Exploring the Necessity of Visual Modality
in Multimodal Machine Translation using Authentic Datasets

Zi Long[1], Zhenhao Tang[2], Xianghua Fu[1], Jian Chen[2], Shilong Hou[2], Jinze Lyu[2]

[1]College of Big Data and Internet, Shenzhen Technology University, Shenzhen, China
[2]College of Application and Technology, Shenzhen University, Shenzhen, China

Abstract

Recent research in the field of multimodal machine translation (MMT) has indicated that the visual modality is either dispensable or offers only marginal advantages. However, most of these conclusions are drawn from the analysis of experimental results based on a limited set of bilingual sentence-image pairs, such as Multi30k. In these kinds of datasets, the content of one bilingual parallel sentence pair must be well represented by a manually annotated image, which is different from the real-world translation scenario. In this work, we adhere to the universal multimodal machine translation framework proposed by Tang et al. (2022). This approach allows us to delve into the impact of the visual modality on translation efficacy by leveraging real-world translation datasets. Through a comprehensive exploration via probing tasks, we find that the visual modality proves advantageous for the majority of authentic translation datasets. Notably, the translation performance primarily hinges on the alignment and coherence between textual and visual contents. Furthermore, our results suggest that visual information serves a supplementary role in multimodal translation and can be substituted.

Keywords: MMT, image retrieval, visual noise filtering, supplementary text retrieval

1. Introduction

With the development of neural machine translation (NMT), the role of visual information in machine translation has attracted researchers' attention (Specia et al., 2016; Elliott et al., 2017; Barrault et al., 2018). Different from those text-only NMT (Bahdanau et al., 2014a; Gehring et al., 2016), a bilingual parallel corpora with manual image annotations are used to train an MMT model by an end-to-end framework, and therefore visual information can assist NMT model to achieve better translation performance (Calixto and Liu, 2017; Calixto et al., 2017; Su et al., 2021).

Concurrently, researchers have also undertaken a diverse range of experiments in an effort to validate the specific role of visual information in NMT. For example, Grönroos et al. (2018a) and Lala et al. (2018) observed that the robustness of MMT systems remains unaffected when the input image lacks direct relevance to the accompanying text. Notably, the absence of visual features, as highlighted by Elliott (2018), also does not yield detrimental effects. Wu et al. (2021) underscores that the utilization of the visual modality serves as a regularization mechanism during training rather than serving as a true complement to the textual modality. Oppositely, Caglayan et al. (2019) delve into the correlation between visual features and text. Their investigation reveals that incorporating the input image aids translation, particularly when certain input words are masked. Li et al. (2022) design more detailed probing tasks and found that stronger vision features strengthen MMT systems.

Note that most of the previous conclusions are drawn from the analysis of experimental results based on a restricted selection of manually annotated bilingual sentence-image pairs, known as the Multi30k dataset (Elliott et al., 2016). Within the Multi30k dataset, as depicted in Table 1, the sentences primarily comprise common and straightforward vocabulary, with each bilingual parallel sentence pair being effectively depicted by a single image. Table 1 also presents an illustration of a bilingual sentence-image pair extracted from a genuine news report from the United Nations News[1], alongside examples of sentence pairs derived from other other authentic translation datasets. Evidently, a substantial disparity exists between the Multi30k dataset and the authentic translation data. Hence, the evidence and findings derived from Multi30k may potentially exhibit inadequate generalizability and offer limited utility when attempting to analyze the role of the visual modality in MMT within real-world translation scenarios. In these scenarios, sentences often incorporate rare and uncommon words and are only partially depicted by accompanying images.

In a recent study, Tang et al. (2022) introduced a universal multimodal neural machine translation model that integrates open-vocabulary image retrieval techniques. In this work, inspired by Tang et al. (2022), we formulate a set of comprehensive probing tasks aimed at assessing the extent to which the visual modality enhances MMT within real-world translation scenarios. In addition to com-

---

[1]https://news.un.org/en/

Data source	Sentences	Image
Multi30k	EN: A dog is running in the snow. DE: Ein Hund rennt im Schnee.	
UN News	EN: Rescue workers look for survivors in a building in Samada, Syria destroyed by the February 6 earthquake. DE: Rettungskräfte suchen nach Überlebenden in einem Gebäude in Samada, Syrien, das durch das Erdbeben vom 6. Februar zerstört wurde.	
Bible	EN: I saw, and behold, there was no man,and all the birds of the sky had fled. DE: Ich sah, und siehe, da war kein Mensch, und alle Vögel unter dem Himmel waren weggeflogen.	no image
MultiUN	EN: Development assistance cannot by itself prevent or end conflict. DE: Entwicklungshilfe allein kann Konflikte weder verhüten noch beenden.	no image

Table 1: Comparison between Multi30k Dataset and Authentic Datasets

monly used Multi30k, we conduct an extensive set of experiments across four authentic text-only translation datasets. We further evaluated two visual noise filtering approaches based on the correlation between textual and visual content. Furthermore, we investigate the necessity of visual modality in the current multimodal translation process by substituting visual data with closely equivalent textual content. To summarize, our findings are:

(1) Visual modality is mostly beneficial for translation, but its effectiveness wanes as text vocabulary becomes less image-friendly.

