Leveraging Visual Scene Graph to Enhance Translation Quality in Multimodal Machine Translation

Ali Hatami[1], Mihael Arcan[2], Paul Buitelaar[1]

[1]Insight Research Ireland Centre for Data Analytics,
Data Science Institute, University of Galway, Ireland
[2]Lua Health, Galway, Ireland,

Correspondence: ali.hatami@insight-centre.org

Abstract

Despite significant advancements in Multimodal Machine Translation, understanding and effectively utilising visual scenes within multimodal models remains a complex challenge. Extracting comprehensive and relevant visual features requires extensive and detailed input data to ensure the model accurately captures objects, their attributes, and relationships within a scene. In this paper, we explore using visual scene graphs extracted from images to enhance the performance of translation models. We investigate this approach for integrating Visual Scene Graph information into translation models, focusing on representing this information in a semantic structure rather than relying on raw image data. The performance of our approach was evaluated on the Multi30K dataset for English into German, French, and Czech translations using BLEU, chrF2, TER and COMET metrics. Our results demonstrate that utilising visual scene graph information improves translation performance. Using information on semantic structure can improve the multimodal baseline model, leading to better contextual understanding and translation accuracy.

1 Introduction

Neural Machine Translation (NMT) has significantly advanced translation quality compared to earlier methods, showcasing remarkable improvements in fluency and precision (Cho et al., 2014). Transformer-based models enhanced performance by effectively capturing semantic dependencies and producing fluent, contextually relevant translations (Vaswani et al., 2017).

However, despite these advancements, text-only NMT models face persistent challenges in translating the input text (Wang and Xiong, 2021; Zhao et al., 2022). Resolving ambiguity in the input sentence is one of these challenges (Futeral et al., 2023; Bowen et al., 2024; Hatami et al., 2024).

To address these limitations, researchers have explored Multimodal Machine Translation (MMT), a subfield of NMT that integrates visual information from images or videos to enhance translation models (Yao and Wan, 2020; Wang and Xiong, 2021; Zhao et al., 2022). MMT leverages visual content as a complementary source of information to aid in understanding the source text and resolving ambiguities. Text-only NMT models might struggle to translate ambiguous sentences, but an accompanying image can provide crucial visual cues for disambiguation, enabling the model to select the correct translation.

Despite its potential, MMT presents its own challenges. Visual resources, such as images, often contain a large amount of information, not all of which is relevant to the translation task. This extra information can not only fail to improve translation quality but may even degrade it. In addition, training an MMT model requires a vast amount of visual information covering different objects and their relationships.

To address these challenges, recent studies have focused on identifying and incorporating the most relevant visual information into translation models (Lala and Specia, 2018; Fei et al., 2023; Yin et al., 2023; Hatami et al., 2023). These papers examine the importance of using visual information by focusing on lexical ambiguity in the input text to find relevant information on the visual side.

In this paper, we study the impact of using Visual Scene Graphs (VSGs), which represent objects and their relationships within an image, as a means to enhance MMT models. First, we extract VSGs as a semantic structure from images and then utilize this information as triples to train our translation model. Our work differs from previous studies by directly leveraging VSGs to represent objects and their relationships, providing a structured se-

Proceedings of Machine Translation Summit XX Volume 1, pages 353–364, June 23-27, 2025

mantic context for translation. We evaluated our approach on the Multi30K dataset for English into German, French and Czech translations. The results demonstrate that the use of VSGs in MMT leads to notable improvements in both quantitative metrics and qualitative evaluations, highlighting the potential of this approach for advancing the field of multimodal translation.

2 Related Work

In recent years, MMT has gained significant attention to enhance traditional text-only translation by incorporating visual information. MMT models primarily relied on image features extracted from vision-based transformers to improve translation quality, particularly in cases of ambiguity or lexical uncertainty (Delbrouck and Dupont, 2017). Early approaches to MMT incorporated joint multimodal embeddings to fuse textual and visual features. Calixto et al. (2017) proposed an attention-based framework that used convolutional neural networks (CNNs) to extract image features, which were then integrated into a sequence-to-sequence NMT model. Similarly, Libovický and Helcl (2017) introduced hierarchical attention mechanisms to balance contributions from different modalities dynamically.

Some other papers explored transformer-based architectures to enhance multimodal fusion. Wu et al. (2021) adapted the Transformer model by introducing multimodal self-attention, enabling better integration of visual and textual features. Caglayan et al. (2019) demonstrated that incorporating region-based visual features (e.g., using object detectors like Faster R-CNN) improved MMT performance by focusing on semantically relevant image regions.

Despite advancements, challenges remain in effectively integrating multimodal information without introducing noise. Elliott (2018) found that while images help in specific cases, text-only models often outperform multimodal ones when trained on large-scale datasets. This has led to investigations into adaptive multimodal fusion techniques, where the model selectively uses visual information only when beneficial (Hatami et al., 2024).

Recent advancements in MMT have explored the integration of structured visual knowledge to enhance translation quality. Yin et al. (2020) proposed a graph-based multimodal fusion encoder for NMT, leveraging Graph Neural Networks (GNNs) to encode multimodal information more effectively. By structuring both visual and textual inputs into a graph representation, their model captures semantic relationships between objects, improving the contextual grounding of translations. These studies highlight the growing importance of structured vision-language representations, such as scene graphs and graph-based encoders, in addressing the challenges of multimodal translation, particularly in ambiguous and resource-constrained settings.

