Enhancing Translation Quality by Leveraging Semantic Diversity in Multimodal Machine Translation

Ali Hatami ali.hatami@insight-centre.org
Insight SFI Research Centre for Data Analytics,
Data Science Institute, University of Galway, Ireland

Mihael Arcan mihael@luahealth.io
Lua Health, Galway, Ireland

Paul Buitelaar paul.buitelaar@insight-centre.org
Insight SFI Research Centre for Data Analytics,
Data Science Institute, University of Galway, Ireland

Abstract

Despite advancements in neural machine translation, word sense disambiguation remains challenging, particularly with limited textual context. Multimodal Machine Translation enhances text-only models by integrating visual information, but its impact varies across translations. This study focuses on ambiguous sentences to investigate the effectiveness of utilizing visual information. By prioritizing these sentences, which benefit from visual cues, we aim to enhance hybrid multimodal and text-only translation approaches. We utilize Latent Semantic Analysis and Sentence-BERT to extract context vectors from the British National Corpus, enabling the assessment of semantic diversity. Our approach enhances translation quality for English-German and English-French on Multi30k, assessed through metrics including BLEU, chrF2, and TER.

1 Introduction

Neural Machine Translation (NMT) has significantly improved translation quality with transformer-based models (Cho et al., 2014; Vaswani et al., 2018), integrating cross-attention for better semantic understanding (Vaswani et al., 2017). Despite focusing on the broader context in the text-only translation model, resolving word ambiguity persists as a challenge. In natural language, lexical ambiguity (Gonzales et al., 2017) refers to the occurrence where a single word possesses multiple meanings or interpretations, thereby complicating comprehension of the text. For example, in the domain of finance and economy, the word ”*bank*” almost always refers to a financial institution rather than the side of a river.

Multimodal Machine Translation (MMT), a subset of NMT, incorporates visual information to enhance translations. Recent studies highlight the potential of leveraging both textual and visual data to improve accuracy and contextuality (Yao and Wan, 2020; Zhao et al., 2022; Wang and Xiong, 2021; Hatami et al., 2023). MMT utilises visual cues to disambiguate input words and select appropriate translations, particularly beneficial for ambiguous sentences or when visual context provides crucial details not explicit in the text. Despite the benefits of integrating visual information into MMT, this can sometimes result in degraded translation quality, particularly when there is insufficient data, including parallel visual and textual data, to adequately train the model. For sentences with unambiguous interpretations, textual context alone might suffice for accurate translation. Unlike NMT, MMT can be susceptible to noise or irrelevant information in the visual data, which may introduce errors or distractions, leading to inaccurate translations.

This paper aims to explore the correlation be-

Proceedings of the 16th Conference of the Association for Machine Translation in the Americas,
Chicago, USA, September 30 - October 2, 2024. Volume 1: Research Papers

tween sentence ambiguity and translation quality, focusing on effectively integrating visual cues into the translation process to enhance overall quality. We assess sentence ambiguity using semantic diversity in Latent Semantic Analysis (LSA) and Sentence-BERT (S-BERT) vector embedding spaces, investigating the impact of visual information across varying levels of ambiguity. By experimenting with different ambiguity scores, we determine the optimal value where visual cues enhance translation quality, comparing outcomes with text-only and multimodal models. For sentences with low ambiguity, we employ a text-only approach, while for those with higher ambiguity, we utilize a multimodal approach.

2 Related Work

Lexical ambiguity presents a major hurdle in machine translation, making it challenging to discern the correct word meaning and translation due to multiple senses and contextual variations. While Multimodal Machine Translation (MMT) leverages visual cues to aid disambiguation, the efficacy of visual features varies, particularly when textual context is sufficient. Despite the potential of visual cues to improve accuracy, their impact may be constrained when textual information is already rich. This underscores the importance of seamlessly integrating visual and textual data for optimal translation outcomes (Caglayan et al., 2016, 2019).

Various methodologies have been proposed to enhance the quality of the visual modality in MMT. For example, Yao and Wan (2020) introduced a multimodal transformer-based self-attention mechanism to encode relevant image information. To capture diverse relationships, Yin et al. (2020) proposed a graph-based multimodal fusion encoder. Ive et al. (2019) devised a translate-and-refine mechanism, employing images in a second-stage decoder to refine text-only NMT models for ambiguous words. Additionally, Calixto et al. (2019) utilised a latent variable model to extract multimodal relationships between images and text. Recent methods aim to mitigate visual information noise and select relevant visual features correlated with text. For instance, Wang and Xiong (2021) employed object-level visual modeling to mask irrelevant objects and specific words in the source text, facilitating visual feature analysis. Similarly, Zhao et al. (2022) integrated object detection into the image encoder to extract visual features of object regions and applied them to a doubly-attentive decoder model.