(2) The MMT performance depends on the consistency between textual and visual contents, and utilizing filters based on the textual-visual correlation can enhance the performance.

(3) Visual information plays a supplementary role in the multimodal translation process and can be substituted by the incorporation of additional textual information.

2. Related Work

The integration of extra knowledge to build fine-grained representations is a crucial aspect in language modeling (Li et al., 2020a,b; Zhang et al., 2020). Incorporating the visual modality into language modeling has the potential to enhance the machine's understanding of the real world from a more comprehensive perspective. Inspired by the studies on the image description generation task (Elliott et al., 2015; Venugopalan et al., 2015; Xu et al., 2015), MMT models have gradually become a hot topic in machine translation research. In some cases, visual features are directly used as supplementary information to the text presentation. For example, Huang et al. (2016) take global visual features and local visual features as additional information for sentences. Calixto and Liu (2017) initializes the encoder hidden states or decoder hidden states through global visual features. Calixto et al. (2017) use an independent attention mechanism to capture visual representations. Caglayan et al. (2016) incorporate spatial visual features into the MMT model via an independent attention mechanism. On this basis, Delbrouck and Dupont (2017b) employs compact bilinear pooling to fuse two modalities. Lin et al. (2020) attempt to introduce the capsule network into MMT, they use the timestep-specific source-side context vector to guide the routing procedure. Su et al. (2021) introduce image-text mutual interactions to refine their semantic representations.

Researchers have also come to recognize the potential redundancy of the visual modality. Inconsequential images exhibit minimal impact on translation quality, and the absence of images does not yield a significant drop in BLEU scores, as noted by Elliott (2018). Encouraging findings emerged in the study by Caglayan et al. (2019). They high-
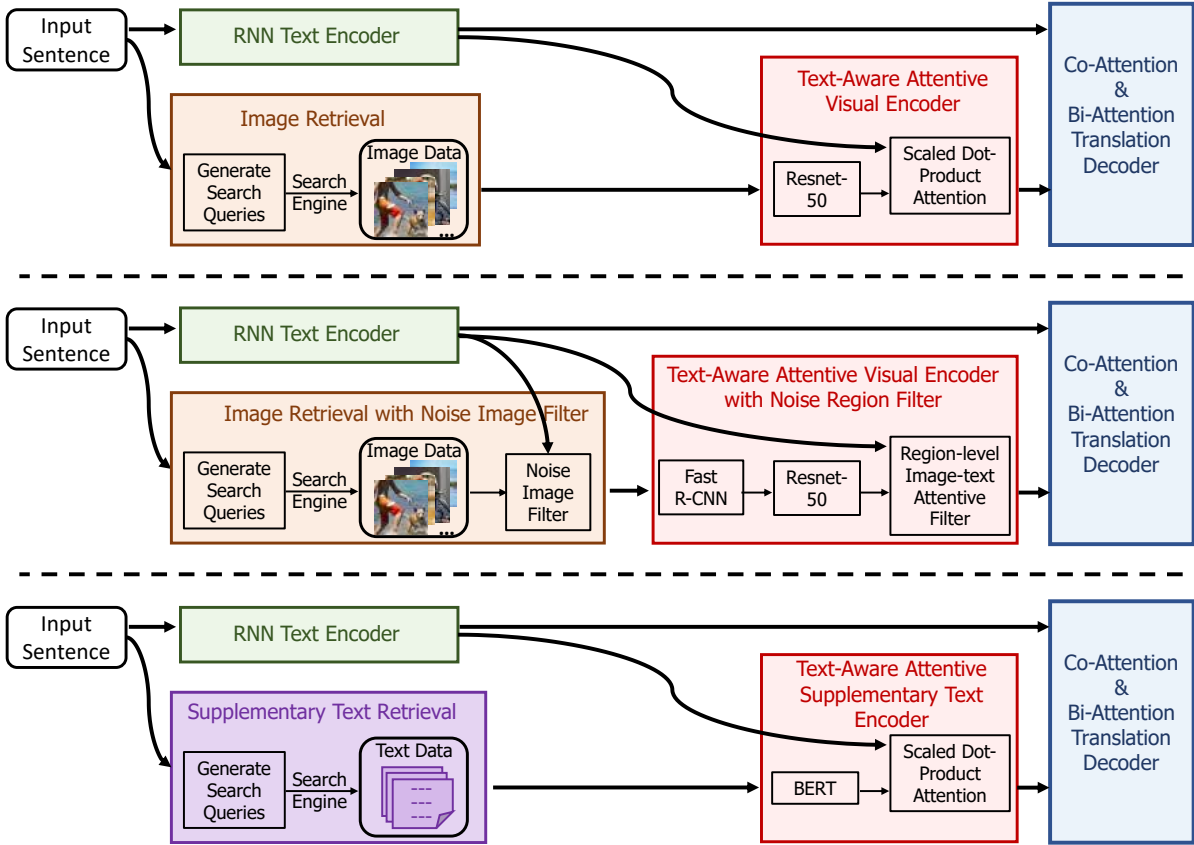
Figure 1: Frameworks of three probing methods

lighted the continuing utility of the visual modality in scenarios where linguistic context is limited but noted its diminished sensitivity when exposed to complete sentences. In a more recent investigation, Wu et al. (2021) attributed the observed BLEU improvement in MMT tasks to training regularization. They underscored the importance of constructing appropriate probing tasks with inadequate textual input. It's important to highlight that the proposed probing task represents an enhanced iteration building upon prior research (Caglayan et al., 2019; Wu et al., 2021). Li et al. (2022) made a systematic study on whether stronger vision features are helpful. All the preceding research has been conducted exclusively on the Multi30k dataset, which has limitations in scale and considerably differs from real-world translation scenarios. In this study, we employ the framework introduced by Tang et al. (2022) to systematically examine the influence of visual information across various authentic translation datasets, extending our analysis beyond the limitations of the small and specialized Multi30k dataset.