Incorporating knowledge graphs into NMT has proven effective in improving the translation of named entities and specialized terminology, as demonstrated by Moussallem et al. (2019). Their approach introduced two strategies: Entity Linking with Knowledge Bases, which enriched NMT embeddings through multilingual entity linking, and Surface Form Initialization, which optimized entity vector values without explicit linking. By leveraging structured knowledge representations, their method enhanced translation accuracy, particularly in handling domain-specific terms and low-resource scenarios.

Unsupervised MMT (UMMT) system introduced by Fei et al. (2023) that utilises scene graphs as a pivoting mechanism to perform inference-time image-free translation through visual scene hallucination. Their method generates synthetic scene graphs from textual input, enabling multimodal translation even in the absence of actual image inputs. This approach effectively bridges the gap between vision and language representations, demonstrating improved translation performance in low-resource and zero-resource scenarios.

Although VSGs are widely used in various multimodal tasks such as image captioning (Yang et al., 2018), visual question answering (Hildebrandt et al., 2020), and image retrieval (Johnson et al., 2018), they remain underexplored in the multimodal translation task. VSGs provide a powerful representation for understanding image semantics by capturing objects, their attributes, and relationships in a structured graph format. In the context of MMT, leveraging the structured and interpretable visual information provided by scene graphs has the potential to enhance the translation process by improving contextual grounding and disambiguating visually dependent terms.

In our work, we propose an approach by leveraging VSGs extracted using a Multimodal Large Language Model (MLLM) to improve translation quality in MMT systems. By using MLLMs, we

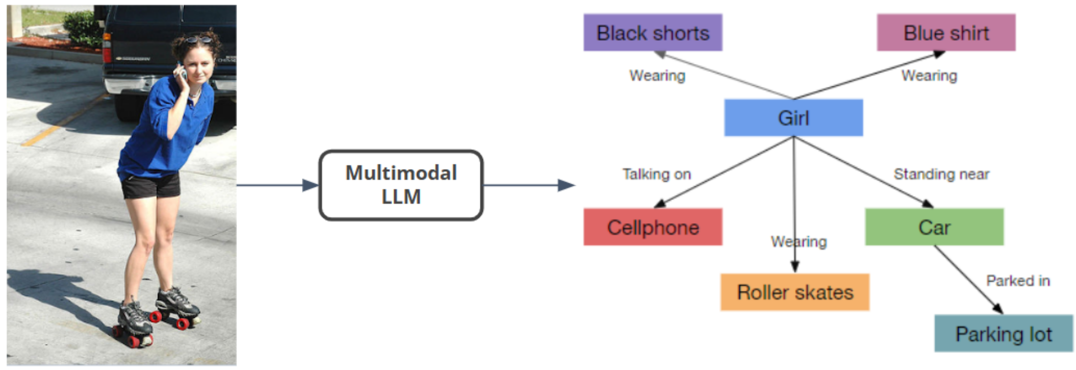

Figure 1: Example for extracting a Visual Scene Graph (VSG) from an image.

ensure accurate and detailed scene graph extraction, capturing not only objects and their relationships but also contextual nuances often missed by conventional visual models. This structured visual information is then incorporated into the translation pipeline, enabling our model to produce translations that are more contextually appropriate and semantically accurate. Figure 1 shows an example of the VSGs extracted from an image using Gemini 1.5 Flash.

To the best of our knowledge, few studies focus on extracting object relationships in MMT (Fei et al., 2023; Yin et al., 2023). By integrating scene-graph information into translation models, we aim to address the limitations of raw visual inputs and provide meaningful context for disambiguation and improved translations. Unlike prior approaches that focus on multimodal fusion without explicit scene-graph extraction or rely on hallucinated visual representations during inference, we extract VSGs from images and utilize them as triples to enhance translation quality through structured semantic learning. The integration of triples aims to provide contextual information about the scene, potentially disambiguating lexical or syntactic ambiguities in the text. Our results demonstrate that incorporating the VSG information yields better performance compared to using raw images as visual input.

3 Methodology

In this section, we explain our methodology for extracting scene graph information from images and utilising it in the translation process.

3.1 Visual Scene Graph Extraction

To integrate visual information into the translation model, we extract Visual Scene Graphs (VSGs)

Prompt

```
Extract the visual scene graph as triples from the
provided image and save it in a Python list.

Number identical objects if more than one exists,
and ensure the visual scene graph is in English.
```

```
[("girl", "wearing", "black shorts"), ("girl",
"wearing", "blue shirt"), ("girl", "talking
on", "cellphone"), ("girl", "wearing", "roller
skates"), ("girl", "standing near", "car"),
("car", "parked in", "parking lot")]
```

Figure 2: Prompt example for extracting a Visual Scene Graph (VSG) from an image in triples format using Gemini.

in English from images. VSGs provide structured representations of images in a triple format (subject, relationship, object), capturing object relationships and semantic context. This structure encodes visual information in a textual format, covering all objects and their relationships within the scene.

We use Gemini 1.5 Flash as a multimodal LLM to generate Visual Scene Graphs (VSGs) from images. Gemini includes parameters such as temperature, top_P, and safety settings to control generating the output. These parameters are explained in Section 4.2 in more detail. After configuring these parameters, the model generates VSGs from images for the training, validation, and test sets based on the provided prompt. Figure 2 shows the prompt used to extract VSG from the given image.