The Multimodal Lexical Translation (MLT) approach aims to accurately translate ambiguous words within both visual and textual contexts. Introduced with the MLT dataset, which includes 4-tuples of ambiguous words, visual and textual contexts, and translations aligned with both, this resource facilitates the evaluation of lexical disambiguation within Multimodal Machine Translation (MMT) (Lala and Specia, 2018). The study by Lala et al. (2018) examines the effectiveness of multimodal re-ranking methods in improving a standard sequence-to-sequence attention-based Neural Machine Translation (NMT) system. By integrating cross-lingual word sense disambiguation and data augmentation techniques, the authors aim to enhance translation quality and develop an image-based, cross-lingual approach for accurately predicting translation candidates for ambiguous words in the source sentence.

The translate-and-refine approach (Ive et al., 2019), introduced to improve upon previous MMT model, employs images in a second-stage decoder to refine translation drafts by incorporating both textual and visual contexts. This method achieves state-of-the-art results, demonstrating superior performance over text-only models, especially in complex linguistic scenarios, by refining translations only when necessary through deliberation networks. In their analysis, Tang et al. (2018) examine how encoder-decoder attention mechanisms in Neural Machine Translation (NMT) models handle ambiguous nouns during word sense disambiguation (WSD). Contrary to expectations, attention tends to focus more on the ambiguous noun itself rather than surrounding context tokens, suggesting that contextual information for WSD is primarily encoded in the encoder's hidden states. This study sheds light on the challenges of WSD in NMT models, particularly due to data sparsity, and offers insights into the learning process of attention mechanisms in Transformers.

In addressing ambiguity in Multimodal Machine Translation (MMT), Futeral et al. (2023) propose a novel approach incorporating neural adapters, guided self-attention mechanisms, and a visually conditioned masked language modeling objective.

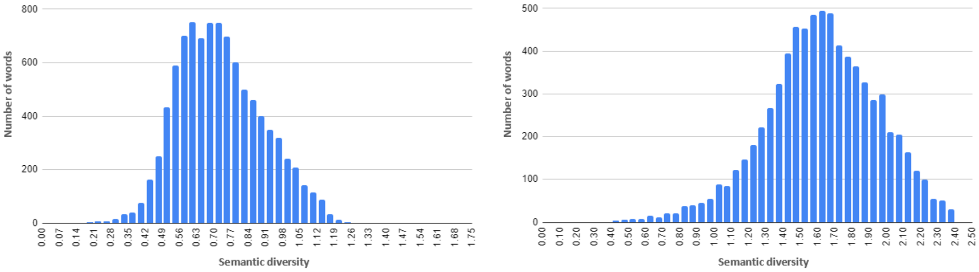


Figure 1: Histogram showing the distribution of lexical ambiguity scores across words in BNC (left: S-BERT and right: LSA)

Their study underscores the importance of using image context to improve translation quality, introducing the CoMMuTE dataset as a tool to evaluate and enhance multimodal translation. The dataset includes 155 English sentences with two possible translations in French, German, and Czech, facilitating assessment of MMT models in leveraging visuals for accurate translations, especially with ambiguous content. In Bowen et al. (2024), techniques for identifying visually and contextually relevant tokens in Multimodal Machine Translation (MMT) systems are explored, employing natural language processing (NLP), object detection, and deterministic selection strategies. The study, conducted using the GRAM MMT architecture (Vijayan et al., 2024), reveals performance improvements over baseline models by training on synthetically collated datasets of masked sentences and images, emphasizing the importance of visual context in enhancing translation accuracy within MMT systems. In Hatami et al. (2022), an approach utilizing *WordNet* synsets to gauge sentence ambiguity was proposed to evaluate the effect of incorporating visual information in translation models, demonstrating the potential of visual cues to improve translation accuracy, especially in challenging tasks like English-German translation, as observed in the analysis of the Multi30k dataset.

This paper investigates how integrating visual elements affects translation quality by examining the relationship between sentence ambiguity and accuracy, using semantic diversity in sentence vector spaces to quantify ambiguity and assessing the impact of visual information on translation quality across different levels of ambiguity scores.