3. Preliminary

We start with a description of three probing methods employed in this work, which encompass the approach introduced by Tang et al. (2022) and two additional methods derived from it. Figure 1 shows frameworks of these three methods.

3.1. MMT with Search Engine Based Image Retrieval

As depicted in the top section of Figure 1, Tang et al. (2022) introduced a search engine-based image retrieval technique and a text-aware attention image encoder. This innovation enables the handling of authentic text-only translation data within MMT systems. We implement this approach across multiple authentic translation datasets to examine the influence of visual information across datasets with varying styles. To ensure the comprehensiveness of this paper, this section will provide a brief overview of the approach proposed by Tang et al. (2022).

Text Encoder In this work, we employ a commonly utilized bi-directional LSTM as the RNN text encoder. For a given sentence denoted as X,

the output of the text encoder is represented as $C = (\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_N)$, where N denotes the length of the sentence X.

Image Retrieval To emphasize the core components of the sentence and mitigate the impact of noise, including stopwords and infrequent words, Tang et al. (2022) utilized the TF-IDF method (Witten et al., 2005) to generate search queries for image search engines. Subsequently, the generated search queries are utilized in image search engines to retrieve the first available image associated with each query. For each given sentence X, M search queries denoted as $(q_1, q_2, \ldots, q_M)$ are generated, and subsequently M images represented as $(A_1, A_2, \ldots, A_M)$ are retrieved from search engines.

Text-Aware Attentive Visual Encoder Each image A_m ($m = 1, \ldots, M$) is transformed into a 196×1024 dimensional feature vector using ResNet-50 (He et al., 2016) . A simple but effective scaled dot-product attention in visual encoder is subsequently employed in the visual encoder to derive a resultant visual representation. Here, we utilize the average pooling C' of the text representation $C = (\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_N)$ as the query, while the visual feature vectors $A_1, A_2, \ldots, A_M$ serve as the keys and values in this attention mechanism. The resultant visual representation A is also expressed as a 196×1024 dimensional feature vector, which can be regarded as a matrix $A = (\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_L)$, where $L = 196$ and each $\boldsymbol{a}_l \in R^{1024}$ ($l = 1, \ldots, L$). Visual representation $A = (\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_L)$ and text representation $C = (\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_N)$ are then used as the inputs of translation decoder.

Translation Decoder For the decoder, we adopt the approach introduced by Su et al. (2021), implementing both a bidirectional attention network and a co-attention network to effectively capture the underlying semantic interactions between textual and visual elements. Based on the results of the preliminary experiment, it was evident that transformer-based models did not confer a performance advantage on datasets like Global Voices and other smaller ones. Consequently, we followed the approach of Tang et al. (2022) and selected LSTM as our foundational model. The bidirectional attention network enhances the representations of both text and image. These enhanced representations are subsequently input into the co-attention network to obtain the time-dependent context vector c_t and the visual vector v_t. Within the co-attention network, we calculate the probability distribution for the next target word y_t using the previous hidden state s_{t-1}, the previously generated target word y_{t-1}, the time-dependent context vector c_t, and the time-dependent visual vector v_t.

3.2. MMT with Visual Noise Filtering

Considering that the noise images obtained from search engines could have a substantial impact on the performance of the MMT system, we further evaluated two visual noise filtering approaches based on the correlation between textual and visual content, as depicted in the central part of Figure 1. One approach utilizes the pretrained CLIP model to filter out noise images, while the other employs a region-level image-text attentive filter module to filter out noisy image regions.

Noise Image Filter In the CLIP-based noise image filtering approach, we begin by retrieving M' ($M' > M$) images from search engines for each input sentence. Following this, we calculate the correlation between the input text and the retrieved images using a pretrained CLIP model (Radford et al., 2021). Subsequently, we select only the top-M images with the highest correlation to the input source text as the output of the image retrieval process.

