Beyond Triplet: Leveraging the Most Data for Multimodal Machine Translation

Yaoming Zhu, Zewei Sun, Shanbo Cheng, Luyang Huang, Liwei Wu, Mingxuan Wang
ByteDance
{zhuyaoming,sunzewei.v,chengshanbo}@bytedance.com
{huangluyang,wuliwei.000,wangmingxuan.89}@bytedance.com

Abstract

Recent work has questioned the necessity of visual information in Multimodal Machine Translation (MMT). This paper tries to answer this question and build a new benchmark in this work. As the available dataset is simple and the text input is self-sufficient, we introduce a challenging dataset called EMMT, whose testset is deliberately designed to ensure ambiguity. More importantly, we study this problem in a real-word scenario towards making the most of multimodal training data. We propose a new framework 2/3-Triplet which can naturally make full use of large-scale image-text and parallel text-only data. Extensive experiments show that visual information is highly crucial in EMMT. The proposed 2/3-Triplet outperforms the strong text-only competitor by 3.8 BLEU score, and even bypasses a commercial translation system. [1]

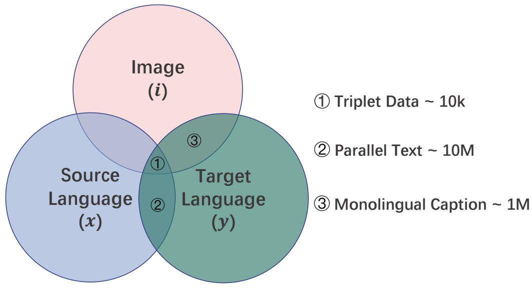


Figure 1: Triple data, although widely utilized in multimodal machine translation, is quite scarce. We emphasize the importance of other two kinds of data: parallel text and image captions. The numbers represent the size of commonly used datasets for the corresponding data type.

1 Introduction

Multimodal Machine Translation (MMT) is a machine translation task that utilizes data from other modalities, such as images. Previous studies propose various methods to improve translation quality by incorporating visual information and showing promising results (Lin et al., 2020; Caglayan et al., 2021; Li et al., 2022a; Jia et al., 2021). However, manual image annotation is relatively expensive; at this stage, most MMT work is applied on a small and specific dataset, Multi30K (Elliott et al., 2016). The current performance of the MMT system still lags behind the large-scale text-only Neural Machine Translation (NMT) system, which hinders the real-world applicability of MMT.

We summarize the limitations of the current MMT in two aspects. The first limitation is the size of the training data. Usually, the performance of MMT heavily relies on the triple training data: parallel text data with corresponding images. The triplets are much rarer for collection and much more costly for annotation than monolingual image-text and parallel text data, as in Figure 1. Considering that current MT systems are driven by a massive amount of data (Aharoni et al., 2019), the sparsity of multimodal data hinders the large-scale application of these systems. Some researchers have proposed retrieve-based approaches (Zhang et al., 2020; Fang and Feng, 2022), aiming to construct pseudo-multimodal data through text retrieval. However, their constructed pseudo-data face problems like visual-textual mismatches and sparse retrieval. Besides, the models still cannot take advantage of monolingual image-text pairs.

The second limitation is the shortage of proper benchmarks. Although several researchers have examined the benefit of visual context upon the translation when textural information is degradated (Caglayan et al., 2019; Wang and Xiong, 2021), the improvements remain questionable. Wu et al. (2021) and Li et al. (2021) argue that vision contributes minor in previous MMT systems, and the images in the previous benchmark dataset provide limited additional information. In many cases, the translation of sentences relies on textual other than image information. The texts contain com-

[1] Codes and data are available at https://github.com/Yaoming95/23Triplet

Findings of the Association for Computational Linguistics: ACL 2023, pages 2679–2697
July 9-14, 2023 ©2023 Association for Computational Linguistics

plete contexts and are unambiguous, leaving the usage of images doubtful. Therefore, a benchmark in that the sentences can not be easily translated without visual information is much needed.

To address these limitations, we propose models to make the most of training data and build a challenge and real-world benchmark to push the real-world application of MMT research. At first, we propose a new framework, named `2/3-Triplet`, which can use both parallel text and image-text data. It provides two different ways of exploiting these data based on the continuous vision feature and discrete prompt tokens, respectively. The two approaches are not mutually exclusive and can be used jointly to improve performance within the same framework. It is also worth mentioning that the prompt approach is easy to deploy without modifying the model architecture.

In addition, we present a new real-world dataset named EMMT. We collect parallel text-image data from several publicly available e-commerce websites and label the translation by 20 language experts. To build a challenge test set, we carefully select ambiguous sentences that can not be easily translated without images. This high-quality dataset contains 22K triplets for training and 1000 test examples, along with extra image-text and parallel text data.

Comprehensive experiments show that `2/3-Triplet` rivals or surpasses text-only and other MMT competitors on EMMT, as well as previous benchmarks. Especially, `2/3-Triplet` consistently improves the strong text-only baseline by more than 3 BLEU scores in various settings, showing the importance of visual information.

2 Related Work

Researchers applied multimodal information to enhance machine translation systems since the statistical machine translation era (Hitschler et al., 2016; Afli et al., 2016). With the rise of neural networks in machine translation, researchers have focused on utilizing image information more effectively. Early work used image features as initialization for neural MT systems (Libovický and Helcl, 2017). More recent studies proposed multimodal attention mechanisms (Calixto et al., 2017; Yao and Wan, 2020), enhanced text-image representations using graph neural networks (Lin et al., 2020), latent variable models or capsule networks (Yin et al., 2020), and used object-level visual grounding information to align text and image (Wang and Xiong, 2021). Li et al. (2022a) found that a stronger vision model is more important than a complex architecture for multimodal translation.

As we discussed earlier, these methods are limited to bilingual captions with image data, which is scarce. Therefore, some researchers (Zhang et al., 2020; Fang and Feng, 2022) also design retrieval-based MMT methods that retrieve images with similar topics for image-free sentences. Alternatively, Elliott and Kádár (2017) proposed visual "imagination" by sharing visual and textual encoders.

Recently, Wu et al. (2021) and Li et al. (2021) have questioned whether the most common benchmark Multi30K (Elliott et al., 2016) is suited for multimodal translation since they found images contribute little to translation. Song et al. (2021) have contributed a new dataset of the e-commercial product domain. However, we find their datasets still have similar drawbacks.

Several relevant studies about translation and multimodality are noteworthy. Huang et al. (2020) used visual content as a pivot to improve unsupervised MT. Wang et al. (2022b) proposed a pre-training model by using modality embedding as prefix for weak supervision tasks. Li et al. (2022c) introduced the VALHALLA, which translates under guidance of hallucinated visual representation.