3 Methodology

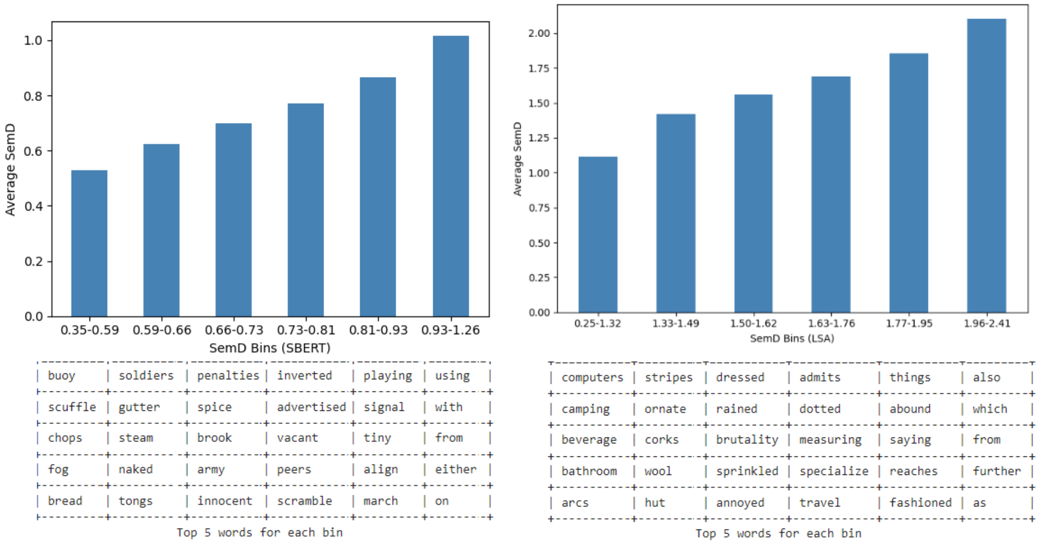
This section details the methodology for enhancing translation quality in MMT by utilizing semantic diversity. It involves computing lexical ambiguity scores for nouns, extending to sentence-level ambiguity, and exploring sentence ambiguity to optimize translation scores for text-only and MMT models.

3.1 Lexical Ambiguity Score

We computed the lexical ambiguity score for all words in the British National Corpus (BNC) by tokenizing sentences from the Multi30k dataset training set, resulting in a word list with 10,105 unique words, including morphological variants to capture potential differences in ambiguity scores based on their roles in sentences. Utilizing Latent Semantic Analysis (Landauer and Dumais, 1997) and S-BERT (Reimers and Gurevych, 2019), we derived lexical ambiguity scores based on distributional semantics (Harris, 1954), which infer word meanings from contextual usage, considering that words appearing in the same context likely share the same meaning, while differing contexts may lead to varied interpretations.

In adopting LSA based on Hoffman's work (Hoffman et al., 2012), we segmented the BNC corpus into 1,000-word texts to construct a co-occurrence matrix, applying singular value decomposition (SVD) to reveal latent semantic structures and word associations. Concurrently, for S-BERT, we segmented the BNC into sentences, preprocess them for quality, and utilized the pre-trained S-BERT model[1] to generate 768-dimensional sentence embeddings, capturing semantic information comprehensively. To compute lexical ambiguity, we as-

[1] https://www.sbert.net/



buoy	soldiers	penalties	inverted	playing	using
scuffle	gutter	spice	advertised	signal	with
chops	steam	brook	vacant	tiny	from
fog	naked	army	peers	align	either
bread	tongs	innocent	scramble	march	on

Top 5 words for each bin

computers	stripes	dressed	admits	things	also
camping	ornate	rained	dotted	abound	which
beverage	corks	brutality	measuring	saying	from
bathroom	wool	sprinkled	specialize	reaches	further
arcs	hut	annoyed	travel	fashioned	as

Top 5 words for each bin

Figure 2: Average Semantic Diversity (SemD) across different bins, displaying the top 5 words in each bin from BNC (left: S-BERT and right: LSA)

sessed textual similarity through LSA and S-BERT, leveraging Semantic Diversity (*SemD*) scores to represent ambiguity. By measuring cosine similarity between vectors, we determined ambiguity levels, with higher similarity indicating lower ambiguity and vice versa, enabling precise ambiguity scoring for individual words. To do this, we first calculated the mean of the similarity of all pairwise combinations of texts or sentences including the word (*w*). Then we took the logarithm of this mean and reversed the sign to obtain the *SemD* value of the word (*w*). The equation for *SemD* of word w is:

$$SemD_w = -log(\frac{\sum_{i,j \; \epsilon \; V_w} cos\_sim(v_i, v_j)}{n})$$

where V_w is the set of all context vectors for word *w*, and $v_i, v_j \; \epsilon \; V_w$.