Noise Region Filter In the noise image region filtering approach, we begin by extracting convolutional feature maps from the top-O most confident regions denoted as $(r_1, \ldots, r_O)$ in each collected image. This is achieved using a pretrained Faster R-CNN model (Ren et al., 2015), aiding in the initial filtration of visual information that may be challenging to distinguish as distinct regions in the images. The image region of each collected image is then represented as a 1024 dimensional feature vector using ResNet-50. For all the retrieved M images, we extract a total of $M \times O$ regions $(r_1, \ldots, r_{M \times O})$, resulting in $M \times O$ feature vectors $(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_{M \times O}, \boldsymbol{a}_o \in R^{1024})$. Subsequently, we compute the correlation score between each image region and the input text using the following equation:

$$S(\boldsymbol{a}_o, C') \quad = \quad V_a \tanh(W_a \boldsymbol{a}_o + U_a C')$$

Here, C' represents the average pooling of the text representation $C = (\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_N)$. We retain only the visual information from the top-O most relevant regions out of the initially extracted $M \times O$ regions. This preserved visual information serves as the visual representation for the given input sentence, denoted as $A = \{\boldsymbol{a}_o | S(\boldsymbol{a}_o, C')$ ranks in the top-$O, 1 \leq o \leq M \times O\}$, and it is subsequently fed into the translation decoder module. Less relevant regions are discarded during this process.

3.3. Supplementary Text Enhanced NMT

As discussed by Caglayan et al. (2019), multimodal translation models typically view visual information as a complementary component to textual information. However, we raise the question of whether this complementary role can also be achieved by incorporating additional textual information, potentially obviating the need for images in the process. Hence, our investigation aims to assess the necessity of visual information in the existing multimodal translation process by substituting visual data with nearly equivalent textual information. As illustrated in the lower section of Figure 1, we replace the image retrieval module with a supplementary text retrieval module and substitute the text-aware attentive visual encoder with a similar text-aware attentive supplementary text encoder.

Supplementary Text Retrieval Similar to the process of retrieving images from search engines, we collected supplementary textual data from search engines. For every input source sentence X, we follow the same approach as outlined in Section 3.1 to generate M search queries $(q_1, \ldots, q_M)$. Subsequently, we collect M sentences $(T_1, \ldots, T_M)$ that contains all the terms present in the respective search queries $(q_i \subseteq T_i, 1 \leq i \leq M)$.

Text-Aware Attentive Supplementary Text Encoder Each gathered supplementary text T_m $(m = 1, \ldots, M)$ is transformed into a $N \times 1024$ dimensional feature vector using BERT (Devlin et al., 2018), where N denotes the length of the gathered text data. To ensure consistency, these textual feature vectors are subsequently padded to match the dimensions of $L \times 1024$ $(L = 196)$, aligning them with the visual feature vectors. These feature vectors are then integrated into the scaled dot-product attention module as keys and values, with the average pooling C' representing the input text serving as the query. The resultant supplementary text representation is then passed to the translation decoder.

4. Experiment Setup

4.1. Dataset

We conducted experiments on five commonly used machine translation datasets, including multimodal machine translation dataset Multi30k (Elliott et al., 2016) English-to-German, Global Voices (Tiedemann, 2012) English-to-German , and WMT' 16 (100k) English-to-German (Newstest2016 as the test set)[2] , Bible (Christodouloupoulos and

dataset	training set	dev set	test set
Multi30k	29,000	1,014	1,000
Global Voices	69,227	2,000	2,000
WMT'16 (100k)	100,000	2,000	3,000
Bible	56,734	1,953	1,821
MultiUN	56,235	4,000	4,000

Table 2: Statistics of datasets

Steedman, 2015) English-to-German, and MultiUN (Eisele and Chen, 2010) English-to-German. The statistics for each dataset are presented in Table 2.

4.2. Model Implementation

For image retrieval, we used the Microsoft Bing[3] as the image search engine. In contrast, for supplementary text retrieval, we gathered sample sentences that included all the terms found in the respective search queries by referencing the Microsoft Bing Dictionary[4]. As described in Section 3.1 and Section 3.3, we set M to 5. This choice signifies that we formulated 5 search queries and procured 5 images or supplementary text instances[5] for every source language sentence.

Regarding the text encoder, we used a bidirectional RNN with GRU to extract text features. Specifically, we used a 256 dimensional single-layer forward RNN and a 256 dimensional single-layer backward RNN. For the translation decoder, we adhered to the approach proposed by Su et al. (2021) and utilized a modified cGRU with hidden states of 256 dimensions. Furthermore, we configured the embedding sizes for both source and target words to be 128.

As described in Section 3.1, the visual encoder we employed leveraged the $\mathrm{res4f}$ layer of a pretrained ResNet-50(He et al., 2016) model to extract visual features of dimensions 196×1024. Furthermore, as described in Section 3.3, the supplementary text encoder utilized a BERT model pretrained on the BooksCorpus(Zhu et al., 2015) and English Wikipedia[6]. This model was employed to extract

---

	Method	BLEU Score
Text-only NMT	Bi-LSTM (Calixto et al., 2017)	33.70
	Transformer (Zhang et al., 2019)	36.86
MMT with Original Images	Zhang et al. (2019)	36.86
	Zhao et al. (2021)	38.40
	Su et al. (2021)	39.20
	Tang et al. (2022) (Section 3.1)	38.14
MMT with Retrieved Images	Zhang et al. (2019)	36.94
	Tang et al. (2022) (Section 3.1)	38.43
	MMT with Visual Noise Filtering (Section 3.2)	38.51
NMT with Retrieved Supplementary Text (Section 3.3)		**39.13**

Table 3: Results on Multi30K

Method	Dataset				
	Multi30k	Global Voices	WMT'16 (100k)	Bible	MultiUN
Text-only NMT	33.70	9.22	7.99	35.23	39.49
MMT with Random Images	37.65	9.29	8.11	35.31	39.48
MMT with Blank Images	37.79	9.46	8.31	35.39	39.52
MMT with Retrieved Images	**38.43**	**9.81**	**8.41**	**35.42**	**39.53**

Table 4: Translation performance across diverse datasets under varied image conditions (BLEU score)

textual features of dimensions $N \times 1024$, where N represents the length of the retrieved supplementary text.