To ensure a consistent output format, we enforced the model to generate VSGs in a Python list, preventing variations in format. We also restricted the model to generate VSGs strictly in English to reduce hallucinations, as it sometimes defaulted to other languages based on the image context. Ad-

Prompt 1

```
Translate the following English sentence to German:

A trendy girl talking on her cellphone while gliding
slowly down the street.
```

Prompt 2

```
Translate the following English sentence to German:

A trendy girl talking on her cellphone while gliding
slowly down the street.

Use the following triples to ensure the translation
is correct:

girl | wearing | black shorts
girl | wearing | blue shirt
girl | talking on | cellphone
girl | wearing | roller skates
girl | standing near | car
car | parked in | parking lot
```

Prompt 4

```
Translate the following English sentence to German:

A trendy girl talking on her cellphone while gliding
slowly down the street.

Use the following triples and image to ensure the
translation is correct:

girl | wearing | black shorts
girl | wearing | blue shirt
girl | talking on | cellphone
girl | wearing | roller skates
girl | standing near | car
car | parked in | parking lot
```

Prompt 3

```
Translate the following English sentence to German:

A trendy girl talking on her cellphone while gliding
slowly down the street.

Use the following image to ensure the translation is
correct:
```

Figure 3: Prompt examples that we used for T5 and Gemini to translate the input text from English to German; Prompt 1: Text-to-Text translation, Prompt 2: Text+Triples-to-Text translation, Prompt 3: Text+Image-to-Text translation, Prompt 4: Text+Triples+Image-to-Text translation.

ditionally, we numbered identical objects in the VSGs to improve scene comprehension when multiple identical objects were present.

3.2 Training Text-to-Text Model

Text-to-Text (T2T) translation is a baseline approach in which the model is used to translate the input text from the source language into the target language. For T2T translation, we utilise four models: NMT-T2T, mT5_Base, NLLB-200, and Gemini. NMT-T2T is a transformer-based model trained on the dataset, while mT5_Base and NLLB-200 are fine-tuned on the dataset. Additionally, we use Gemini for zero-shot translation of the test sets.

Prompt 1 in Figure 3 illustrates an example prompt used for mT5 and Gemini to translate the input sentence from English into German. Unlike mT5 and Gemini, which are multitask models requiring prompt instructions for translation, NLLB-200 is specifically trained for translation tasks. Therefore, we simply provide the input sentence to the fine-tuned NLLB-200 model to generate the translation.

3.3 Training Text+Triplets-to-Text Model

To investigate the impact of incorporating VSG, we enriched the source text with the information extracted from VSG. In Text+Triples-to-Text (TT2T) translation, we incorporate this information (Section 3.1) into the training process of the translation model. By augmenting the text with structured visual-contextual information, we aimed to assess whether the inclusion of triples improves the ability of the models to capture implicit meanings and context that are otherwise absent in text-only inputs.

For TT2T translation, we concatenate these triples with the English input text to provide additional context, helping the model better understand the input. This approach leverages semantic insights from visual relationships in a textual format, enhancing translation quality without directly using images. Similar to T2T, in TT2T, we utilise four models: NMT-T2T, mT5_Base, NLLB-200, and Gemini. We train NMT-T2T on input text enriched with triple information, along with the corresponding output text. We also fine-tune mT5_Base and NLLB-200 on input text enriched with triple information. For Gemini, we apply zero-shot translation to translate test set sentences while incorporating triple information to ensure accurate translation.

Prompt 2 in Figure 3 presents an example prompt used for mT5 and Gemini to translate an input sentence from English to German. By adding triples extracted from the paired image, we guide the model to consider semantic information from the image when translating. This approach ensures that the translation aligns correctly with the visual context.

3.4 Training Text+Image-to-Text Model

In Text+Image-to-Text (TI2T) translation, we use the input text along with an image to train the model. For TI2T translation, we utilise two models: MMT-TI2T and Gemini. MMT-TI2T is a gated fusion multimodal model trained on the training and validation sets. For Gemini, we use zero-shot translation of the given sentence, considering the paired image. Prompt 3 in Figure 3 indicates the example prompt in TT2T translation from English to German. In the prompt, we provide an instruction to the model to use the given image to make sure the translation is correct.

3.5 Training Text+Triplets+Image-to-Text Model

For Text+Triples+Image-to-Text (TTI2T) translation, we add triples extracted from Visual Scene Graphs (VSGs) as additional information to the translation model alongside the input text and image. The reason behind this approach is that using images alone may introduce noise and degrade the performance of the translation model. By incorporating structured semantic information from the scene graph along with the image, enables the model to incorporate both low-level visual details and high-level relational knowledge into the translation process.

For TTI2T, we employ two multimodal translation models: MMT-TI2T and Gemini. We explain both models in Section 3.4. The only difference is that TTI2T additionally provides extracted triples along with the input text and image.

Prompt 4 in Figure 3 shows an example prompt for TTI2T translation from English to German. In the prompt, we instruct the model to use the given image and triples to ensure the translation is accurate.

4 Experimental Setup

In this section, we provide insights into the dataset used in this work, extracting VSG from images, settings for text-only and multimodal models, and the translation evaluation metrics BLEU, ChrF2, TER and COMET.

4.1 Multi30k Dataset

Multi30K (Elliott et al., 2016) is an extension of the Flickr30K Entities dataset that consists of 29,000 images paired with descriptions in English, along with translated sentences in German, French, and Czech (Elliott et al., 2017). The dataset is specifically designed for evaluating MMT systems, where both textual and visual information are utilised for translation tasks. Multi30K also provides three test sets: the 2016 and 2017 test sets, each with 1,000 images, and the 2018 test set with 1,071 images.