The histograms in Figure 1 compare the distribution of words across different ambiguity score ranges for LSA and S-BERT. S-BERT shows a positively skewed distribution, with most words having lower ambiguity scores, while LSA displays a negatively skewed distribution, indicating a higher prevalence of words with higher ambiguity scores.

Figure 2 illustrates the average SemD across 6 bins, showcasing the lexical ambiguity scores along with the top 5 words in each bin. These results, derived from the BNC, compare S-BERT and LSA in assigning SemD to each word.

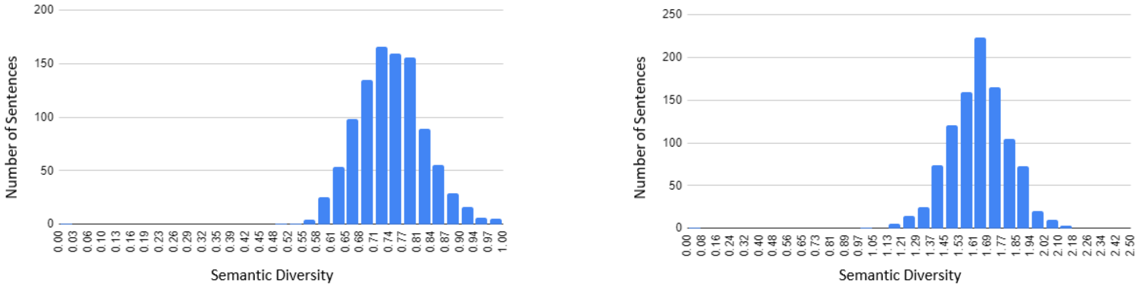
3.2 Sentence Ambiguity Score

After computing *SemD* values for all words in the vocabulary, we utilize these values to derive ambiguity scores for sentences in the test set, focusing solely on nouns, which carry specific semantic content and are extracted using *SpaCy*[2].

To compute the ambiguity score at the sentence level, two mathematical functions, the arithmetic mean (*Mean*) and the geometric mean (*G-Mean*), are utilized. The arithmetic mean aggregates and divides the lexical ambiguity scores of all nouns in a sentence by the total number of content words, giving equal weight to each score, while the geometric mean calculates the *n-th* root of the product of lexical ambiguity scores, assigning less weight to larger values and mitigating the influence of outliers. These methods enable the quantification of ambiguity within sentences, facilitating comparisons based on their ambiguity scores.

The histograms in Figure 3 display sentence ambiguity scores calculated using *Mean* and *G-Mean* for LSA and S-BERT. LSA exhibits a normal distribution of scores between 1.13 and 2.18, while S-BERT shows a positively skewed distribution be-

[2] https://spacy.io/usage/linguistic-features



(a) S-BERT (G-Mean)

(b) LSA (G-Mean)

Figure 3: Histograms showing the distribution of sentence ambiguity scores across sentences in Multi30k

tween 0.55 and 1.00, indicating lower ambiguity. These ambiguity scores are used to sort sentences in the test set in ascending order and then apply a hybrid approach to translate the sentences. In Section 3.3, we explain the details of this approach.

3.3 Translation Quality Measure

Despite the benefits of incorporating visual data into multimodal machine translation (MMT), its use can sometimes lead to reduced translation quality compared to text-only approaches. This decline may occur due to the presence of noise or irrelevant visual information, which could introduce errors or distractions, ultimately resulting in inaccurate translations (see Figure 4).

We utilize sentence ambiguity scores based on *SemD* to decide between using Text-only or Multimodal models for translation. By adopting a hybrid approach, we determine whether visual information enhances translation quality, leveraging the ambiguity score to select the most suitable model for sentences in a specific ambiguity range. After computing ambiguity scores for all sentences in the test set, we ranked the sentences based on the sentence ambiguity score calculated using *Mean* and *G-Mean* for both LSA and S-BERT. Then we divided the test set into 20 sets, each including 50 sentences. The first set in the sorted sentence list has the lowest ambiguity score, and the last set has the highest ambiguity score. The hybrid approach aims to employ the Text-only MT model for sentence sets with lower ambiguity and utilize Multimodal models for those with higher ambiguity. By using a Hybrid model, we explore the effectiveness of visual information in translating sentences with higher ambiguity scores, thereby evaluating translation quality to determine the optimal range of ambiguity for leveraging visual information.