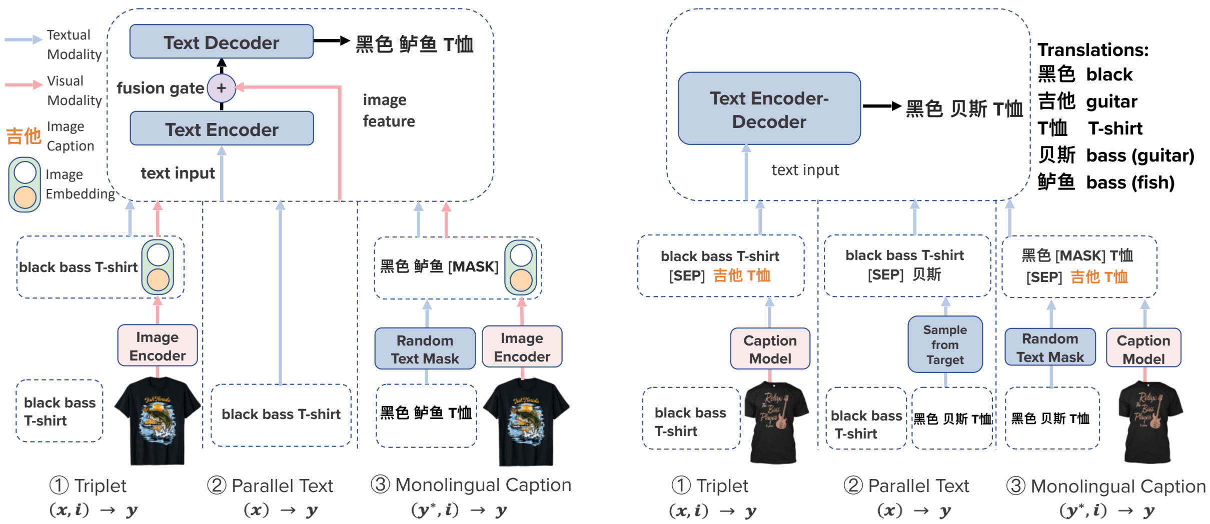
3 Approach

For the fully supervised condition in MMT, we have triplet $\{(x, y, i)\}$, where x is the source text, y is the target text, and i is the associated image. Since the triplet is rare, we attempt to utilize partially parallel data like $\{(y, i)\}$ and $\{(x, y)\}$, which are referred as monolingual image-text data and parallel text data in this paper.

In this section, we propose a new training framework `2/3-Triplet` with two approaches to utilize triple and non-triple data at the same time. We name these two approaches as **FUSION-BASED** and **PROMPT-BASED**, as shown in Figure 2.

For each approach, the model can conduct a mix training with three kinds of data: $((x, i) \to y)$, $((x) \to y)$, and $((y^*, i) \to y)$, where y^* indicates the masked target text.

FUSION-BASED approach resembles the conventional models where the encoded vision information is taken as model input and the model is trained in end2end manners, and our design makes it possible to utilize bilingual corpus and image-text pairs



(a) FUSION-BASED (b) PROMPT-BASED

Figure 2: The illustration of our framework $^2/_3$-Triplet in FUSION-BASED and PROMPT-BASED given ambiguous sentences. For each approach, the model can conduct a mixed training with three kinds of data: $\{(x, y, i)\}$, $\{(y, i)\}$, and $\{(x, y)\}$. ① means using triple data ($(x, i) \to y$); ② means using parallel text data ($(x) \to y$); ③ means using monolingual image-text data ($(y^*, i) \to y$), where y^* indicates the masked target text.

other than multilingual triplets.

PROMPT-BASED approach is inspired by the recent NLP research based on prompts (Gao et al., 2021; Li and Liang, 2021; Wang et al., 2022a; Sun et al., 2022), where we directly use the image caption as a prompt to enhance the translation model without any modification to the model.

3.1 FUSION-BASED

The common practice to utilize image information is to extract vision features and use them as inputs of the multimodal MT systems. Typically, it's common to cooperate vision and textual features to get a multimodal fused representation, where the textual features are the output state from the Transformer encoder and the vision feature is extracted via a pre-trained vision model.

We incorporate textual embedding and image features by simple concatenation:

$$H^{\text{fused}} = [H^{\text{text}}; \mathbf{h}^{\text{img}}] \tag{1}$$

where H^{text} is the encoded textual features of Transformer encoder, and $\mathbf{h}^{\text{img}}$ is the visual representation of [CLS] token broadcasted to the length of the text sequence.

Then, we employ a gate matrix Λ to regulate the blend of visual and textual information.

$$\Lambda = \tanh(f([H^{\text{text}}; H^{\text{fused}}])) \tag{2}$$

Finally, we add the gated fused information to the origin textual feature to get the final multimodal fused representation:

$$H^{\text{out}} = H^{\text{text}} + \mathbb{1}(\text{img})\Lambda H^{\text{fused}} \tag{3}$$

$\mathbb{1}(\text{img})$ indicates whether the image exists. The value is set to zero when image is absent.

It is worth noting that in Eq.2, we employ the hyperbolic tangent (tanh) gate instead of the traditional sigmoid gate (Wu et al., 2021; Li et al., 2022a) in the multimodal translation scenario. The new choice has two major advantages: (a) The output of the tanh can take on both positive and negative values, thereby enabling model to modulate the fused features H^{fused} in accordance with the text H^{text}; (b) The tanh function is centered at zero, thus, when the fused feature is close to zero, the output of the gate is also minimal, which aligns with the scenario where the image is absent naturally (*i.e.* $\tanh(0) = \mathbb{1}(\text{no img}) = 0$).

The next paragraphs illustrate how to utilize three types of data respectively.

Using Triple Data ($(x, i) \to y$) Figure 2a ①: Based on the basic architecture, we take in the source text for the text encoder and the image for the image encoder. By setting $\mathbb{1}(\text{img}) = 1$, we naturally leverage vision context for translation. The inference procedure also follows this flow.

Using Parallel Text ($(x) \to y$) Figure 2a ②: We utilize the same architecture as the triple data setting. By setting $\mathbb{1}(\text{img}) = 0$, we can adapt to the text-only condition. For the image-free bilingual data, the fused term is absent, and the final representation H^{out} is reduced to textual only, consistent with the learning on unimodal corpus.

Using Monolingual Caption ($(y^*, i) \to y$) Figure 2a ③: Inspired by Siddhant et al. (2020)'s strategy on leveraging monolingual data for translation, we adapt the mask de-noising task for utilizing monolingual image-text pairs. In a nutshell, we randomly mask some tokens in the text, and force the model to predict the complete caption text based on the masked text and image as input.

3.2 PROMPT-BASED

As prompt-based methods have made great success in NLP tasks(Gao et al., 2021; Li et al., 2022b; Wang et al., 2022a; Sun et al., 2022), we also consider whether the image information can be converted to some prompt signals for guiding sentence generation.

The general idea is quite straight: our translation system accepts a sentence of source language along with some keywords of target language, and translates the source sentence into the target language under the instruction of the target keywords. The keywords can be any description of the image that can help disambiguate the translation.

Using Triple Data ($(x, i) \to y$) Figure 2b ①: First, we generate the prompt from the image with a pre-trained caption model (we will introduce the caption model later). The source sentence is concatenated with the The original source sentence and the prompt are concatenated together to compose the training sources, with a special token `[SEP]` as a separator between the two.

Using Parallel Text ($(x) \to y$) Figure 2b ②: Since PROMPT-BASED approach adopts a standard Transformer and involves no modification on architecture, it is natural to train on unimodal parallel corpus. We use the parallel data to strengthen the ability to take advantage of the prompt. Without any image, we randomly select several words from the target sentence as the pseudo vision prompt. For translation training, we append the keyword prompt to the end of the original sentence and use a special token as a separator (Li et al., 2022b). After inference, we extract the translation result by splitting the separator token.

Using Monolingual Caption ($(y^*, i) \to y$) Figure 2b ③: Like FUSION-BASED approach, we use the de-noising auto-encoder task. By randomly masking some tokens and combining the caption result from the image as the prompt, we make the model learn to predict the original target text.

Training Caption Model ($(i) \to$ **keywords**(y)) Meanwhile, we train an caption model to generate the guiding prompt from images for translation, We formulate image-text pairs from both triple data and target-side monolingual caption. The input and output of the model are the image and extracted keywords of the corresponding target sentence.

3.3 Comparison and Combination of FUSION-BASED and PROMPT-BASED

Under the same training framework 2/3-`Triplet`, we propose two approaches, FUSION-BASED and PROMPT-BASED, for utilizing non-triple data. The FUSION-BASED approach preserves the complete visual context, providing more information via model fusion. In contrast, the PROMPT-BASED approach has the advantage of not requiring any modifications to the model architecture. Instead, all visual information is introduced by the prompt model, making deployment more straightforward.