4.2 Gemini 1.5 Flash

To extract VSGs from the Multi30K dataset, we used Gemini 1.5 Flash [1], a pre-trained LLM to analyse the multimodal data. For our experiment, we used Gemini through the free-tier API, which provides a rate limit of 15 requests per minute (RPM) and 1,500 requests per day (RPD). We set the default inference parameters for the model. These defaults included a temperature of 1.0, ensuring a balanced mix of randomness and determinism in responses, a Top-p sampling set to 0.95, allowing diverse but high-probability token selections, and a maximum output length of 8,192 tokens. The default Top-k setting was automatically adjusted by the system. To ensure comprehensive processing of all images in the dataset, we configured the model's safety settings, including thresholds for "Harassment", "Hate Speech", "Sexually Explicit Content", and "Dangerous Content" to "BLOCK_NONE". This adjustment allows the model to generate responses for every image ensuring that outputs are returned in full without being restricted by safety mechanisms. After setting the parameters, the model generated VSG from the image in our dataset based on the given prompt (Figure 2).

Gemini 1.5 Flash is capable of processing both text and visual information. For text-only and multimodal translation, we also employed Gemini, maintaining the same parameter settings and safety configurations as described in VSG extraction. The model was used for zero-shot translation from English into German, French, and Czech on the Multi30k dataset, covering both text-only and multimodal translation under different configurations. These configurations included T2T (En $\rightarrow$ De, Fr, Cs), TT2T (En + triples $\rightarrow$ De, Fr, Cs), TI2T (En + image $\rightarrow$ De, Fr, Cs), and TIT2T (En + image + triples $\rightarrow$ De, Fr, Cs). This setup allowed us to assess Gemini's capability in handling both textual and multimodal inputs across multiple

[1] https://deepmind.google/technologies/gemini/

languages.

4.3 OpenNMT

A text-only transformer model serves as the baseline in our experiment, utilising solely the textual captions of images for translation. Trained using the OpenNMT toolkit (Klein et al., 2018) on the Multi30k dataset for English to German, French, and Czech translations, the model comprises a 6-layer transformer architecture with attention mechanisms in both encoder and decoder stages, trained for 50K steps. Sentencepiece (Kudo and Richardson, 2018) is employed to segment words into subword units, offering a language-independent approach to tokenization without necessitating preprocessing steps, thus enhancing the model's adaptability and versatility in handling raw text.

4.4 Gated Fusion Multimodal

In the MMT model, we adopt the gated fusion MMT model (Wu et al., 2021) as a multimodal basline model. Gated fusion is a mechanism that is used to integrate visual information from images with textual information from source sentences by fusing visual and text representations by employing a gate mechanism.. The main idea behind gated fusion is to control the amount of visual information that is blended into the textual representation using a gating matrix. The source sentence x is fed into a vanilla Transformer encoder to obtain a textual representation H_{text} of dimension $T \times d$. The image z is processed using a pre-trained ResNet-50 CNN which has been trained on the ImageNet dataset (Deng et al., 2009) to extract a 2048-dimensional average-pooled visual representation, denoted as $Embed_{image}(z)$. The visual representation $Embed_{image}(z)$ is projected to the same dimension as H_{text} using a weight matrix W_z. A gating matrix Λ of dimension $T \times d$ is generated to control the fusion of the textual and visual representations. The gating matrix Λ is computed as:

$$\Lambda = \text{sigmoid}(W_\Lambda \text{Embed}_{\text{image}}(z) + U_\Lambda H_{\text{text}})$$

where W_Λ and U_Λ are model parameters.

4.5 NLLB-200

In this section, we outline the setup used the No Language Left Behind (NLLB) model. This model is a transformer-based multilingual NMT model designed for covering 200 languages. Due to our GPU limitation, we fine-tune NLLB-200 with 600M model on our dataset. The process involved data preprocessing, model training, hyperparameter tuning, and evaluation.

Similar to mT5, the fine-tuning process was conducted using two NVIDIA A6000 GPUs (2×48GB GPU memory). We set the learning rate to 2e-5 and used the Adam optimizer with a weight decay of 0.01 to prevent overfitting. The model was trained for 10 epochs with a per-device batch size of 16 for both training and evaluation. To ensure efficient monitoring, logging was performed every 500 steps. The training leveraged Automatic Mixed Precision (AMP) for optimized memory usage and performance.

4.6 Multilingual T5

Multilingual Text-to-Text Transfer Transformer (mT5) is a transformer-based language model designed specifically for multilingual Natural Language Processing (NLP) tasks. It extends the T5 model, which frames all NLP tasks as text-to-text problems (Raffel et al., 2020). We fine-tuned the mT5 model on the Multi30K dataset to optimise its performance in translation tasks, focusing solely on the textual modality without any information from the visual side.

One of the key features of mT5 is its support for 101 languages, making it a powerful model for multilingual applications such as translation tasks (Xue et al., 2021). The model is pretrained on mC4 (Multilingual Common Crawl), a large-scale dataset containing filtered web text from a wide range of languages. This extensive training allows mT5 to perform well in both high-resource and low-resource languages. Additionally, since mT5 is trained on a diverse dataset, it is more capable of handling syntactic and grammatical variations across different languages (Raffel et al., 2020). Supporting multiple languages makes it well-suited for machine translation, allowing us to leverage a single model without the need for separate models for different languages.