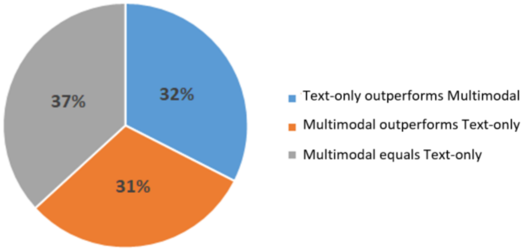


Figure 4: Comparing sentence-level BLEU scores of Text-only and Multimodal MT models for English to German translation on the Multi30k 2016 test set.

4 Experimental Setup

This section provides insights into the dataset used in this work, neural architectures involving text-only and multimodal models, and context vector embedding methods: LSA and S-BERT, and the translation evaluation metrics BLEU, ChrF2 and TER.

4.1 Dataset

During our experiment, we employed two datasets: the British National Corpus (BNC) and Multi30k. The BNC facilitated the extraction of sentence vectors for computing lexical ambiguity, while the Multi30k dataset served for training and evaluating our translation models.

4.1.1 British National Corpus (BNC)

The British National Corpus (BNC) (Aston and Burnard, 1998) is a vast collection comprising 100 million words of both written and spoken British English texts, designed to represent the language comprehensively. It encompasses diverse sources such as newspapers, periodicals, academic books, fiction, letters, and spoken conversations, offering insights into contemporary British English usage across various contexts. As a rich repository of language patterns and expressions, the BNC serves as a valuable resource for linguistic research and language analysis, facilitating a deeper understanding of British English in its diverse forms.

4.1.2 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 provides validation and test sets, each containing 1,000 images aligned with the descriptions.

4.2 Semantic Representation Techniques

In this section, we explain two techniques for extracting semantic representation vectors from given contexts: Latent Semantic Analysis and S-BERT. These approaches offer sophisticated methods to capture the underlying semantic meanings embedded within texts, which can be used to determine the similarity between them.

4.2.1 Latent Semantic Analysis (LSA)

LSA is a technique that leverages patterns of word co-occurrence to construct high-dimensional semantic spaces. To implement LSA, the BNC is divided into text samples, each representing a different context. A co-occurrence matrix is generated, tracking which words appear in each context. Each word is represented as a vector, with elements corresponding to its frequency in a context. Using singular value decomposition (SVD), LSA extracts the underlying structure in the co-occurrence matrix, revealing higher-order relationships between words based on their co-occurrence patterns. SVD reduces the dimensionality of the word vectors (to 300 dimensions), with the similarity structure of these vectors approximating the original matrix. Consequently, word representations can be interpreted as points in a high-dimensional space, where proximity indicates similarity in meaning based on context. Additionally, LSA places individual contexts in the same semantic space, enabling comparisons between contexts based on their content similarity.

4.2.2 S-BERT

S-BERT extends the capabilities of the BERT model by focusing on generating high-quality sentence embeddings. Unlike traditional BERT models, which are primarily trained on word-level tasks like next-sentence prediction and masked language modeling, S-BERT fine-tunes the BERT architecture to produce embeddings at the sentence level. It was trained on a combination of two Natural Language Inference (NLI) datasets: the Stanford NLI (SNLI) dataset and the Multi-Genre NLI (MultiNLI) dataset. S-BERT typically employs Siamese or triplet network architectures during fine-tuning, enabling it to capture contextual information and nuances in meaning. By considering the surrounding context, S-BERT generates embeddings that are suitable for tasks such as semantic textual similarity. Cosine similarity between associated sentence vectors indicates the similarity between word meanings in different sentences, with higher similarity indicating lower ambiguity in word meaning.

4.3 Neural Machine Translation

4.3.1 Text-only Machine Translation

A text-only transformer model serves as the baseline in our experiment, utilizing 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.