The two methods, FUSION-BASED and PROMPT-BASED, are not mutually exclusive, and we can jointly utilize them. Specifically, the model simultaneously utilizes the fused feature in Eq. 3 as an encoder representation and the prompted-concatenated source as text input. The combination enables the model to benefit from our framework in the most comprehensive way, and as a result, the performance gains significant improvements.

4 Dataset

As mentioned before, in previous test sets, many sentences can be easily translated without the image context, for all information is conveyed in the text and has no ambiguity. To deeply evaluate visual information usage, we propose a multimodal-specific dataset.

We collect the data based on real-world e-commercial data crawled from TikTok Shop and Shoppee. We crawled the product image and title on two websites, where the title may be in English or Chinese. We filter out redundant, duplicate samples and those with serious syntax errors. Based on this, we conduct manual annotations. We hired a team of 20 professional translators. All translators are native Chinese, majoring in English. In addition, another translator independently samples the annotated corpus for quality control. We let the annotators select some samples specifically for the test set, which they found difficult to translate or had some confusion without images. The total number of triples annotated is 22, 500 of which are carefully selected samples as testset. We also randomly selected 500 samples as devsets among the full-set while the remaining as training set.

Besides the annotated triplets, we clean the rest of the crawled data and open sourced it as the monolingual caption part of the data. Since our approach features in utilizing bilingual data to enhance multimodal translation, we sample 750K CCAlign (El-Kishky et al., 2020) English-Chinese as a bilingual parallel text. The selection is motivated by the corpus's properties of its diversity in sources and domains, and it is more relevance to real-world compared to other corpus. The sampled data scale is decided based on both the model architecture and the principles of the neural scaling law (Kaplan et al., 2020; Gordon et al., 2021). We also encourage future researchers to explore the use of additional non-triple data to further enhance performance, as detailed in the appendix. We summarize the dataset statistics in Table 1. We discuss ethic and copyright issue of the data in the appendix.

Train			Test	Dev
Triplet	PT	MC		
22K	750K	103K	1000	500

Table 1: EMMT statistics. "PT" stands for parallel text data. "MC" stands for monolingual caption.

5 Experiments

5.1 Datasets

We conduct experiments on three benchmark datasets: Multi30K (Elliott et al., 2016), Fashion-MMT (Clean) (Song et al., 2021), and our EMMT. **Multi30K** is the most common benchmark on MMT tasks, annotated from *Flickr*, where we focus on English-German translation. To validate the effectiveness of parallel text, we add 1M English-German from CCAlign and COCO (Lin et al., 2014; Biswas et al., 2021). **Fashion-MMT** is built on fashion captions of FACAD (Yang et al., 2020).

5.2 Baselines

We compare our proposed 2/3-`Triplet` with the following SOTA MT and MMT systems:

Transformer (Vaswani et al., 2017) is the current *de facto* standard for text-based MT.

UPOC[2] (Song et al., 2021) introduced cross-modal pre-training tasks for multimodal translation.

Selective-Attention (SA) (Li et al., 2022a) investigated strong vision models and enhanced features can enhance multimodal translation with simple attention mechanism.

UVR-NMT (Zhang et al., 2020) retrieves related images from caption corpus as the pseudo image for sentences.

Phrase Retrieval (Fang and Feng, 2022) is an improved version of retrieval-based MMT model that retrieve images in phrase-level.

In addition, we report the results of Google Translate, which helps to check whether the translation of the test set actually requires images. All baselines reported use the same number of layers, hidden units and vocabulary as 2/3-`Triplet` for fair comparison.

We mainly refer to BLEU (Papineni et al., 2002) as the major metric since it is the most commonly used evaluation standard in various previous multimodal MT studies.

5.3 Setups

To compare with previous SOTAs, we use different model scales on Multi30K and the other two datasets. We follow Li et al. (2022a)'s and Li et al. (2021)'s setting on Multi30K, where the model has 4 encoder layers, 4 decoder layers, 4 attention heads, hidden size and filter size is 128 and 256, respectively. On the other two datasets, we set the model has 6 encoder layers, 6 decoder layers, 8 attention heads, hidden size and filter size is 256 and 512, respectively (*i.e.* Transformer-base setting). We apply BPE (Sennrich et al., 2016) on tokenized English and Chinese sentences jointly to get vocabularies with 11k merge operations. We use Zeng et al. (2022)'s method to get the caption model. The vocabularies, tokenized sentence and caption models will be released for reproduction. Codes are based on Fairseq (Ott et al., 2019).

When training models on various domains (+PT and +MC in Tab. 2), we upsample small-scale data (*i.e.* E-commercial Triplet) because of the massive

ID	Test set	EMMT			Multi30k-Test16		Multi30k-Test17	
	Training Data	Triplet Only	+PT	+ PT + MC	Triplet Only	+PT	Triplet Only	+PT
1	Plain Transformer♡	39.07	40.66	42.71	39.97	44.13	31.87	40.46
2	Selective Attention♠	41.27	/	/	40.63	/	33.80	/
3	UPOC^2◇	40.60	/	44.81	40.8	/	34.1	/
4	UVR-NMT♣	37.82	41.13	/	38.19	/	31.85	/
5	Phrase Retrieval♣	/	/	/	40.30	/	33.45	/
6	FUSION-BASED	41.74	44.22	45.93	40.95	/	34.03	/
7	PROMPT-BASED	41.70	43.35	46.28	40.17	/	33.87	/
8	FUSION+PROMPT	**42.03**	**45.20**	**46.55**	40.48	44.60	34.62	40.07
	Google Translate	44.27			41.9		42.0	

♡ We also train plain Transformer on monolingual captions via Siddhant et al. (2020)'s method for fair comparison on textual data.
♠ We use their open source code to reproduce Multi30K's results.
◇ Multi30K's results copy from Song et al. (2021). We add all MC and PT data for its pre-training in +PT+MC column for fair comparison on data. The complete UPOC^2 also utilize product attributes besides images, which is removed from our replication.
♣ Multi30K's results copy from Fang and Feng (2022). Phrase Retrieval is not reported on EMMT since they haven't released the phrase extraction scripts. We conduct the retrieval for all parallel sentences with top 5 images as candidate in +PT column of UVR-NMT.

Table 2: Results of 2/3-Triplet and related work on EMMT and Multi30k. "PT" indicates parallel text data $\{(x, y)\}$, "MC" indicates monolingual caption data $\{(y, i)\}$. On the one hand, 2/3-Triplet outperforms previous studies. On the other hand, extra non-triple data brings significant improvements. The reported improvement on **EMMT** dataset is examined with Dror et al. (2018)'s significance test with $p < 0.05$.

disparity of data scale in different domains. We follow Wang and Neubig (2019)'s and Arivazhagan et al. (2019)'s temperature based data sampling strategy and set the sampling temperature at 5. We empirically find that the model gains by simply randomly dropping some images during the training, where we set the drop ratio at 0.3 . Interestingly, such the method is also observed in other multimodal research topics (Abdelaziz et al., 2020; Alfasly et al., 2022). We evaluate the performance with tokenized BLEU (Papineni et al., 2002).

5.4 Main Results

We list the main results in Tab. 2. We get three major findings throughout the results:

1. In traditional multimodal MT settings (*i.e.* Triplet only and Multi30K), whose training and inference are on triple data, 2/3-Triplet rivals or even surpasses the previous SOTAs.
2. Parallel text and monolingual captions significantly boost the performance of multimodal translation models. With these additional data, even the plain Transformer model outperforms SOTA multimodal baselines. Given the scarcity of multimodal data, we argue that the use of extra data, especially the parallel text, is more crucial for multimodal translation than the use of multimodal information.
3. FUSION and PROMPT generally achieve the best performance when used together. This suggests two approaches are complementary.