We used mT5-Base which has around 220 M parameters. When fine-tuning mT5, common settings include a learning rate of 2e-5, which helps to ensure stable convergence during training while avoiding overfitting. The batch size is set to 16 for both training and evaluation, which balances efficiency and memory constraints, though it can be adjusted depending on GPU availability. Additionally, a weight decay of 0.01 is used to reduce

	English → German				English → French				English → Czech			
	BLEU ↑	ChrF2 ↑	TER ↓	COMET ↑	BLEU ↑	ChrF2 ↑	TER ↓	COMET ↑	BLEU ↑	ChrF2 ↑	TER ↓	COMET ↑
Text-to-Text (T2T)												
NMT-T2T	41.1	65.4	43.8	0.8604	60.6	71.4	31.8	0.8765	31.8	56.4	49.8	0.8852
mT5_Base	36.8	62.1	46.7	0.8072	52.7	70.5	32.4	0.8255	27.4	50.7	54.5	0.8109
NLLB-200	44.0*†	68.7*†	41.2*	0.862	66.4*†	80.3*†	22.3*†	**0.8916**	**37.6***†	**61.3***†	**44.7***†	0.8867
Gemini 1.5 Flash	43.7*†	68.7*†	41.2*	0.8657	54.5	73.2*	30.9	0.8755	35.0*†	59.9*	47.4*	0.8929
Text+Triplets-to-Text (TT2T)												
NMT-TT2T	41.3	65.7	43.6	0.8618	60.5	71.3	31.6	0.8779	31.9	56.6	49.7	0.8854
mT5_Base	37.2	62.5	46.0	0.8107	52.7	70.5	32.8	0.8266	27.7	51.1	54.4	0.8167
NLLB-200	44.6*†	69.1*†	40.7*†	0.8626	**67.0***†	**80.5***†	**21.9***†	0.8912	36.9*†	60.7*†	45.5*†	0.8828
Gemini 1.5 Flash	43.9*	68.7*†	40.8*†	0.8688	54.5	73.2	30.6	0.8803	34.5*†	59.2*	48.0	0.8923
Text+Image-to-Text (TI2T)												
MMT-TI2T	42.3*	66.6*	42.1*	0.8672	62.1*	72.6	31.1	0.8786	32.7	58.2*	47.6*	0.8864
Gemini 1.5 Flash	44.1*†	68.7*†	40.3*†	0.868	55.0	73.5*	30.8	0.8738	35.0*†	59.7*	48.4	0.8917
Text+Triplets+Image-to-Text (TTI2T)												
MMT-TTI2T	42.6*	66.8*	41.8*	0.8681	62.2*	72.5	30.9	0.8791	32.9	58.1*	47.8*	0.8862
Gemini 1.5 Flash	**45.1***†	**69.2***†	**40.1***†	**0.8696**	54.6	73.5*	30.4*	0.8767	34.8*†	59.7*	48.3	**0.8964**

Table 1: BLEU, ChrF2, TER and COMET scores for baseline and proposed models for English to German, French and Czech on the 2016 test set ($*$ and † represent a statistically significant results compared to baseline NMT and MMT respectively at a significance level of $p < 0.05$).

the risk of overfitting by penalizing excessively large model weights. We fine-tuned the model for 10 epochs by monitoring the validation loss during training to prevent unnecessary computations and potential overfitting. During training, logging every 500 steps provides periodic updates on performance, ensuring that any issues can be quickly identified and addressed.

4.7 Evaluation Metrics

We use four evaluation metrics: BLEU (Papineni et al., 2002), ChrF2 (Popović, 2015), TER (Snover et al., 2006), and COMET (Rei et al., 2020). BLEU assesses translation precision by comparing candidate translations to reference translations based on *n-grams*. ChrF2 evaluates the similarity between character *n-grams* in machine-generated and reference translations, particularly beneficial for languages with complex writing systems. TER quantifies the number of edits needed to align machine translations with human-generated references. COMET [2] is a neural-based metric that leverages both source and reference sentences to produce quality assessments aligned with human judgments. We conduct statistical significance testing using the *sacrebleu*[3] toolbox.

5 Results

In this section, we present the results of different translation models for language pairs of English into German, French and Czech. The evaluation is based on four metrics: BLEU, ChrF2, TER and COMET. In the first part, we focus on quantitative analysis, and in the second part, we conduct a qualitative analysis to manually evaluate the translation outputs of the models.

5.1 Quantitative Analysis

Table 1 presents the evaluation scores for our proposed multimodal and text-only translation models across English to German, French, and Czech translation tasks for the 2016 test set from the Multi30k dataset. For English to German translation, the Gemini (TTI2T) model achieved the highest scores in BLEU (45.1), ChrF2 (69.2), and COMET (0.8696) while also maintaining the lowest TER (40.1). This indicates that the inclusion of both triples and images in the input significantly enhanced translation quality. The NLLB-200 (TT2T) model closely followed, showing competitive results, particularly in ChrF2 (69.1) and COMET (0.8626). This suggests that leveraging structured data, even without images, is beneficial. Meanwhile, for English to French, the NLLB-200 (TT2T) model outperformed others with the highest BLEU (67.0) and lowest TER (21.9), showcasing its efficiency in maintaining fluency and adequacy. However, Gemini (TTI2T) scored the highest in COMET (0.8767), indicating that it produced the most human-like translations despite slightly lower BLEU. For English to Czech, NLLB-200 (T2T) led in all metrics, except COMET, where Gemini (TI2T) achieved the highest score (0.8929), emphasizing the benefit of incorporating multimodal