4.3.2 Multimodal Machine Translation

In the MMT model, we adopt the Gated Fusion MMT model Wu et al. (2021), which fuses visual and text representations by employing a gate mechanism. Gated Fusion is a mechanism used to integrate visual information from images with textual

En → De	Test 2016						Test 2017					
	LSA			S-BERT			LSA			S-BERT		
	BLEU ↑	chrF2 ↑	TER ↓	BLEU ↑	chrF2 ↑	TER ↓	BLEU ↑	chrF2 ↑	TER ↓	BLEU ↑	chrF2 ↑	TER ↓
Baseline (MMT)	40.1	64.6	40.6	40.1	64.6	40.6	31.9	59.8	49.6	31.9	59.8	49.6
Hybrid (50)	40.7	**65.0**	39.9	40.7	**65.0**	39.8*	**32.6**	**60.6***	**48.5***	**32.6**	60.5*	**48.5***
Hybrid (100)	40.8	**65.0**	**39.7***	**40.9**	**65.0**	**39.7***	32.5	60.4*	48.6*	**32.6**	60.5*	48.6*
Hybrid (150)	40.8	**65.0***	39.9*	**40.9**	**65.0**	39.8*	32.4	60.3	48.8*	32.4	60.4*	48.8*
Hybrid (200)	40.6	**65.0**	40.1	40.7	64.9	39.9	32.4	60.3	48.6*	32.3	60.4*	48.8*
Hybrid (250)	40.6	64.9	40.2	40.5	64.9	40.1	32.2	60.2	48.8*	32.1	60.3	49.0
Hybrid (300)	40.6	**65.0**	40.2	40.5	64.8	40.1	32.0	60.1	49.0*	32.1	60.1	49.0
Hybrid (350)	40.5	64.8	40.3	40.4	64.7	40.1	32.1	60.1	48.9*	32.0	59.9	49.1
Hybrid (400)	40.5	64.8	40.2	40.5	64.8	40.0	32.3	60.1	48.8*	31.9	59.9	49.1
Hybrid (450)	40.4	64.7	40.3	40.5	64.8	40.0	32.1	59.9	49.0*	32.0	60.0	49.0*
Hybrid (500)	40.4	64.7	40.3	40.5	64.7	40.2	32.2	60.0	49.0*	32.1	60.0	49.0*
Hybrid (550)	40.3	64.7	40.4	40.4	64.8	40.1	32.2	59.9	49.0*	32.0	59.9	49.1*
Hybrid (600)	40.2	64.6	40.5	40.4	64.7	40.1	32.1	59.8	49.2	31.9	59.9	49.2
Hybrid (650)	40.3	64.7	40.4	40.3	64.6	40.1*	32.0	59.8	49.2	32.2	59.9	49.1*
Hybrid (700)	40.3	64.7	40.3	40.2	64.5	40.4	32.2	59.8	49.3	32.1	59.8	49.2*
Hybrid (750)	40.3	64.7	40.3	40.2	64.6	40.3	32.3	59.9	49.3	32.2	59.9	49.3
Hybrid (800)	40.3	64.7	40.4	40.2	64.7	40.3	32.3	59.9	49.3	32.2	59.8	49.4
Hybrid (850)	40.1	64.6	40.4	40.3	64.7	40.2*	32.2	59.9	49.5	32.4*	59.9	49.4
Hybrid (900)	40.1	64.5	40.5	40.3	64.7	40.2*	32.1	59.9	49.6	32.2*	59.9	49.5
Hybrid (950)	40.2	64.7	40.5	40.2	64.7	40.4	31.9	59.8	49.7	32.0	59.8	49.6