We also list results on Multi30k for dataset comparison. Google Translate achieves the best results, while all other models are close in performance with **no** statistical significant improvement. It indicates that images in Multi30K are less essential and a strong text translation model is sufficient to handle the majority of cases. Moreover, we find that by incorporating non-imaged parallel text, the model's performance improves significantly, while narrows the gap between plain transformer models MMT ones. Hence, the parallel text rather than images may be more essential for improving performance on the Multi30k. In contrast, 2/3-Triplet surpass Google's on EMMT with visual infomation, providing evidence that ours serves as a suitable benchmark.

We also report the results of 2/3-Triplet and baselines on FashionMMT in Appendix along with BLEURT (Sellam et al., 2020) and word accuracy as supplementary metrics. The results show that 2/3-Triplet also rivals the SOTA MMT systems on various benchmarks and metrics.

5.5 Performance on Triplet-unavailable Setting

In more scenarios, annotated triple data is rather scarce or even unavailable, *i.e.* only bilingual translation or monolingual image caption is available in the training data, while we wish the model can still translate sentences in multimodal manners.

Since our proposed 2/3-Triplet utilize not only triplets, we examine whether our model can conduct inference on multimodal triple testset while only trained on the non-triple data, as triplet might be unavailable in real scenarios. In this experiment, we discard all images of EMMT's triples during the training stage, while the trained model is still evaluated on the multimodal test set. We compare the triplet-unavailable results to triplet only and full data training set settings in Figure 3

We can see that $^2/_3$-`Triplet` still preserves a relatively high performance and even sharply beats the triplet-only setting. This fully illustrates that involving parallel text and monolingual caption is extremely important for MMT.

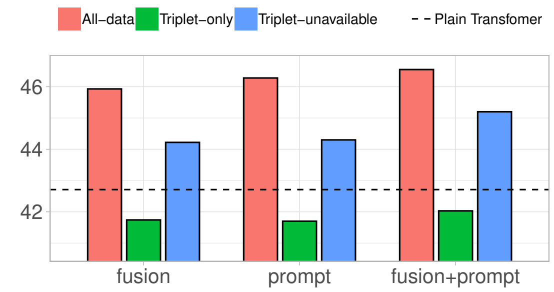


Figure 3: $^2/_3$-`Triplet`'s performance on using all-data, use triplet-only and triplet-unavailable cases. The horizontal line is the transformer baseline trained on all-data(*i.e.* +PT+MC of Row 1 in Table 2)

6 Discussion

As plenty of previous studies have discussed, the current multimodal MT benchmarks are biased, hence the quality gains of previous work might not actually derive from image information, but from a better training schema or regularization effect (Dodge et al., 2019; Hessel and Lee, 2020). This section gives a comprehensive analysis and sanity check on our proposed $^2/_3$-`Triplet` and EMMT: we carefully examine whether and how our model utilize images, and whether the testset of EMMT has sufficient reliability.

6.1 Visual Ablation Study: Images Matter

We first conduct ablation studies on images to determine how multimodal information contributes to the performance. Most studies used **adversarial** input (*e.g.* shuffled images) to inspect importance of visual information. However, effects of adversarial input might be opaque (Li et al., 2021). Hence, we also introduce **absent** input to examine whether $^2/_3$-`Triplet` can handle source sentences without image by simply zeroing the image feature for FUSION or striping the prompt for PROMPT.

We list the results of a vision ablation study of both adversarial and absent respectively in Figure 4,

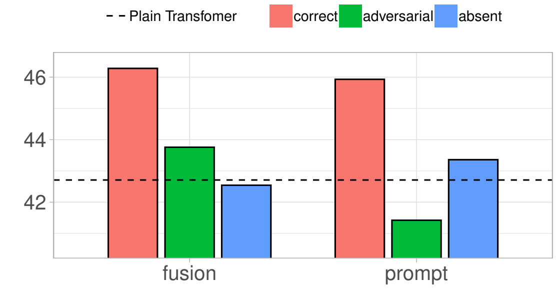


Figure 4: The results of ablation study by given empty image (absent) and wrong image (adversarial) as input.

where we select FUSION-BASED and PROMPT-BASED approaches trained with full data(last columns in Table 2) for comparison.

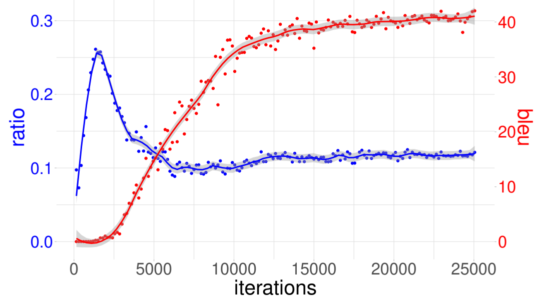


Figure 5: Ratio-BLEU on testset during the training

In the absent setting, both the FUSION and PROMPT degrade to the baseline, confirming the reliance of $^2/_3$-`Triplet` on image information. In the adversarial setting, the PROMPT performs worse than the baseline, which is in line with the expectation that incorrect visual contexts lead to poor results. However, while the FUSION also exhibits a decline in performance, it still surpasses the baseline. This aligns with the observations made by Elliott (2018); Wu et al. (2021) that the visual signal not only provides multimodal information, but also acts as a regularization term. We will further discuss this issue in Section 7.

6.2 How Visual Modality Works

We further investigate how the visual signal influence the model.

FUSION-BASED We verify how much influence the visual signal imposes upon the model. Inspired by Wu et al. (2021), we quantify the modality contribution via the L2-norm ratio (ΛH^{fused} for vision over H^{text} for text, in Eq. 3). We visualize the whole training process along with BLEU as a reference in Figure 5. Wu et al. (2021) criticize that previous studies do not utilize visual signal, for the

	source:	ready stock , cheese grains , pets only
	human:	现货 商品 奶酪 粒 (宠物 专用)
	Plain:	现货 起司 ~~谷物~~ 宠物 专用 ready stock cheese ~~cereal grains~~ pet only
	Ours (Triplet-only):	现货 奶酪 <u>颗粒</u> 宠物 ~~仅限~~ ready stock cheese <u>granular</u> pets ~~only for~~
	Ours (All-data):	现货 奶酪 粒 宠物 专用 ready stock cheese **grains** for pets only
	source:	ready stock kids medical surgical face mask 3-ply 20pcs
	human:	现货 儿童 医疗 手术 口罩 3 层 20 个
	Plain:	现货 儿童 医用 ~~面具~~ 3-~~ply~~ 20~~pcs~~ ready stock kids medical ~~(opera) mask~~ 3-~~ply~~ 20~~pcs~~
	Ours (Triplet-only):	现货 儿童 医用 口罩 3~~ply~~ 20~~pcs~~ ready stock kids medical **mask** 3-~~ply~~ 20~~pcs~~
	Ours (All-data):	现货 儿童 医用 外科 口罩 3 层 20 片 ready stock kids medical surgical **mask** 3-**ply** 20**pcs**

Table 3: Qualitative examples from two complex scenarios. Plain and Ours (Triplet-only) respectively indicate plain Transformer and 2/3-Triplet trained on triple data only, and Ours (All-data) indicate 2/3-Triplet trained on all data. The ~~strikethrough~~, <u>underline</u> and **bold** indicates inappropriate, reluctant and excellent choices respectively. The more detailed comments on the translation are in Appendix.

