at Figure 4.

Sentences augmented by LAMBDA provide more detailed descriptions along with additional relevant visual details of corresponding images, enabling the visual model to explore the visual elements thoroughly. Here, we validate the effectiveness of LAMBDA in the visual model improvement. We use VGAMT as our foundation model due to its excellent cross-modality alignment capabilities. To obtain a comprehensive comparison and analysis, we divide VGAMT into three different settings: ① The visual model is trained with monolingual pairs from Multi30K dataset (VO). ② The visual model is trained with monolingual pairs from augmented Multi30K dataset (VA). ③ The visual model is trained with monolingual pairs from augmented Multi30K dataset and original Multi30K dataset (VAO). Please note that text-only models like GPT-

3.5 have a 50% chance level accuracy in this test, as they cannot incorporate image information. The results are presented at Table 5.

According to the results, we have the following findings: First, compared with the text-only baseline, VGAMT can make use of visual information to complement text modality to reduce ambiguity. Secondly, both VA and VAO models demonstrate superior performance over VO. This improvement likely stems from the enriched text descriptions covering more parts of the corresponding images, which allows the visual model to further explore the correlation between text and the image. Thirdly, VAO outperforms VA in terms of accuracy because enriched sentences can also introduce noise, such as incorrect objects and descriptions. Integration of original data helps to mitigate such noise and improve model reliability, which can be seen as a residual link. Lastly, LAMBDA outperforms the AttrPrompt method because synthetic images often exhibit different distribution patterns compared to authentic images, as described in Section 4.2, which shows the superiority of LAMBDA over image-generation-based methods in the MMT field. In conclusion, the superior performance demonstrates the effectiveness of LAMBDA in enhancing the capability of the visual model.

Model	CoMMuTE	
	English→German	***English→French***
AttrPrompt-VAO	53.40%	54.74%
LAMBDA-VAO	**55.13%**	**56.69%**
LAMBDA-VA	53.73%	55.20%
LAMBDA-VO	52.80%	54.93%

Table 5: This table shows the performance (Accuracy) of the VGAMT model on the CoMMuTE dataset, trained with different configurations of monolingual data for the En→De and En→Fr translation tasks.

4.6 Performance for Different Lengths

Previous research has proven that long sentences are always difficult to translate (Mishra et al., 2013; Shao et al., 2019). To further demonstrate the effectiveness of LAMBDA, we report the translation performance of MMT models trained on the original dataset and augmented&original dataset for sentences with different lengths. Here, we select the pure-transformer-based baselines, Noise-Robust, SA, and Gated-Fusion, to better demonstrate the advantages of enriched descriptions. Because pre-trained language model-based MMT methods have already been trained on a large amount of text data with varying lengths. Focusing on pure-transformer-based baselines allows us to clearly observe the impact of the augmented data on models that do not benefit from extensive pre-training. The result of Noise-Robust is represented in Figure 5 and others are displayed at Appendix B. Here the BLEU score and METEOR score are calculated by sacrebleu package and NLTK[4] toolkit, respectively. We can see that the pure-transformer-based baselines achieve considerable gains on long sentences when trained with the augmented data. The reason might be that LAMBDA extends the length of the original sentences and enriches their content, enhancing the contextual comprehension and semantic capture capabilities of these MMT models. This allows them to process and maintain long-term dependency relationships effectively.

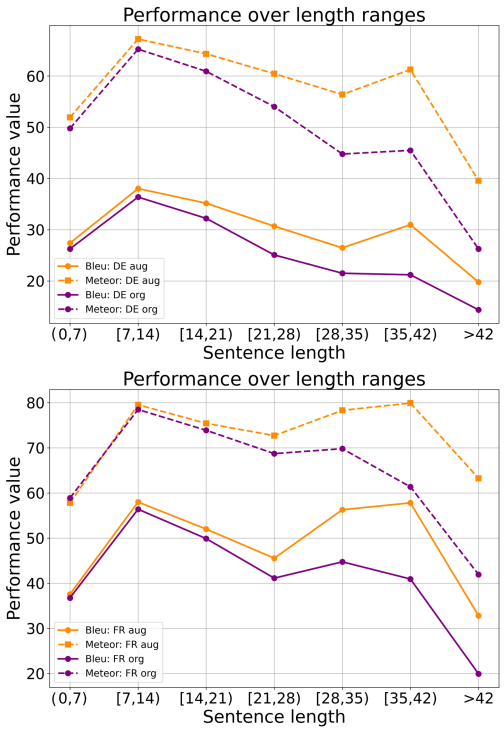

Figure 5: BLEU and METEOR scores of the Noise-Robust under different sentence lengths, trained on both the original dataset and the augmented dataset.