En → Fr	Test 2016						Test 2017					
	LSA			S-BERT			LSA			S-BERT		
	BLEU ↑	chrF2 ↑	TER ↓	BLEU ↑	chrF2 ↑	TER ↓	BLEU ↑	chrF2 ↑	TER ↓	BLEU ↑	chrF2 ↑	TER ↓
Baseline (MMT)	**62.3**	75.3	25.5	**62.3**	75.3	25.5	55.6	70.7	30.9	55.6	70.7	30.9
Hybrid (50)	62.1	75.4	25.2	62.2	**75.5**	**25.1**	55.9	71.0	30.8	55.9	**71.1**	30.8
Hybrid (100)	62.0	75.3	25.3	62.2	**75.5**	25.2	55.9	**71.1**	30.7	**56.0**	**71.1**	30.7
Hybrid (150)	61.8	75.2	25.4	62.1	75.4	25.3	55.9	71.0	30.7	55.8	71.0	30.9
Hybrid (200)	61.7	75.2	25.3	62.0	75.3	25.4	55.9	**71.1**	**30.6**	55.9	**71.1***	30.7
Hybrid (250)	61.8	75.2	25.4	61.9	75.2	25.4	55.7	70.9	30.7	55.9	**71.1***	30.7
Hybrid (300)	61.8	75.2	25.4	61.9	75.2	25.5	55.5	70.9	30.8	55.7	70.9	30.7
Hybrid (350)	61.7*	75.1	25.6	61.7	75.1	25.5	55.5	70.8	30.9	55.8	71.0	30.7
Hybrid (400)	61.7*	75.1	25.6	61.6*	75.0	25.7	55.6	70.8	30.8	55.9	**71.1***	30.7
Hybrid (450)	61.7*	75.1	25.6	61.7*	75.0	25.8	55.5	70.8	30.8	55.9	71.0	30.7
Hybrid (500)	61.6*	75.1	25.7	61.7*	75.0	25.7	55.5	70.7	31.0	55.8	70.9	30.7
Hybrid (550)	61.8*	75.1	25.6	61.8*	75.0	25.8	55.3	70.6	31.0	55.7	70.8	30.9
Hybrid (600)	61.9	75.1	25.6	61.9	75.1	25.7	55.3	70.6	31.0	55.5	70.7	31.0
Hybrid (650)	62.0	75.2	25.6	62.0	75.2	25.5	55.5	70.6	31.0	55.5	70.7	31.0
Hybrid (700)	62.1	75.3	25.5	62.1	75.2	25.5	55.4	70.6	31.1	55.5	70.7	31.1
Hybrid (750)	62.0	75.2	25.7	62.0	75.2	25.6	55.5	70.7	31.0	55.4	70.7	31.1
Hybrid (800)	62.0*	75.1*	25.7	62.1	75.2	25.6	55.5	70.6	31.0	55.3	70.6	31.1
Hybrid (850)	62.0*	75.1*	25.6	62.1	75.2	25.6	55.6	70.7	31.0	55.5	70.7	31.1
Hybrid (900)	62.1	75.2	25.6	62.2	75.2	25.6	55.6	70.7	30.9	55.6	70.7	31.0
Hybrid (950)	**62.3**	75.3	25.5	62.1	75.2*	25.7*	55.6	70.7	31.0	55.6	70.7	31.0

Table 1: BLEU, chrF2, and TER scores for baseline and Hybrid models for English-to-German and English-to-French translations. Numbers in parentheses show sentences where the model uses visual information (e.g., Hybrid (50) refers to the top 50 ambiguous sentences using Multimodal, while the remaining 950 use a text-only model). * indicates a statistically significant result compared to the baseline multimodal at a significance level of $p < 0.05$. Bold numbers indicate the best results in each test dataset for each score.

information from source sentences during the translation process. 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 (Russakovsky et al., 2015) 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 of dimension $T{\times}d$ is generated to control the fusion of the textual and visual representations. The gating matrix is computed as:

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

where W and U are model parameters.

4.4 Evaluation Metrics

We use three evaluation metrics: BLEU (Papineni et al., 2002), ChrF2 (Popović, 2015), and TER (Snover et al., 2006). 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. We conduct statistical significance testing using the *sacrebleu*[3] toolbox.

5 Results

In this section, we analyze the results of our experiments. We present the findings for both LSA and S-BERT approaches on the 2016 and 2017 Multi30k test sets for English to German and English to French translations. Table 1 provides a comprehensive comparison of different models' performance in terms of BLEU, chrF2, and TER metrics, offering insights into the effectiveness of integrating sentence ambiguity scores with a multimodal setting in English to German and French translations. We report translation scores for the baseline multimodal and the Hybrid models for LSA and S-BERT using *G-Mean*[4].