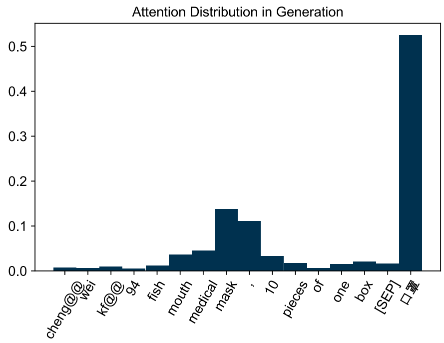


Figure 6: The attention distribution when predicting the token "口罩" ("mask" in English). The model shows a strong attention preference to the prompt words.

final ratio converge to zero. Our method shows a different characteristic: as the BLEU becomes stable, the ratio of visual signal and textual signal still remains at around 0.15, showing the effectiveness of the visual modality.

PROMPT-BASED We also look into the influence caused by the prompts. We sample an ambiguous sentence: "*chengwei kf94 fish mouth medical mask, 10 pieces of one box*". The keyword "mask" can be translated into "口罩" ("*face mask*" in English) or "面膜" ("*facial mask*" in English) without any context. We visualize the attention distribution when our PROMPT-BASED model is translating "mask" in Figure 6. We can see that the a high attention is allocated to the caption prompt. Therefore, our method correctly translates the word. We also visualize the detailed attention heatmaps for source, prompts and generated sentences in Appendix.

6.3 Qualitative Case Study

We also compare several cases from EMMT testsets to discuss how multimodal information and external bilingual data help the translation performance. Meanwhile, we regard the case study as a spot check for the multimodal translation testset itself. We here choose plain Transformer, our methods trained on triplet only and all data, as well as human reference for comparison.

Table 3 presents the qualitative cases and major conclusions are as follows: 1) Visual information plays a vital role in disambiguating polysemous words or vague descriptions. 2) Non-triple data improves translation accuracy, particularly in translating jargons and enhancing fluency in the general lack of multimodal data. 3) Our test set is representative in real-world seniors as it includes product titles that are confusing and require image, in contrast to previous case studies on Multi30k where researchers artificially mask key words (Caglayan et al., 2019; Wu et al., 2021; Wang and Xiong, 2021; Li et al., 2022a).

7 Conclusion

This paper devises a new framework 2/3-Triplet for multimodal machine translation and introduces two approaches to utilize image information. The new methods are effective and highly interpretable. Considering the fact that current multimodal benchmarks are limited and biased, we introduce a new dataset EMMT of the e-commercial domain. To better validate the multimodal translation systems, the testset is carefully selected as the image are crucial for translation accuracy. Experimental results and comprehensive analysis show that $^2/_3$-Triplet makes a strong baseline and EMMT can be a promising benchmark for further research.

Limitation

First, there are studies (Wu et al., 2021) claiming visual information only serves as regularization. In our ablation study, we find the adversarial setting of FUSION-BASED approach outperforms the plain Transformer. Combined with observations from previous studies, we suggest that fusion-based architectures may apply some images information as regularization terms, yet the further quantitative analysis is needed to confirm this phenomenon.

Second, though our testset is carefully selected to ensure the textual ambiguity without image data, we encounter difficulties in designing a suitable metric for quantifying the degree to which the models are able to resolve the ambiguity. Specifically, we find that conventional metrics, such as word-level entity translation accuracy, exhibit significant fluctuations and do not effectively quantify the extent to which the model effectively resolves ambiguity. We discuss this metric in more details in the Appendix, and offer a glossary of ambiguous words used in the test set. We acknowledge that the evaluation of multimodal ambiguity remains an open problem and an area for future research.

In addition, there are some details regarding the dataset that we need to clarify: the dataset is collected after COVID-19, so some commodities will be associated with the pandemic. We collect data by category in order to cover various products to reduce the impact of the epidemic on product types.

References

Ahmed Hussen Abdelaziz, Barry-John Theobald, Paul Dixon, Reinhard Knothe, Nicholas Apostoloff, and Sachin Kajareker. 2020. Modality dropout for improved performance-driven talking faces. In *ICMI '20: International Conference on Multimodal Interaction, Virtual Event, The Netherlands, October 25-29, 2020*, pages 378–386. ACM.

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

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 3874–3884. Association for Computational Linguistics.

Saghir Alfasly, Jian Lu, Chen Xu, and Yuru Zou. 2022. Learnable irrelevant modality dropout for multimodal action recognition on modality-specific annotated videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 20176–20185. IEEE.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George F. Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *CoRR*, abs/1907.05019.

Rajarshi Biswas, Michael Barz, Mareike Hartmann, and Daniel Sonntag. 2021. Improving german image captions using machine translation and transfer learning. In *Statistical Language and Speech Processing - 9th International Conference, SLSP 2021, Cardiff, UK, November 23-25, 2021, Proceedings*, volume 13062 of *Lecture Notes in Computer Science*, pages 3–14. Springer.

Ozan Caglayan, Menekse Kuyu, Mustafa Sercan Amac, Pranava Madhyastha, Erkut Erdem, Aykut Erdem, and Lucia Specia. 2021. Cross-lingual visual pre-training for multimodal machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 1317–1324. Association for Computational Linguistics.

Ozan Caglayan, Pranava Madhyastha, Lucia Specia, and Loïc Barrault. 2019. Probing the need for visual context in multimodal machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4159–4170. Association for Computational Linguistics.

Iacer Calixto, Qun Liu, and Nick Campbell. 2017. Doubly-attentive decoder for multi-modal neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1913–1924. Association for Computational Linguistics.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. 2019. Show your work: Improved reporting of experimental results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2185–2194. Association for Computational Linguistics.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 1383–1392. Association for Computational Linguistics.

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. Ccaligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 5960–5969. Association for Computational Linguistics.

Desmond Elliott. 2018. Adversarial evaluation of multimodal machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2974–2978. 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, hosted by the 54th Annual Meeting of the Association for Computational Linguistics, VL@ACL 2016, August 12, Berlin, Germany*. The Association for Computer Linguistics.

Desmond Elliott and Ákos Kádár. 2017. Imagination improves multimodal translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017 - Volume 1: Long Papers*, pages 130–141. Asian Federation of Natural Language Processing.

Qingkai Fang and Yang Feng. 2022. Neural machine translation with phrase-level universal visual representations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 5687–5698. Association for Computational Linguistics.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 3816–3830. Association for Computational Linguistics.

Mitchell A. Gordon, Kevin Duh, and Jared Kaplan. 2021. Data and parameter scaling laws for neural machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 5915–5922. Association for Computational Linguistics.

Jack Hessel and Lillian Lee. 2020. Does my multimodal model learn cross-modal interactions? it's harder to tell than you might think! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 861–877. Association for Computational Linguistics.

Julian Hitschler, Shigehiko Schamoni, and Stefan Riezler. 2016. Multimodal pivots for image caption translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Po-Yao Huang, Junjie Hu, Xiaojun Chang, and Alexander G. Hauptmann. 2020. Unsupervised multimodal neural machine translation with pseudo visual pivoting. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8226–8237. Association for Computational Linguistics.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *CoRR*, abs/2001.08361.

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), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 6327–6337. Association for Computational Linguistics.

Jiaoda Li, Duygu Ataman, and Rico Sennrich. 2021. Vision matters when it should: Sanity checking multimodal machine translation models. In *Proceedings* *of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 8556–8562. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 4582–4597. Association for Computational Linguistics.

Yafu Li, Yongjing Yin, Jing Li, and Yue Zhang. 2022b. Prompt-driven neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2579–2590. Association for Computational Linguistics.

Yi Li, Rameswar Panda, Yoon Kim, Chun-Fu Richard Chen, Rogério Feris, David D. Cox, and Nuno Vasconcelos. 2022c. VALHALLA: visual hallucination for machine translation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 5206–5216. IEEE.