Regarding the noise image filter, we set $M' = 10$ and used a CLIP model (Radford et al., 2021) pretrained on the YFCC100M dataset (Thomee et al., 2016) to filter out noisy images. For the noise region fitler, we configured it with $O = 128$. Here, we utilized a pretrained Faster R-CNN model (Ren et al., 2015) that had been trained on the Open Images dataset (Kuznetsova et al., 2020). This model was employed to identify and filter noisy regions in images effectively.

4.3. Training Parameters

We initiated the word embeddings and other associated model parameters randomly, following a uniform distribution with a range of -0.1 to 0.1. During training, we employed the Adam optimizer with a mini-batch size of 32 and set the learning rate to 0.001. Additionally, a dropout strategy with a rate of 0.3 was applied to further enhance the models. The training process continued for up to 15 epochs, with early stopping activated if the BLEU (Papineni et al., 2002) score on the development set did not exhibit improvement for 3 consecutive epochs. The model with the highest BLEU score on the dev set was selected for evaluation on the test set. To minimize the impact of random seeds on experimental results and ensure result stability, we conducted the experiment 5 times with fixed random seeds and reported the macro-average of BLEU scores as the final result.

4.4. Baselines

In the case of the Multi30k dataset, we conducted a quantitative comparison of the probing methods with several recent MMT models (Zhang et al., 2019; Zhao et al., 2021; Su et al., 2021; Tang et al., 2022). However, the main focus of this research is to evaluate the necessity of visual information within real-world translation scenarios. Four out of the five datasets utilized in our evaluation experiments are authentic text-only translation datasets without any visual annotation. Consequently, for each dataset, we exclusively employed the text-only Bi-LSTM (Calixto et al., 2017) as a baseline.

The baseline model and the models detailed in Section 3 were all trained using the same training set and identical training parameters. For all these models, we present the 4-gram BLEU score (Papineni et al., 2002) as the primary evaluation metric.

5. Results and Analysis

Table 3 presents the experimental results of the Multi30k dataset. Compared to various baseline models, all three probing methods mentioned in Section 3 have achieved promising results. Notably, the MMT model with visual noise filtering (Section 3.2) achieved a BLEU score of 38.51, while the NMT model with retrieved supplementary text (Section 3.3) achieved an impressive BLEU score of 39.13. In comparison to text-only NMT models (Calixto et al., 2017; Vaswani et al., 2017), the NMT model with retrieved supplementary text significantly outperforms them, showcasing a sub-

stantial increase in BLEU score. When compared to existing MMT methods that utilize original images (Zhang et al., 2019; Zhao et al., 2021; Su et al., 2021), the NMT model with retrieved supplementary text obtains a comparable BLEU score. Furthermore, in contrast to the MMT methods with retrieved images (Zhang et al., 2019; Tang et al., 2022), the NMT model with retrieved supplementary text demonstrates performance gains of approximately 2.2 and 0.7 BLEU points, respectively.

Further experimental results and analysis will be presented in the following sections.

5.1. Translation Performances across Varied Datasets

We firstly quantitatively compared text-only NMT (Calixto et al., 2017) with MMT utilizing retrieved images (Section 3.1) across five diverse datasets mentioned in Section 4.1. As demonstrated in Table 4, MMT achieved significantly higher BLEU scores on Multi30k, higher BLEU scores on Global Voices and WMT'16 (100k), and slightly higher BLEU scores on Bible and MultiUN. It is intriguing to note that the improvement in translation performance is substantial on Multi30k, with an increase of approximately 4.7, whereas the gain on MultiUN is relatively modest, at approximately 0.04.

We speculate that the variations in results among the aforementioned translation datasets, such as Multi30k and other datasets, may be attributed to the differing qualities of images collected through the search engine. To evaluate the influence of the quality of collected images, we train the MMT model with randomly retrieved unrelated images, blank images, and retrieved images from image search engines, respectively.

The evaluation results are shown in table 4. It is obvious that MMT models with retrieved images achieves the highest BLEU score on all Multi30k and other four datasets, demonstrating the effectiveness of visual information from retrieved images. Compared with the model with random images and blank images, the performance gain of collected images is approximately 0.7 & 0.6 BLEU score on Multi30k, and 0.5 & 0.3 BLUE score on Global Voices. However, on WMT'16 (100k), Bible, and MultiUN datasets, models with retrieved images achieve almost the same BLEU score as the model with blank images.