5 Conclusion

In this work, we devise an innovative data augmentation method, LAMBDA, to address faced challenges in the current MMT field. It involves a fine-grained image captioning module, a caption

[4] https://www.nltk.org

filter, a sentence augmentation component, and two specific prompts to effectively augment the original samples from the Multi30K dataset without relying on external image and text data. Comprehensive experiments conducted on three benchmark datasets for En→De and En→Fr translation tasks demonstrate significant improvements in translation performance with current representative MMT methods, validating the effectiveness. Furthermore, results on the CoMMuTE dataset confirm that LAMBDA enhances the MMT model's ability to exploit the visual modality. In future work, we will explore effective methods to reduce noise in the generation stage and incorporate a broader range of large LLMs, including BLOOM, OPT, and GPT-4, to comprehensively compare our method's performance across these diverse models.

Acknowledgements

We would like to thank all the reviewers for their valuable suggestions, which helped us improve the quality of our manuscript. This work was supported by the Key-Area Research and Development Program of Guangdong Province (Grants No. 2021B0909060002), and National Natural Science Foundation of China (Grants No.62204140). Finally, Dongyuan Li acknowledges the support from the China Scholarship Council (CSC).

Limitations

Although LAMBDA has achieved obvious performance improvement across three benchmark datasets with the existing MMT baselines, we still face several limitations in this study, which brightens the way for our future work. Firstly, the effectiveness of LAMBDA is influenced by the availability and quality of pre-trained models such as image-caption models and LLMs. These models cannot achieve perfect accuracy and the noise persists in the augmented data. Secondly, for the caption filter, we set the discarding thresholds based on the average value from our experience and manual evaluation, as thoroughly exploring the optimal settings would incur significant time and cost. In the future, we might explore automated optimization techniques to find the optimal setting. Moreover, one issue with LLMs-based training data generation is the phenomenon of hallucination. The judgment, selection, and combination of fine-grained captions rely on the knowledge of LLMs, which may not always align with accuracy or human logic. This also exists in the translation stage, resulting in inaccurately sequenced words within the translated text. To mitigate this problem, it is possible to leverage additional fact-checking mechanisms to cross-verify the generated text with a reliable knowledge base. This research focuses on whether generated augmented data can benefit downstream translation tasks. Exploring and comparing more effective prompts and their relationship to augmented data and human annotations is not central to this research but will be considered in future work to complement our findings.

References

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Yew Ken Chia, Lidong Bing, Soujanya Poria, and Luo Si. 2022. RelationPrompt: Leveraging prompts to generate synthetic data for zero-shot relation triplet extraction. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 45–57, Dublin, Ireland. Association for Computational Linguistics.

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, Copenhagen, Denmark. Association for Computational Linguistics.

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

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

Irving John Good. 1952. Rational decisions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 14(1):107–114.

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

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc.

Yichong Huang, Xiaocheng Feng, Xinwei Geng, Baohang Li, and Bing Qin. 2023. Towards higher Pareto frontier in multilingual machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3802–3818, Toronto, Canada. Association for Computational Linguistics.

Harshit Joshi, José Cambronero Sanchez, Sumit Gulwani, Vu Le, Gust Verbruggen, and Ivan Radiček. 2023. Repair is nearly generation: Multilingual program repair with llms. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(4):5131–5140.

Zhibin Lan, Jiawei Yu, Xiang Li, Wen Zhang, Jian Luan, Bin Wang, Degen Huang, and Jinsong Su. 2023. Exploring better text image translation with multimodal codebook. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3479–3491, Toronto, Canada. 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, Dublin, Ireland. Association for Computational Linguistics.

Dongyuan Li, Yusong Wang, Kotaro Funakoshi, and Manabu Okumura. 2023a. Joyful: Joint modality fusion and graph contrastive learning for multimoda emotion recognition. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16051–16069, Singapore. Association for Computational Linguistics.