Jindrich Libovický and Jindrich Helcl. 2017. Attention strategies for multi-source sequence-to-sequence learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers*, pages 196–202. Association for Computational Linguistics.

Huan Lin, Fandong Meng, Jinsong Su, Yongjing Yin, Zhengyuan Yang, Yubin Ge, Jie Zhou, and Jiebo Luo. 2020. Dynamic context-guided capsule network for multimodal machine translation. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, pages 1320–1329. ACM.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer.

Benjamin Marie, Atsushi Fujita, and Raphael Rubino. 2021. Scientific credibility of machine translation research: A meta-evaluation of 769 papers. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 7297–7306. 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: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Demonstrations*, pages 48–53. 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, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. BLEURT: learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7881–7892. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Aditya Siddhant, Ankur Bapna, Yuan Cao, Orhan Firat, Mia Xu Chen, Sneha Reddy Kudugunta, Naveen Arivazhagan, and Yonghui Wu. 2020. Leveraging monolingual data with self-supervision for multilingual neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 2827–2835. Association for Computational Linguistics.

Yuqing Song, Shizhe Chen, Qin Jin, Wei Luo, Jun Xie, and Fei Huang. 2021. Product-oriented machine translation with cross-modal cross-lingual pre-training. In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, pages 2843–2852. ACM.

Zewei Sun, Qingnan Jiang, Shujian Huang, Jun Cao, Shanbo Cheng, and Mingxuan Wang. 2022. Zero-shot domain adaptation for neural machine translation with retrieved phrase-level prompts. *CoRR*, abs/2209.11409.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Dexin Wang and Deyi Xiong. 2021. Efficient object-level visual context modeling for multimodal machine translation: Masking irrelevant objects helps grounding. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 2720–2728. AAAI Press.

Xinyi Wang and Graham Neubig. 2019. Target conditioned sampling: Optimizing data selection for multilingual neural machine translation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5823–5828. Association for Computational Linguistics.

Yifan Wang, Zewei Sun, Shanbo Cheng, Weiguo Zheng, and Mingxuan Wang. 2022a. Controlling styles in neural machine translation with activation prompt. *CoRR*, abs/2212.08909.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2022b. Simvlm: Simple visual language model pretraining with weak supervision. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Zhiyong Wu, Lingpeng Kong, Wei Bi, Xiang Li, and Ben Kao. 2021. Good for misconceived reasons: An empirical revisiting on the need for visual context in multimodal machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 6153–6166. Association for Computational Linguistics.

Xuewen Yang, Heming Zhang, Di Jin, Yingru Liu, Chi-Hao Wu, Jianchao Tan, Dongliang Xie, Jue Wang, and Xin Wang. 2020. Fashion captioning: Towards generating accurate descriptions with semantic rewards. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XIII*, volume 12358 of *Lecture Notes in Computer Science*, pages 1–17. Springer.

Shaowei Yao and Xiaojun Wan. 2020. Multimodal transformer for multimodal machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 4346–4350. Association for Computational Linguistics.

Yongjing Yin, Fandong Meng, Jinsong Su, Chulun Zhou, Zhengyuan Yang, Jie Zhou, and Jiebo Luo. 2020. A novel graph-based multi-modal fusion encoder for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 3025–3035. Association for Computational Linguistics.

Yan Zeng, Xinsong Zhang, and Hang Li. 2022. Multi-grained vision language pre-training: Aligning texts with visual concepts. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 25994–26009. PMLR.

Zhuosheng Zhang, Kehai Chen, Rui Wang, Masao Utiyama, Eiichiro Sumita, Zuchao Li, and Hai Zhao. 2020. Neural machine translation with universal visual representation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Appendix

7.1 Ethic Consideration About Data Annotators

We hire 20 professional translators in a crowd-source platform and pay them according to the market wage and work within 8 hours a day. All translators are native Chinese and have graduated with an English major. The ethics review is done while in data acceptance stage.

7.2 Data Copyright

In our study, we present a new dataset of public e-commercial products from Shoppee and TikTok Shop. To address copyright concerns, we provide a detailed description of how we collect the data and ensure that our usage complies with all relevant policies and guidelines.

For the Shoppee dataset, we obtain the data from their Open Platform API[2]. We carefully review their Data Protection Policy [3] and Privacy Policy guidelines [4], which provide clear instructions for using data through the Shopee Open Platform. We strictly follow their requirements and limitations, ensuring that we did not access any personal data and that we only use open information provided by the API. We also adhere to their robot guidelines [5], avoiding full-site scraping.

For the TikTok Shop dataset, we access the data using robots, as scraping is allowed according to their robots.txt file [6]. We also review TikTok Shop Privacy Policy and TikTok for Business Privacy Policy [7] to ensure that we only collect data from merchants under their policy.

It is important to note that all data we publish is publicly available on the Internet and only pertains to public e-commercial products. We do not access or publish any user information, and we take all necessary steps to respect the intellectual property and privacy rights of the original authors and corresponding websites. If any authors or publishers express a desire for their documents not to be included in our dataset, we will promptly remove that portion from the dataset. Additionally, we certify that our use of any part of the datasets is limited to non-infringing or fair use under copyright law. Finally, we affirm that we will never violate anyone's rights of privacy, act in any way that might give rise to civil or criminal liability, collect or store personal data about any author, infringe any copyright, trademark, patent, or other proprietary rights of any person.

[2]https://open.shopee.com/documents
[3]https://open.shopee.com/developer-guide/32
[4]https://careers.shopee.tw/privacy-policy
[5]https://shopee.tw/robots.txt
[6]https://shop.tiktok.com/robots.txt
[7]https://tiktokfor.business/privacy-policy/

7.3 Results on Fashion-MMT

We list the testset performance on Fashion-MMT in Table 4.

	FashionMMT（C）	Triplet Only	+ Parallel Text
	Transformer	40.12	/
UPOC^2	MTLM+ISM	41.38	/
	MTLM+ISM+ATTP	41.93	/
Ours	FUSION	41.19	42.38
	PROMPT	40.97	42.02
	FUSION+PROMPT	41.38	42.33

Table 4: Results on Fashion-MMT(C) testset.

Fashion-MMT is divided into two subset according to the source of the Chinese translation: "Large" subset for the machine-translated part and "Clean" subset for the manually annotated part. As its authors also found the Large subset is noisier and different from the human annotated data, our experiments focused on the Clean subset with Fashion-MMT(*i.e.* Fashion-MMT(c)).

We compare the model performance on training on Triplet Only and adding Parallel Text settings. As the original dataset does not provide a parallel corpus without pictures, we used Parallel Text from EMMT for our experiments.

Note that the UPOC^2 model relies on three sub-methods, namely MTLM, ISM, and ATTP. The ATTP requires the use of commodity attributes, whereas our model does not use such information. Hence, we also list results of UPOC^2 without ATTP in the table.

The results show that our model rivals UPOC^2 on triplet only settings. And by using parallel text, ours gain further improvement, even if the parallel text does not match the domain of the original data. The results demonstrate the potential of our training strategy over multiple domains.

7.4 Evaluation with various metrics

Recent studies have indicated that the sole reliance on BLEU as an evaluation metric may be biased (Marie et al., 2021). We hence evaluate models with machine learning-based metric BLEURT (Sellam et al., 2020) and list the results

ID	Metric	BLEURT			Accuracy		
	Training Data	Triplet Only	+ PT	+ PT + MC	Triplet Only	+ PT	+ PT + MC
1	Plain Transformer	0.5424	0.5559	0.5662	0.765	0.754	0.761
2	Selective Attention	0.5619	/	/	0.782	/	/
3	UPOC^2	0.4855	/	0.5788	0.792	/	0.798
4	UVR-NMT	0.5299	0.5866	/	0.795	0.791	/
5	Phrase Retrieval	/	/	/	/	/	/
6	FUSION-BASED	**0.5760**	0.5782	0.5923	0.771	0.778	0.812
7	PROMPT-BASED	0.5600	0.5772	0.5980	0.792	0.775	0.792
8	FUSION+PROMPT	0.5647	**0.5917**	**0.6018**	0.809	**0.791**	0.799
	Google Translate		**0.6108**			0.741	

Table 5: Results of 2/3-`Triplet` and baselines on EMMT evaluated by BLEURT and word-level accuarcy of ambiguous words.

in Table 5[8].