One of the possible reason is that sentences from those three datasets contains fewer entity words that can be represented by images, and therefore, the search engine based image retrieval method collects numbers of noise images. Sentences from WMT'16 (100k), Bible, and MultiUN datasets describe abstract concepts and complex events, while sentences from Multi30k and Global Voices describe real objects and people, which is more reliable for image retrieval. [7]

To validate the hypotheses, we manually analyzed the image retrieval outcomes of each dataset. In detail, we initially conducted a random sampling of 1,000 sentences and employed the image retrieval methods outlined in Section 3.1 to gather keywords and images for each sentence. Regarding the extracted keywords, we conducted manual assessments to identify whether each keyword qualifies as an entity word. Regarding the collected images, we carried out manual evaluations to determine if an image could offer pertinent visual information for the search query, and those that could not. Images in the latter category were categorized as noise images. Lastly, we tallied the quantity of sentences containing at least half of non-entity keywords and the quantity of sentences harboring at least half of noise images among the collected images.

As presented in Table 5, for the Multi30k dataset, out of 1000 sentences, only 27 sentences contained half or more non-entity keywords, and 61 sentences gathered half or more noise images from search engines. However, in the WMT'16 (100k) dataset, there are 796 sentences with half or more non-entity keywords and 685 sentences with half or more noise images, accounting for more than half of the sampled sentences. Consequently, our method shows poor performance on the WMT'16 (100k) dataset. The Bible dataset and MultiUN dataset exhibit a similar situation. For the Global Voices dataset, there are 94 sentences with half or more non-entity keywords and 228 sentences with half or more noise images. This falls between the Multi30K and WMT'16 (100k) datasets. It is interesting to note that the Multi30k dataset, which has the smallest proportions of non-entity keywords and noise images, achieves the most significant gain in translation performance. On the other hand, datasets with the largest proportions of non-entity keywords and noise images show the smallest gain in translation performance.

5.2. Influence of the Correlation between Text and Images

Table 6 shows the evaluation results of applying two filtering approaches described in Section 3.2 in MMT. It is obvious that MMT models with both noise image filter and noise region filter achieves the highest BLEU score across all datasets, including Multi30k and the other four,underscoring the

---

[7]Examples of retrieved images from various datasets are presented in Table 8.

	Multi30k	Global Voices	WMT'16 (100k)	Bible	MultiUN
Number of sentences with half or more non-entity keywords	27	94	796	398	818
Number of sentences with half of more noise images	61	228	685	761	663

Table 5: Summary of manual analysis of image retrieval outcomes for each dataset

Method	Dataset				
	Multi30k	Global Voices	WMT'16 (100k)	Bible	MultiUN
MMT with retrieved images (Tang et al., 2022)	38.43	9.81	8.41	35.42	39.53
+ noise image filter	38.50	10.12	8.89	36.12	39.91
+ noise region filter	38.46	9.95	8.78	35.84	39.72
+ noise image & region filter	**38.51**	**10.23**	**8.93**	**36.38**	**39.95**

Table 6: Results of image and region filtering method across diverse datasets (BLEU score)

effectiveness of these two filtering approaches. [8]

Notably, it is intriguing to note that the noise filtering techniques exhibited more substantial enhancements in translation performance for the WMT'16 (100k), Bible, and MultiUN datasets, in contrast to the improvements observed in the Multi30k and Global Voices datasets. This further underscores the significant impact of the correspondence between image and text content on the translation performance the alignment and coherence between image and text content on the translation performance of the MMT system. It also elucidates why noise filtering methods yield marginal improvements on the Multi30K dataset.

In conclusion, the translation performance of the multimodal model primarily hinges on the consistency between textual and visual content. In other words, the more alignment exists between textual and visual content, the greater enhancement in translation performance with multimodal translation compared to text-only translation. Hence, we arrive at a conclusion that aligns closely with (Caglayan et al., 2019), which suggest that multimodal translation models predominantly treat visual information as a complement to textual information.

5.3. Exploring the Necessity of Visual Modality

We conducted a quantitative comparison between MMT with retrieved images (Section 3.1) and NMT with retrieved supplementary texts on the Multi30k dataset. Table 7 shows the experimental results. In comparison to MMT model employing images for translation enhancement, the approach integrating

Method	BLEU score
text-only NMT	33.70
+ visual information (MMT with retrieved images) (Tang et al., 2022)	38.43
+ textual information	**39.13**
+ visual & textual information	38.55

Table 7: Results on Multi30k using visual information or textual information enhanced NMT

supplementary textual data for translation enhancement demonstrated a significantly higher BLEU score of 39.13. Remarkably, the combined utilization of both images and supplementary texts for translation enhancement yielded a BLEU score of 38.55, positioning itself between image-enhanced NMT and text-enhanced NMT.

This demonstrates that both additional visual and supplementary textual information play an entirely equivalent supplementary role in the translation process. Moreover, in most cases, the utilization of supplementary textual information assists the translation process more effectively. [9]

Therefore, we speculate that multimodal translation models trained on a large volume of data might face challenges in outperforming text-only translation models trained on comparable data volumes. This is because as the volume of data used in multimodal model training increases, the potential impact of visual information could diminish. We will verify this in future work.

---

[8]A correct example generated by MMT with visual noise filtering is presented in Table 9.

[9]A correct example comparing NMT with retrieved supplementary texts to MMT with retrieved images is presented in Table 10.