Dongyuan Li, Jingyi You, Kotaro Funakoshi, and Manabu Okumura. 2022b. A-TIP: Attribute-aware text infilling via pre-trained language model. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5857–5869, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Generating training data with language models: Towards zero-shot language understanding. In *Advances in Neural Information Processing Systems*, volume 35, pages 462–477. Curran Associates, Inc.

Abhijit Mishra, Pushpak Bhattacharyya, and Michael Carl. 2013. Automatically predicting sentence translation difficulty. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 346–351, Sofia, Bulgaria. Association for Computational Linguistics.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania. Association for Computational Linguistics.

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

Chenze Shao, Yang Feng, Jinchao Zhang, Fandong Meng, Xilin Chen, and Jie Zhou. 2019. Retrieving sequential information for non-autoregressive neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3013–3024, Florence, Italy. Association for Computational Linguistics.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Annual Meeting of the Association for Computational Linguistics*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Yusong Wang, Dongyuan Li, Kotaro Funakoshi, and Manabu Okumura. 2023. Emp: Emotion-guided multi-modal fusion and contrastive learning for personality traits recognition. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, ICMR '23, page 243–252, New York, NY, USA. Association for Computing Machinery.

Yusong Wang, Dongyuan Li, and Jialun Shen. 2024a. Inter-modality and intra-sample alignment for multimodal emotion recognition. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8301–8305.

Zhen Wang, Dongyuan Li, Guang Li, Ziqing Zhang, and Renhe Jiang. 2024b. Multimodal low-light image enhancement with depth information. In *ACM Multimedia 2024*.

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

Changrong Xiao, Sean Xin Xu, Kunpeng Zhang, Yufang Wang, and Lei Xia. 2023. Evaluating reading comprehension exercises generated by LLMs: A showcase of ChatGPT in education applications. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 610–625, Toronto, Canada. Association for Computational Linguistics.

Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, WENXUAN PENG, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, Xuefeng Du, Kaiyang Zhou, Wayne Zhang, Dan Hendrycks, Yixuan Li, and Ziwei Liu. 2022. Openood: Benchmarking generalized out-of-distribution detection. In *Advances in Neural Information Processing Systems*, volume 35, pages 32598–32611. Curran Associates, Inc.

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

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

Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander J Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. 2023. Large language model as attributed training data generator: A tale of diversity and bias. In *Advances in Neural Information Processing Systems*, volume 36, pages 55734–55784. Curran Associates, Inc.

Xingdi Yuan, Tong Wang, Yen-Hsiang Wang, Emery Fine, Rania Abdelghani, Hélène Sauzéon, and Pierre-Yves Oudeyer. 2023. Selecting better samples from pre-trained LLMs: A case study on question generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12952–12965, Toronto, Canada. Association for Computational Linguistics.

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

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

A Case Study

In this section, we present additional samples from the Multi30K dataset along with their corresponding augmented versions and the case study is presented at Figure 6. These samples indicate that the original annotated text describes only a small portion of corresponding images. After being augmented by LAMBDA, the annotated text encompasses more aspects of the image, demonstrating the effectiveness of LAMBDA.

B Performance for Different Lengths

In this section, we present the translation performance of Gated-Fusion and SA models trained on both the original dataset and augmented dataset for sentences of varying lengths. The results are shown in Figure 7. Based on the results, we observe that these pure-transformer-based baselines consistently achieve significant gains on long sentences. LAMBDA extends the length of the original sentences and enhances the diversity of descriptions, which benefits the contextual comprehension and semantic capture capabilities of these models. Additionally, different baseline models exhibit varying levels of improvement. This variation is might attributed to their differing abilities in multi-modal alignment. A stronger multi-modal alignment capability enables the model to establish more precise connections between the increased objects in enriched descriptions and their corresponding images. This emphasizes the importance of designing a multi-modal fusion framework to bridge the semantic gap between images and text in the MMT field.

Original description
A girl in a bright pink outfit prepares for a race.
Augmented description (ours)
A young girl in a bright pink outfit prepares for a race on the track, crouching down with her hands on the ground.

Original description
A little girl climbing into a wooden playhouse.
Augmented description (ours)
A little girl climbing into a wooden playhouse with a window and flowers.