[2] https://github.com/Unbabel/COMET
[3] https://github.com/mjpost/sacrebleu

	English → German				English → French			
	BLEU ↑	ChrF2 ↑	TER ↓	COMET ↑	BLEU ↑	ChrF2 ↑	TER ↓	COMET ↑
Text-to-Text (T2T)								
NMT-T2T	35.4	61.7	51.3	0.8548	49.4	68.6	35.8	0.8761
mT5_Base	29.9	57.3	55.8	0.7829	45.3	65.7	38.4	0.8169
NLLB-200	39.4*†	66.5*†	46.4*†	0.8566	**59.9***†	**76.8***†	**26.8***†	**0.8839**
Gemini 1.5 Flash	40.0*†	66.2*†	46.4*†	0.8632	53.1*†	73.2*†	32.0*†	0.8804
Text+Triplets-to-Text (TT2T)								
NMT-TT2T	35.3	61.5	51.6	0.8554	49.5	68.5	36.1	0.8723
mT5_Base	29.8	57.4	55.9	0.7796	45.5	65.7	38.8	0.8134
NLLB-200	38.1*†	65.7*†	48.9*	0.8504	59.5*†	76.4*†	27.9*†	0.8815
Gemini 1.5 Flash	39.8*†	66.2*†	45.8*†	0.863	52.5*	72.7*	32.5*	0.8737
Text+Image-to-Text (TI2T)								
MMT-TI2T	36.8	62.8	49.4	0.8572	51.3	71.5*	33.7	0.8768
Gemini 1.5 Flash	39.9*†	66.3*†	46.2*†	0.8624	54.3*†	73.6*†	31.7*	0.8786
Text+Triplets+Image-to-Text (TTI2T)								
MMT-TTI2T	37.1*	63.3	48.5*	0.8586	51.5	71.4	33.6	0.8781
Gemini 1.5 Flash	**40.6***†	**66.9***†	**45.4***†	**0.865**	53.9*†	73.6*†	31.5*†	0.8814

Table 2: BLEU, ChrF2, TER and COMET scores for baseline and proposed models for English to German and French on the 2017 test set ($*$ and † represent a statistically significant results compared to baseline NMT and MMT respectively at a significance level of $p < 0.05$).

information.

Gemini (TTI2T) consistently achieved top-tier scores, highlighting the advantages of integrating text, triples, and images across all language pairs. The lower BLEU and higher TER for mT5_Base across the board suggest its weaker ability to capture linguistic nuances. Notably, models using additional structured data (TT2T and TI2T) generally performed better than pure text-only models, confirming the effectiveness of multimodal approaches.

Table 2 presents the evaluation scores for our proposed multimodal and text-only translation models across English to German and French translation tasks for the 2016 test set from the Multi30k dataset. For English to German, Gemini (TTI2T) achieved the highest BLEU (40.6), ChrF2 (66.9), and COMET (0.865), along with the lowest TER (45.4). This again confirms the model's ability to leverage triplets and images to improve translation quality. Interestingly, NLLB-200 (T2T) performed best among text-only models, demonstrating its robustness. For English to French, NLLB-200 (T2T) set the highest scores in BLEU (59.9), ChrF2 (76.8), and TER (26.8), suggesting that its architecture excels in handling sentence-level fluency. However, Gemini (TTI2T) achieved the highest COMET (0.8814), implying that its translations were more aligned with human preferences.

Across both language pairs, Gemini (TTI2T) and NLLB-200 (T2T) consistently dominated, with the former benefiting from multimodal inputs and the latter excelling in text-based scenarios. Compared to 2016, TER values increased slightly, indicating a possible complexity shift in the test data. Overall, the performance gaps between text-only and multimodal models further widened, reinforcing the importance of multimodal approaches.

Table 3 presents the evaluation scores for our proposed multimodal and text-only translation models across English to German, French, and Czech translation tasks for the 2016 test set from the Multi30k dataset. For English to German, Gemini (T2T) outperformed all models in BLEU (37.6), TER (49.9), and COMET (0.8519), while Gemini (TI2T) led in ChrF2 (64.0). This suggests that including images provides more lexical coverage, enhancing character-level similarity. In English to French, NLLB-200 (TT2T) obtained the highest BLEU (43.1), while Gemini (TTI2T) dominated COMET (0.8503) and had the lowest TER (40.9), reinforcing the effectiveness of triples-based multimodal training. For English to Czech, NLLB-200 (TT2T) showed the highest BLEU (34.7), but Gemini (TTI2T) again achieved the highest COMET (0.8882), demonstrating improved translation quality with respect to human preferences.

Compared to 2016 and 2017, BLEU scores declined slightly in 2018, suggesting that the 2018 test set was more challenging. However, models incorporating multimodal inputs consistently performed better, emphasizing their enhanced ability to handle complex translation tasks. The consistently strong COMET scores achieved by Gemini

	English → German				English → French				English → Czech			
	BLEU ↑	ChrF2 ↑	TER ↓	COMET ↑	BLEU ↑	ChrF2 ↑	TER ↓	COMET ↑	BLEU ↑	ChrF2 ↑	TER ↓	COMET ↑
Text-to-Text (T2T)												
NMT-T2T	32.4	59.8	54.6	0.8352	38.9	62.7	45.5	0.8418	28.9	52.8	57.4	0.8663
mT5_Base	28.1	55.2	58.9	0.7656	34.1	58.3	48.8	0.778	21.8	46.2	62.6	0.757
NLLB-200	37.3*†	63.5*	50.5*	0.8365	42.8*†	65.7*†	**40.8***†	0.8429	34.4*†	59.2*†	**49.9***†	0.8688
Gemini 1.5 Flash	**37.6***†	63.9*	**49.9***†	**0.8519**	42.3*†	65.6*	41.5*†	0.8475	33.2*†	**59.4***†	51.5*†	0.8877
Text+Triplets-to-Text (TT2T)												
NMT-TT2T	32.2	59.4	54.9	0.8346	39.1	62.8	45.5	0.8407	28.8	52.8	57.2	0.8641
mT5_Base	28.4	55.4	59.2	0.7678	34.3	58.4	48.9	0.7806	22.1	46.5	61.8	0.7628
NLLB-200	37.0*†	63.4*	51.3*	0.8351	**43.1***†	**65.8***†	41.1*†	0.8414	**34.7***†	59.2*†	50.8*†	0.8672
Gemini 1.5 Flash	37.0*†	63.7*	50.2*†	0.85	41.0	64.6*	42.3*	0.844	32.6*	58.5*†	51.8*†	0.8852
Text+Image-to-Text (TI2T)												
MMT-TI2T	33.7	61.2	52.4	0.8364	39.9	63.6	43.8	0.8485	30.1	54.8*	55.4*	0.8687
Gemini 1.5 Flash	37.0*†	**64.0***	50.4*	0.8506	42.4*	65.5*	41.3*	0.8476	33.1*†	58.7*†	52.2*†	0.8851
Text+Triplets+Image-to-Text (TTI2T)												
MMT-TTI2T	33.6	61.3	52.6*	0.8385	40.1	63.4	43.5*	0.847	30.3	54.7*	55.3*	0.8664
Gemini 1.5 Flash	37.2*†	63.3*	50.3*†	**0.8519**	42.6*	65.7*	40.9*†	**0.8503**	32.7*	58.5*†	52.7*†	**0.8882**