6. Conclusions

In this paper, we conduct an in-depth exploration into the role of visual information within the multimodal translation process on Multi30k and other four authentic translation datasets. Our findings emphasize that the substantial correlation between visual and textual content significantly impacts the efficacy of multimodal translation, while employing filtering mechanisms based on the textual-visual correlation can enhance translation performance. Additionally, experimental results reveal that visual information plays a supplementary role in the multimodal translation process. This supplementary function of visual information can be substituted by the incorporation of supplementary textual information. As one of our future work, we plan to assess the impact of the visual modality on more extensive translation datasets, including the complete WMT'16 dataset. We speculate that as multimodal translation models are trained using larger datasets, the impact of visual information is likely to diminish.

Acknowledgements

References

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

American Psychological Association. 1983. *Publications Manual*. American Psychological Association, Washington, DC.

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L_1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014a. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014b. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. 2018. Findings of the third shared task on multimodal machine translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, volume 2, pages 308–327.

Ozan Caglayan, Loïc Barrault, and Fethi Bougares. 2016. Multimodal attention for neural machine translation. *arXiv preprint arXiv:1609.03976*.

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*, pages 4159–4170. Association for Computational Linguistics.

Iacer Calixto and Qun Liu. 2017. Incorporating global visual features into attention-based neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 992–1003.

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

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. Alternation. *Journal of the Association for Computing Machinery*, 28(1):114–133.

Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Christos Christodouloupoulos and Mark Steedman. 2015. A massively parallel corpus: The Bible in 100 languages. *Language Resources and Evaluation*, 49(2):375–395.

Jonathan H Clark, Chris Dyer, Alon Lavie, and Noah A Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 176–181.

James W. Cooley and John W. Tukey. 1965. An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation*, 19(90):297–301.

Jean-Benoit Delbrouck and Stéphane Dupont. 2017a. An empirical study on the effectiveness of

images in multimodal neural machine translation. *arXiv preprint arXiv:1707.00995*.

Jean-Benoit Delbrouck and Stephane Dupont. 2017b. Multimodal compact bilinear pooling for multimodal neural machine translation. *arXiv preprint arXiv:1703.08084*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Andreas Eisele and Yu Chen. 2010. MultiUN: A multilingual corpus from united nation documents. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*.

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

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

Desmond Elliott, Stella Frank, and Eva Hasler. 2015. Multilingual image description with neural sequence models. *arXiv preprint arXiv:1510.04709*.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30k: Multilingual english-german image descriptions. *arXiv preprint arXiv:1605.00459*.

Jonas Gehring, Michael Auli, David Grangier, and Yann N Dauphin. 2016. A convolutional encoder model for neural machine translation. *arXiv preprint arXiv:1611.02344*.

Alex Graves. 2012. Long short-term memory. *Supervised sequence labelling with recurrent neural networks*, pages 37–45.

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

Stig-Arne Grönroos, Benoit Huet, Mikko Kurimo, Jorma Laaksonen, Bernard Merialdo, Phu Pham, Mats Sjöberg, Umut Sulubacak, Jörg Tiedemann, Raphael Troncy, et al. 2018b. The memad submission to the wmt18 multimodal translation task. *arXiv preprint arXiv:1808.10802*.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

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

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

Julia Ive, Pranava Madhyastha, and Lucia Specia. 2019. Distilling translations with visual awareness. *arXiv preprint arXiv:1906.07701*.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. 2020. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981.

Chiraag Lala, Pranava Swaroop Madhyastha, Carolina Scarton, and Lucia Specia. 2018. Sheffield submissions for WMT18 multimodal translation shared task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 624–631, Belgium, Brussels. Association for Computational Linguistics.

Bei Li, Chuanhao Lv, Zefan Zhou, Tao Zhou, Tong Xiao, Anxiang Ma, and JingBo Zhu. 2022. 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.

Zuchao Li, Rui Wang, Kehai Chen, Masao Utiyama, Eiichiro Sumita, Zhuosheng Zhang, and Hai

Zhao. 2020a. Explicit sentence compression for neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8311–8318.

Zuchao Li, Rui Wang, Kehai Chen, Masso Utiyama, Eiichiro Sumita, Zhuosheng Zhang, and Hai Zhao. 2020b. Data-dependent gaussian prior objective for language generation. In *Eighth International Conference on Learning Representations*.

Huan Lin, Fandong Meng, Jinsong Su, Yongjing Yin, Zhengyuan Yang, Yubin Ge, Jie Zhou, and Jiebo Luo. 2020. Dynamic context-guided capsule network for multimodal machine translation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1320–1329.

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

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.

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.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

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

Jinsong Su, Jinchang Chen, Hui Jiang, Chulun Zhou, Huan Lin, Yubin Ge, Qingqiang Wu, and Yongxuan Lai. 2021. Multi-modal neural machine

translation with deep semantic interactions. *Information Sciences*, 554:47–60.

ZhenHao Tang, XiaoBing Zhang, Zi Long, and XiangHua Fu. 2022. Multimodal neural machine translation with search engine based image retrieval. In *Proceedings of the 9th Workshop on Asian Translation*, pages 89–98.

Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Peter D Turney. 2000. Learning algorithms for keyphrase extraction. *Information retrieval*, 2(4):303–336.