Previous multimodal works often set entity nouns in the original sentence into [mask] to quantify model's ability for translating masked items with images (Wang and Xiong, 2021; Li et al., 2022a; Fang and Feng, 2022). While the experiment can measure the effectiveness of multimodal information, text with [mask] is not natural and the setting makes less sense in the real world. Inspired by their settings, we have developed a set of commonly used English-Chinese translation ambiguities by mining frequently used product entity and manual annotating. We have defined an word-level accuracy metric based on those potential ambiguous words in Table 7: if a certain English word appears in the original sentence, we require that the model's translation result in the target language must be consistent with the human reference's corresponding entity translation in order to be considered a correct translation, and thus calculate the word-level accuracy.

The results of BLEURT generally align with BLEU, indicating the effectiveness of 2/3-`Triplet`. However, an exception occurs in the Google Translate system, whose score are highest among all systems. We attribute this deviation to the use of back-translated pseudo corpus in the pre-training of the BLEURT model.

Multimodal models consistently perform better than plain transformer models in word-level accuracy. Additionally, Google Translate obtains the lowest scores in word-level accuracy, indicating that BLEURT may not distinguish ambiguous words in multimodal scenarios. However, the difference between multimodal ones is not significant. We attribute it to the difficulty in quantifying the semantic differences between synonyms, as we will demonstrate in our case study details. Furthermore, given the significant human effort required for mining and annotating ambiguous word list while it is highly domain-specifc to the test set, we suggest that the development of new metrics for evaluating multimodal translation ambiguity shall be a valuable topic of future research.

[8]We use BLEURT-20 model from `https://github.com/google-research/bleurt`

7.5 Translation Details of Case Study

Here we give some detailed explanations about the translation of case study translations:

In the first case, the Plain Transformer fail to recognize whether the word "grains" means cereal crop (谷物) or the cheese of grain sizes(奶酪粒). Triplet-Only 2/3-`Triplet` translate "grains" into 颗粒, which is acceptable, but the word not commonly used to describe food in Chinese, yet the model does not translate "only" grammatically properly.

In the second case, Plain Transformer translates "mask" to 面具, which is more commonly used to refer opera mask in Chinese. Both Plain Transformer and Triplet-Only 2/3-`Triplet` fail to understand "pcs"(件、个、片) and "ply"(层), and directly copy them to targets. The two methods also fail to translate "surgical"(手术、外科) correctly as it is a rare word in Triplet only settings.

In comparison, the translation of 2/3-`Triplet` is more consistent with the images, and more appropriate in terms of grammar and wording.

Metric	BLEU		BLEURT	
Data	+ PT	+ Excessive PT	+ PT	+ Excessive PT
Plain Transformer	40.66	40.86	0.5559	0.5406
FUSION-BASED	44.22	43.91	0.5782	0.5877
PROMPT-BASED	43.35	43.61	0.5772	0.5760
FUSION+PROMPT	45.20	44.87	0.5917	0.6025

Table 6: Results of 2/3-Triplet and plain Transformer on EMMT with parallel text and excessive parallel text (5M).

7.6 Attention Visualization

We visualize one the attention heatmap case of PROMPT-BASED in Figure 8 and Figure 8.

Figure 8 shows the attention alignment of original source (y-axis) and the prompted source (x-axis) in text encoder. Figure 8 shows the generated sentence (y-axis) and the prompted source (x-axis) in text decoder. From the heat map we know that the prompt attends to the most relevant ambiguous words and supports the model translation, both when encoding the source sentence and decoding the infernece. Specifically in our case, "口罩"(face mask) in prompts has high attention with all "masks" occurrence on the source side, and has high attention with all "口罩" generation in decoder side. In contrast, the word "防护"(protective) less prominent in the attention heatmap as it is less ambiguous.

7.7 Details on Data Selection and Mixing

As discussed in Section 5.3, we resort to upsampling the e-commercial triplet data due to the significant disparity in the quantity of data across various domains. As previously proposed by Wang and Neubig (2019) and Arivazhagan et al. (2019), we utilize a temperature-based sampling method, where the i-th data split is assigned a sampling weight proportional to $D_i^{\frac{1}{T}}$, where D_i denotes the number of sentences in the i-th data split, and T is the temperature hyper-parameter. In our implementation, to guarantee the completeness and homogeneity of data across each training iteration, we directly upsample the triplet data or monolingual captions, and subsequently, shuffle them randomly with parallel text to construct the training dataset. The upsampling rate for the triplet data is rounded to 15 and the upsampling rate for the parallel text is rounded to 4, resulting in an actual sampling temperature of 5.11 .

8 Model Performance with Excessive Data

Based on data distribution and scaling laws, we sample 750k parallel text and 103k monolingual captions as non-triple data to validate our methods. To further explore the potential of models with excessive non-triple data, we attempt to increase the data scale of the parallel text corpus to 5M, which are also sampled from CCAlign corpus. We list the results in Table 6. However, we find that excessive parallel text does not further promote model performance on current test sets. We suggest that the lack of improvement in performance may be due to the difference in text domain between the general domain and the e-commerce domain. As we will release the parallel text corpus we used in our experiments, in addition to conducting fair comparisons based on our data, we also encourage future researchers to use more unconstrained external data and techniques to continue to improve performance.

English Word	Chinese Potential Translations	English Word	Chinese Potential Translations
mask	面膜,口罩,面罩,面具,遮垫	tape	胶带,胶布,带子,磁带,薄胶带
bow	琴弓,弓子,弯弓	bar	吧台,酒吧,棒杆
top	上衣,上装,女上装,机顶	basin	盆子,盆器,盆,地盆,盆池
set	套装,把套,撮子,套盒,组套	sheet	被单,棚布,薄板,薄片,片材
clip	卡子,提盘夹,提盘夹子,夹片,取夹	film	贴膜,薄膜,胶片,胶卷,软片
nail	钉子,铁钉,扒钉,指甲,钉钉子	eyeliner	眼线笔,眼线液,眼线,眼线膏
iron	铁,铁艺,电熨斗,熨斗,烫斗	shell	车壳,被壳,贝壳,外壳,壳壳
rubber	胶皮,橡皮	chip	芯片,筹码
brush	刷子,毛笔,毛刷,板刷,锅刷	plug	插头,塞子,胶塞,堵头,地塞
oil	机油,油,油脂,油液	napkin	餐巾纸,餐巾
canvas	餐布,油画布,画布,帆布	grease	润滑脂,打油器
ring	戒指,指环,圆环,圈环,响铃	pipe	管子,烟斗,管材,皮管,排管
pad	护垫,盘垫,踏垫,垫块,贴垫	charcoal	木炭,炭笔,炭,引火炭,炉炭
wipes	湿巾,抹手布,擦地湿巾,擦碗巾,擦奶巾	blade	铲刀,刀片,叶片,刀锋,遮板
face mask	焕颜面膜,护脸面罩,遮脸面罩,脸罩,脸部面膜,口罩	bucket	水桶,面桶,扒斗,漂桶,簸箩
powder	粉饼,散粉,粉掌,修容粉饼,粉剂	lift	升降机,升降梯,举升机,举升器,起重器
tie	扎带,领带	crane	吊车,起重机,吊机,起重吊机,仙鹤
desktop	桌面,台式机	football	足球,橄榄球
jack	千斤顶,插孔	frame	画框,车架,框架,包架,裱画框
collar	项圈,颈圈,套环,领夹	plum	话梅,李子
cement	胶泥,水泥	slide	滑轨,滑梯,滑滑梯,滑道,幻灯片
tank	坦克,料槽,坦克车	keyboard	键盘,钥匙板,小键盘
hood	头罩,遮光罩,机罩,风帽,引擎盖	bass	鲈鱼,贝斯
gum	牙胶,树胶,口香糖	makeup remover	卸妆水,卸妆膏,卸妆液,卸妆乳,卸妆棉棒
screen	屏风,纱窗,滤网,丝网,筛网	counter	计数器,柜台
bell	铃铛,车铃,吊钟,吊铃	separator	隔板,分离器,分液器,隔片,分离机