Table 3: BLEU, ChrF2, TER and COMET scores for baseline and proposed models for English to German, French and Czech on the 2018 test set ($*$ and † represent a statistically significant results compared to baseline NMT and MMT respectively at a significance level of $p < 0.05$).

(TTI2T) across all language pairs further underline its potential to produce translations that align more closely with human judgments.

Across the three test sets, the best-performing models varied depending on the language pair and evaluation metric. For English to German translation, the Gemini model showed the most significant improvement, particularly in the TTI2T setting. In English to French, the NLLB-200 model consistently outperformed others, especially in T2T translation. For English to Czech, the same model demonstrated strong performance. Overall, the results indicate that incorporating multimodal data, such as images and structured triples, enhances translation quality, with the TTI2T setting often achieving the best performance. These findings suggest that advanced multimodal approaches, particularly leveraging large-scale models like Gemini, can efficiently benefit from multimodal information and significantly improve machine translation across multiple languages and evaluation benchmarks.

5.2 Qualitative Analysis

In this section, we present examples from translation outputs to qualitatively analyse the performance of the models. We calculated sentence-level BLEU scores for each translation model and manually compared the translation quality across all sentences. Figure 4 shows two examples from the 2016 test set of the Multi30K data set: one for English to German and one for English to French translation.

In English to German, Gemini (TTI2T) provides the most accurate translation as it is identical to the reference sentence. This indicates that it perfectly preserves the original sentence's word choice, structure, and meaning. Specifically, it correctly translates "A boy wearing a red shirt" as "Ein Junge in einem roten Shirt", maintaining both the phrasing and natural German expression. Gemini (TI2T) is slightly less accurate but still acceptable. The only difference is the phrase "mit rotem Shirt" instead of "in einem roten Shirt." While both are grammatically correct, "in einem roten Shirt" is the more natural way to describe someone wearing a shirt in German. NLLB-200 (T2T) produces the weakest translation compared to Gemini. It translates "red shirt" as "roten Hemd," where "Hemd" usually refers to a button-down shirt rather than the more general "Shirt" in English. Also, NLLB-200 translates "into the sand" as "in den Sand," slightly altering the meaning. The reference phrase "mit einer gelben Schaufel im Sand" correctly implies that the boy is digging within the sand, while "in den Sand" suggests movement into the sand, making it a less precise translation.

In the English to French example, Gemini (TTI2T) offers a perfect translation, maintaining an exact correspondence with the original text. However, Gemini (TI2T) diverges slightly with two key differences that make it less accurate: first, it replaces "maillot" (jersey) with "chemise" (shirt), which, while understandable, is not the proper term in the context of sportswear, where "maillot" is universally used to describe athletic jerseys. Sec-

English (source): *A boy wearing a red shirt digs into the sand with a yellow shovel.*
German (reference): *Ein Junge in einem roten Shirt gräbt mit einer gelben Schaufel im Sand.*

Gemini (TTI2T): *Ein Junge in einem roten Shirt gräbt mit einer gelben Schaufel im Sand.*
Gemini (TI2T): *Ein Junge mit rotem Shirt gräbt mit einer gelben Schaufel im Sand.*
Gemini (T2T): *Ein Junge mit einem roten Shirt gräbt mit einer gelben Schaufel im Sand.*
NLLB-200(T2T): *Ein Junge in einem roten Hemd gräbt mit einer gelben Schaufel in den Sand.*
MMT(TI2T): *Ein Junge in einem roten Hemd gräbt mit einer gelben Schaufel im Sand.*

English (source): *A baseball player in a black shirt just tagged a player in a white shirt.*
French (reference): *Une joueuse de baseball en maillot noir vient de toucher une joueuse en maillot blanc.*

Gemini (TTI2T): *Une joueuse de baseball en maillot noir vient de toucher une joueuse en maillot blanc.*
Gemini (TI2T): *Une joueuse de baseball en chemise noire vient de mettre un joueur en chemise blanche hors jeu.*
Gemini (T2T): *Un joueur de baseball en maillot noir vient de toucher un joueur en maillot blanc.*
NLLB-200(T2T): *Un joueur de baseball en T-shirt noir vient de taguer un joueur en T-shirt blanc.*
MMT(TI2T): *Un joueur de baseball en T-shirt noir vient de tagué un joueur en T-shirt blanc .*

Figure 4: Examples of translations from English to German (top) and English to French (bottom). Green highlights indicate perfect translations, while yellow marks less accurate translations of the source text.