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

Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. 2015. Sequence to sequence-video to text. In *Proceedings of the IEEE international conference on computer vision*, pages 4534–4542.

Ian H Witten, Gordon W Paynter, Eibe Frank, Carl Gutwin, and Craig G Nevill-Manning. 2005. Kea: Practical automated keyphrase extraction. In *Design and Usability of Digital Libraries: Case Studies in the Asia Pacific*, pages 129–152. IGI global.

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.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with

visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR.

Jingyi Zhang, Masao Utiyama, Eiichiro Sumita, Graham Neubig, and Satoshi Nakamura. 2017. Nict-naist system for wmt17 multimodal translation task. In *Proceedings of the Second Conference on Machine Translation*, pages 477–482.

Zhuosheng Zhang, Kehai Chen, Rui Wang, Masao Utiyama, Eiichiro Sumita, Zuchao Li, and Hai Zhao. 2019. Neural machine translation with universal visual representation. In *International Conference on Learning Representations*.

Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. 2020. Semantics-aware bert for language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9628–9635.

Yuting Zhao, Mamoru Komachi, Tomoyuki Kajiwara, and Chenhui Chu. 2021. Word-region alignment-guided multimodal neural machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:244–259.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.

A. Qualitative Examples

In this appendix, we provide examples of retrieved images (Table 8), as well as translation examples for MMT with visual noise filtering (Table 9) and NMT with retrieved supplementary texts (Table 10).

Dataset	English Sentence	One of five retrieved images
Multi30k	The person in the striped shirt is mountain climbing.	
Global Voices	Now the city is under a siege from the security forces.	
WMT'16 (100k)	In the future, integration will be a topic for the whole of society even more than it is today.	
Bible	You are Yahweh, even you alone. You have made heaven. the heaven of heavens, with all their army, the earth and all things that are on it, the seas and all that is in them and you preserve them all.	
MultiUN	Development assistance cannot by itself prevent or end conflict.	

Table 8: Examples of retrieved image from different datasets. For the sentence from Multi30k dataset, our method efficiently retrieves an image that accurately represents the sentence's content "A man is rock climbing". For the sentence from Global Voice dataset, the retrieved image exhibits a degree of alignment with the source sentences, encompassing elements like "city","siege" and"forces". However, for the sentence from WMT'16 (100k), Bible and MultiUN datasets, it becomes evident that the retrieved images offer limited relevant visual information and thus provide little assistance for translation.

Source (En)	But he answered and said, "Every plant which my heavenly Father didn't plant will be uprooted.
Target (De)	Aber er antwortete und sprach: Alle Pflanzen, die mein himmlischer Vater nicht pflanzte, die werden ausgereutet.
Retrieved images	
MMT with retrieved images	Er antwortete aber und sprach: Alle Pflanzen, die mein himmlischer Vater nicht verderbte Quelle.
Retrieved images with noise image filter	
MMT with noise image filter	Er antwortete aber und sprach: Alle Pflanzen, die mein himmlischer Vater nicht pflanzte.
MMT with both noise image and region filter	Er antwortete aber und sprach: Alle Pflanzen, die mein himmlischer Vater nicht pflanzte, wird entwurzelt werden.

Table 9: A correct example generated by MMT with visual noise filtering. Due to its unique characteristics, the Bible dataset contains numerous entity words but is challenging to obtain images that effectively represent the textual content. However, visual noise filtering based on visual-text correlation can partially alleviate this situation. In this example, the filtered visual information has enabled the translation of "uprooted" to be correct.

Source (En)	Group of Asian boys wait for meat to cook over barbecue.
Target (De)	Eine Gruppe asiatischer Jungen wartet am Grill darauf, dass Fleisch gar wird.
Text-only NMT	Eine asiatische Jungen warten auf dem Fleisch, um den Grill zu kochen.
Retrieved images	
MMT with retrieved images	Eine Gruppe von asiatischen Jungen wartet darauf, um Fleisch zu grillen.
Retrieved supplementary texts	(1) Delivery is hardly limited to pizza at this point; everything from sushi to barbecue seems available as a to-go order. (2) While the savory aroma of barbecue filled the air, friends and family gathered around the grill, eagerly sharing stories and laughter as they waited for the delicious meal to be ready. (3) As the sun dipped below the horizon, our group of friends decided to have a barbecue in the backyard, lighting up the grill and eagerly waiting for the charcoal to heat up so that we could start cooking our favorite dishes. (4) At the lively outdoor barbecue gathering, a diverse group of friends, including a talented Asian chef, couldn't wait to cook up a mouthwatering feast. (5) While the enthusiastic Asian group gathered around the barbecue, they took turns to cook their favorite dishes, making everyone else eagerly wait in anticipation of the delicious meal.
MMT with retrieved supplementary texts	Eine Gruppe von asiatischen Jungen wartet darauf, dass Fleisch über Grill zukochen.

Table 10: A correct example generated by NMT with retrieved supplementary texts. In this example, in contrast to text-only NMT without any supplementary information, visual information and supplementary text information play an equivalent role, correctly translating "Group" to "Gruppe". Benefiting from the rich information in the supplementary text, the NMT with retrieved supplementary text achieves more accurate translations compared to MMT with retrieved images.