Original description
This dog doesn't seem to appreciate how awesome his hat is.
Augmented description (ours)
The dog on the couch is not showing any appreciation for the blue hat it is wearing.

Original
A few people wearing green outfits carry a large staff that looks like a cross.
Augmented description (ours)
A group of people wearing green outfits carry a large staff that looks like a cross as they walk down the street.

Original
A man in a suit is running past two other gentleman, also dressed in a suit.
Augmented description (ours)
A man in a suit is running past two other gentlemen, also dressed in a suit, on a city street with tall buildings and cars.

Original
A man with a red jacket is shielding himself from the sun trying to read a piece of paper.
Augmented description (ours)
A man with a red jacket and glasses is shielding himself from the sun trying to read a piece of paper in a crowded city.

Figure 6: These examples are from the Multi30K dataset attached with the original description and the corresponding augmented description generated by LAMBDA.

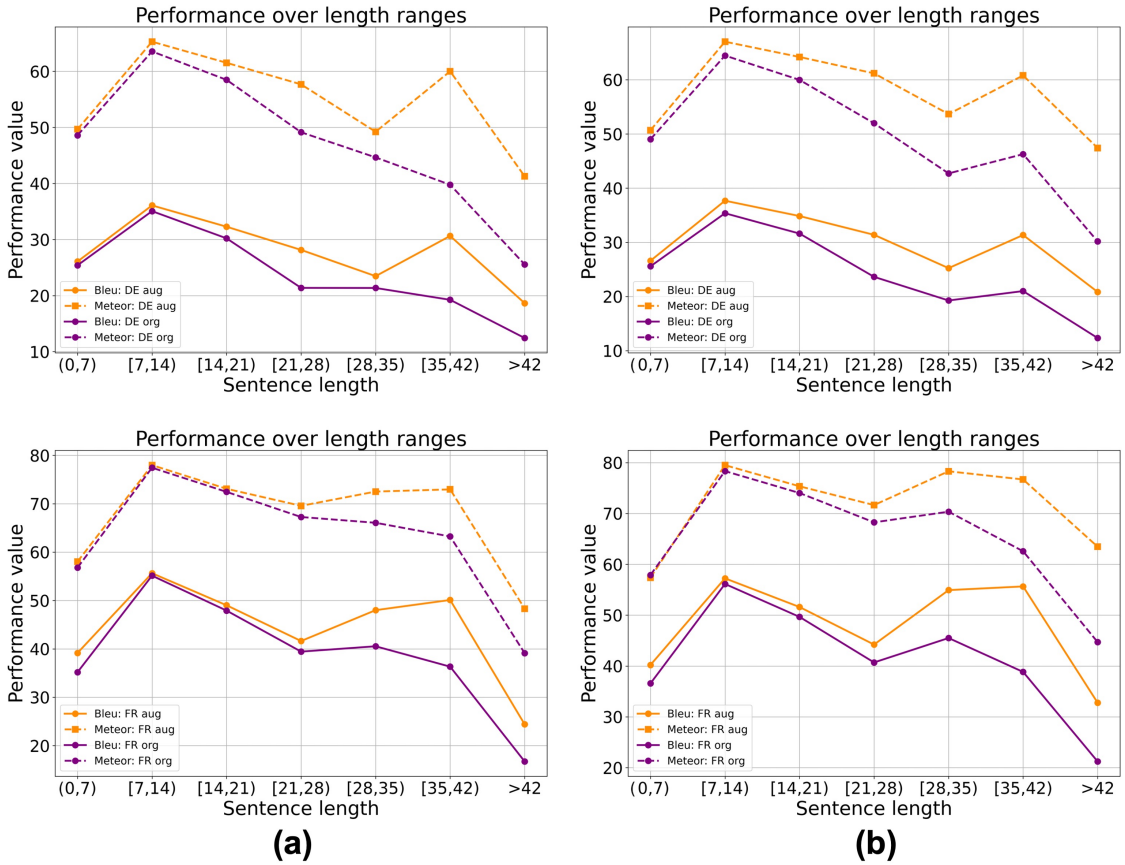

Figure 7: BLEU and METEOR scores of the Gated-Fusion baseline **(a)** and SA baseline **(b)** under different sentence lengths, trained on both the original dataset and the augmented dataset.