ond, it translates "just tagged" as "vient de mettre un joueur hors jeu" (just put a player out of play), which, though conveying the general idea, is less precise than the term "toucher" (to tag) in baseball, where the action refers specifically to a player being touched to be considered out. While this translation remains understandable, these differences make it slightly less accurate than Gemini (TTI2T). The NLLB-200 (T2T) translation introduces additional variations, further straying from the original: it changes "joueuse" (female player) to "joueur" (male player), which introduces an assumption about gender that isn't specified in the source text, and although "joueur" could be used in a gender-neutral sense, "joueuse" would be the more appropriate term in a context where the gender is unclear. It also replaces "maillot" with "T-shirt," a term that, while commonly understood, is less specific and appropriate for sportswear, where "maillot" is the established term. Additionally, the NLLB-200 translation opts for the borrowed English term "taguer" instead of "toucher," a choice that might be understandable in informal or colloquial French, but is not the correct terminology in the context of baseball, where "toucher" is the standard.

6 Conclusion

In this paper, we explored the use of Visual Scene Graphs as a structured and interpretable representation of visual information to enhance translation quality. We focused on integrating these representations into translation models by representing visual content in a semantically structured form rather than relying on raw image data. The results demonstrated that incorporating this information into multimodal machine translation models led to significant improvements in both quantitative metrics and qualitative evaluations, highlighting the potential of this approach to advance multimodal translation.

Given the ability of multimodal Large Language Models (LLMs) to extract Visual Scene Graphs in multiple languages, our approach can be applied to improve translation performance across various language pairs. This capability not only broadens the applicability of visual scene graphs but also facilitates the use of multimodal LLMs in handling diverse languages and domains. However, our approach depends on the language coverage of these models, which constitutes a limitation, restricting applicability to the languages supported by multimodal LLMs. In future work, we plan to refine the integration of Visual Scene Graphs and explore additional language pairs to further validate and extend the applicability of our approach across translation directions.

Acknowledgments

This publication has emanated from research conducted with the financial support of Research Ireland under Grant Number 12/RC/2289_P2 - Insight Research Ireland Centre for Data Analytics. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

References

Braeden Bowen, Vipin Vijayan, Scott Grigsby, Timothy Anderson, and Jeremy Gwinnup. 2024. Detecting concrete visual tokens for multimodal machine translation. In *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 29–38, Chicago, USA. Association for Machine Translation in the Americas.

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, Volume 1 (Long and Short Papers)*, pages 4159–4170, Minneapolis, Minnesota. 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 (Volume 1: Long Papers)*, pages 1913–1924, Vancouver, Canada. Association for Computational Linguistics.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar. Association for Computational Linguistics.

Jean-Benoit Delbrouck and Stéphane Dupont. 2017. An empirical study on the effectiveness of images in multimodal neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 910–919, Copenhagen, Denmark. Association for Computational Linguistics.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

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, Brussels, Belgium. Association for Computational Linguistics.

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

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

Hao Fei, Qian Liu, Meishan Zhang, Min Zhang, and Tat-Seng Chua. 2023. Scene graph as pivoting: Inference-time image-free unsupervised multimodal machine translation with visual scene hallucination. In *In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5980–5994*.

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

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

Ali Hatami, Paul Buitelaar, and Mihael Arcan. 2023. A filtering approach to object region detection in multimodal machine translation. In *Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track*, pages 393–405, Macau SAR, China. Asia-Pacific Association for Machine Translation.

Marcel Hildebrandt, Hang Li, Rajat Koner, Volker Tresp, and Stephan Günnemann. 2020. Scene graph reasoning for visual question answering. *Computing Research Repository (CoRR)*, abs/2007.01072.

Justin Johnson, Agrim Gupta, and Li Fei-Fei. 2018. Image generation from scene graphs. *Computing Research Repository (CoRR)*, abs/1804.01622.

Guillaume Klein, Yoon Kim, Yuntian Deng, Vincent Nguyen, Jean Senellart, and Alexander Rush. 2018. OpenNMT: Neural machine translation toolkit. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 177–184, Boston, MA. Association for Machine Translation in the Americas.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Chiraag Lala and Lucia Specia. 2018. Multimodal lexical translation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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

Diego Moussallem, Axel-Cyrille Ngonga Ngomo, Paul Buitelaar, and Mihael Arcan. 2019. Utilizing knowledge graphs for neural machine translation augmentation. In *Proceedings of the 10th International Conference on Knowledge Capture*, K-CAP '19, page 139–146, New York, NY, USA. Association for Computing Machinery.

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

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer (t5). *Journal of Machine Learning Research*, 21(140):1–67.

Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. *Computing Research Repository (CoRR)*, abs/2009.09025.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Computing Research Repository (CoRR)*, abs/1706.03762.

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.

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. *Computing Research Repository (CoRR)*, abs/2105.14462.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2021)*.

Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. 2018. Auto-encoding scene graphs for image captioning. *Computing Research Repository (CoRR)*, abs/1812.02378.

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

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 *Annual Meeting of the Association for Computational Linguistics*.

Yongjing Yin, Jiali Zeng, Jinsong Su, Chulun Zhou, Fandong Meng, Jie Zhou, Degen Huang, and Jiebo Luo. 2023. Multi-modal graph contrastive encoding for neural machine translation. *Artificial Intelligence*, 323:103986.

Yuting Zhao, Mamoru Komachi, Tomoyuki Kajiwara, and Chenhui Chu. 2022. Region-attentive multimodal neural machine translation. *Neurocomputing*, 476:1–13.