Table 7: Potential ambiguous product entities in English-Chinese translations. The Chinese translations are separated by commas and have different meanings in the alternative translations. Words are sorted by the frequency in e-commercial English corpus

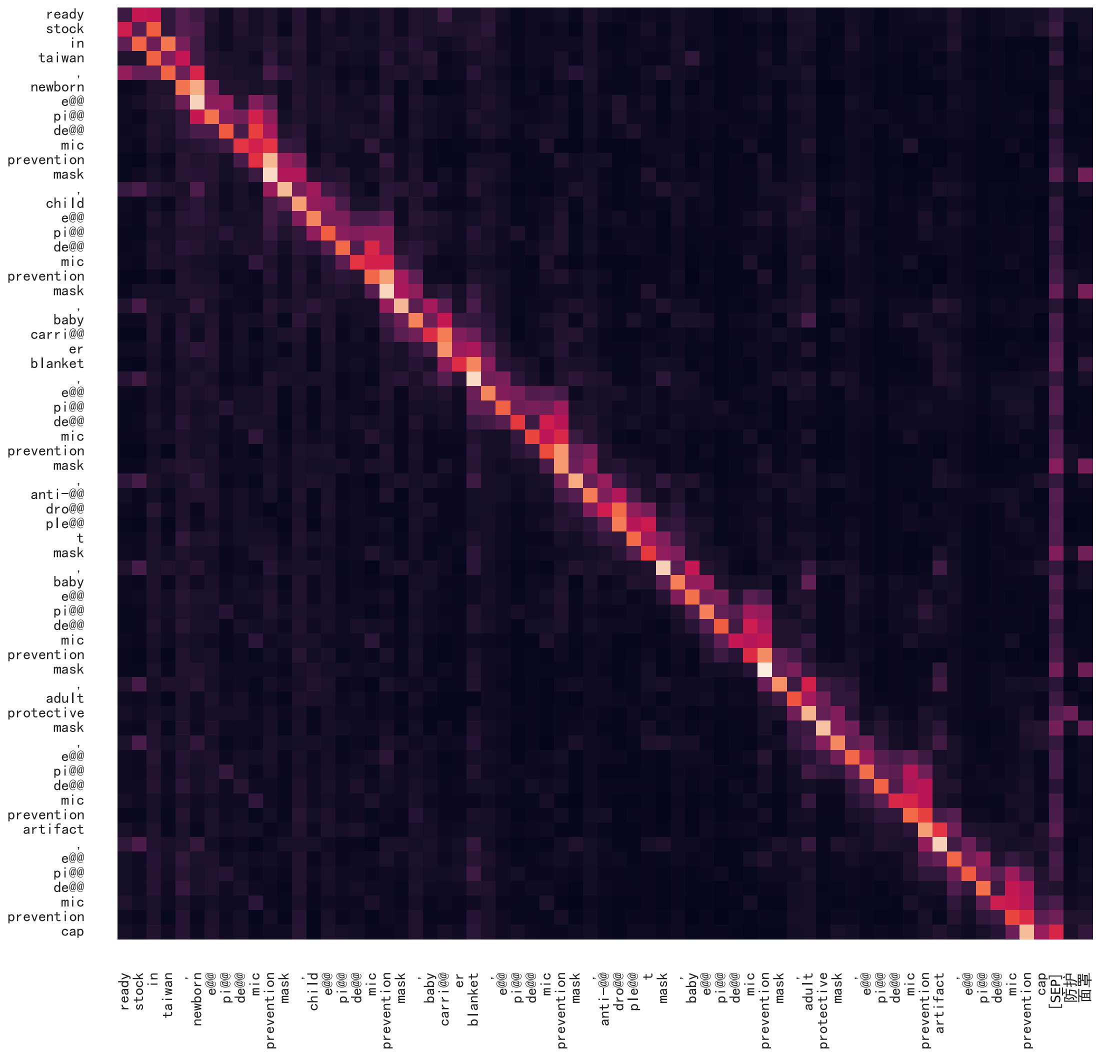


Figure 7: Attention heat map between source sentence and the source with caption prompts

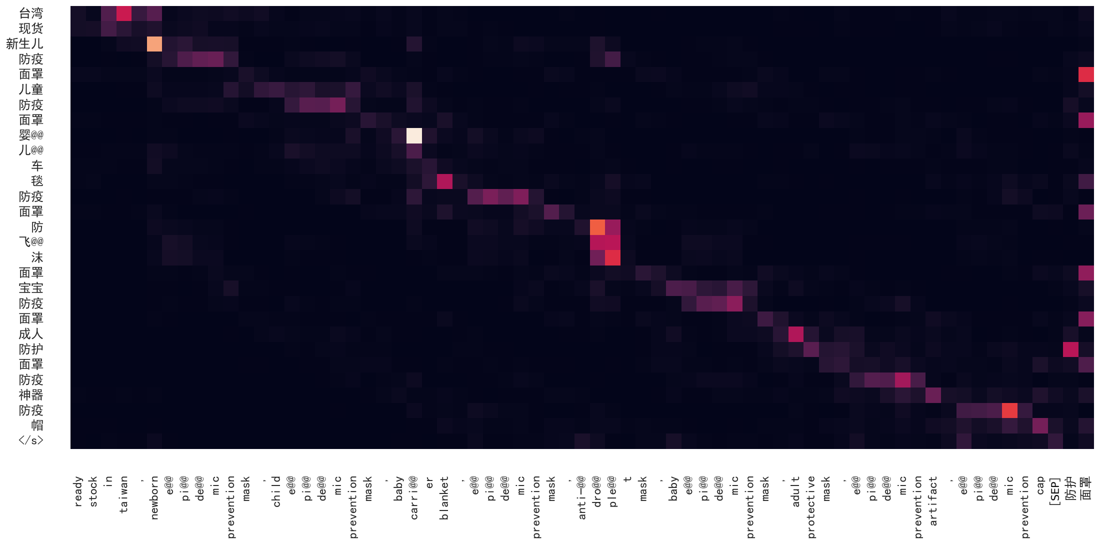


Figure 8: Attention heat map between hypo sentence and the source with caption prompts

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
In Limitation section

☒ A2. Did you discuss any potential risks of your work?
The data set is only on e-commercial domain, which has bit potential risks

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☒ Did you use or create scientific artifacts?

Left blank.

☒ B1. Did you cite the creators of artifacts you used?
Left blank.

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Left blank.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
No response.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
No response.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No response.

☐ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
No response.

C ☑ Did you run computational experiments?

5

☒ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
We illustrate the model parameters. The model is small (6-layer Transformer-base) and is friendly to researchers with low resources

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
5

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
5.4. In table

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
5.4

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
4 and in Appendix

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
in Appendix

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
in Appendix

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
in Appendix

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
in Appendix

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
in Appendix