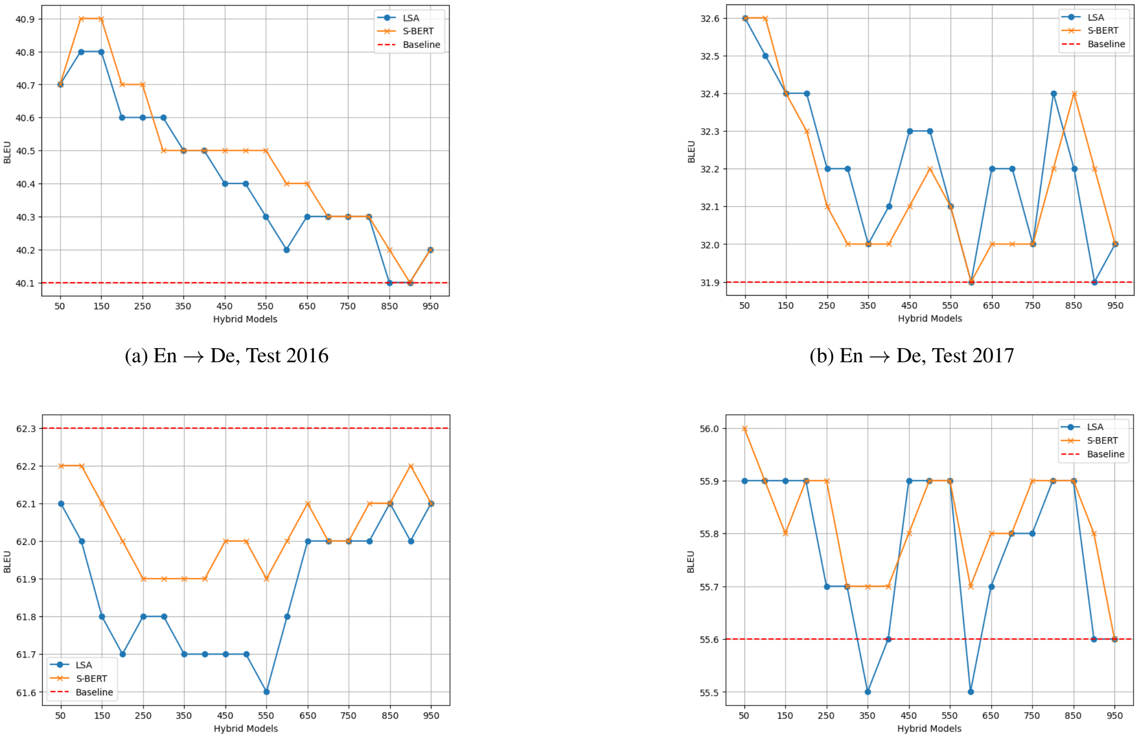
Table 1 presents the translation performance of baseline and Hybrid models for English-to-German (En → De) and English-to-French (En → Fr) across different test datasets and Hybrid configurations. The table highlights metrics including BLEU score (higher is better), chrF2 score (higher is better), and TER score (lower is better). Each Hybrid model variant is indicated by the number of sentences (in parentheses) where visual information aids translation, with the remainder utilizing a text-only model[5]. Statistically significant improvements over the baseline multimodal model at $p<0.05$ are marked with *, while the best-performing scores in each dataset are indicated in bold.

The results indicate that the sentence ambiguity score plays an important role in determining the importance of using visual information in English-to-German translation compared with English-to-French translation. In English-to-German translation for Test 2016, the baseline multimodal model achieves a performance with a BLEU score of 40.1, a chrF2 score of 64.6, and a TER of 40.6 using LSA and S-BERT. In contrast, the Hybrid models show improvements over the baseline. In LSA, Hybrid (50) achieves a BLEU score of 40.7, a chrF2 score of 65.0, and a TER of 39.9. Hybrid (100) and Hybrid (150) continue to outperform the baseline across all metrics. In S-BERT, similar to LSA, Hybrid (100) and Hybrid (150) achieved a BLEU score of 40.9, a chrF2 score of 65.0, and notably reduced the TER to 39.7. For both LSA and S-BERT, Hybrid (50) to Hybrid (150) achieve statistically significant improvements in chrF2 and TER in some configurations. By increasing the number of sentences that the Hybrid model uses visual information for, the results get close to the baseline multimodal model (see Hybrid (950)). For Test 2017, the performance of the Hybrid models remains consistent with Test 2016, indicating stability in the proposed approach for English-to-German translation. For this test set, Hybrid (50) maintains improvements over the baseline with a BLEU score of 32.6, a chrF2 score of 60.6, and a TER of 48.5, representing a statisti-

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

[4] Based on our experiments, *G-Mean* shows better results compared with *Mean*. Therefore, due to limited space, we reported the results for *G-Mean*.

[5] The total number of sentences in the test sets is 1,000.



(a) En → De, Test 2016

(b) En → De, Test 2017

(c) En → Fr, Test 2016

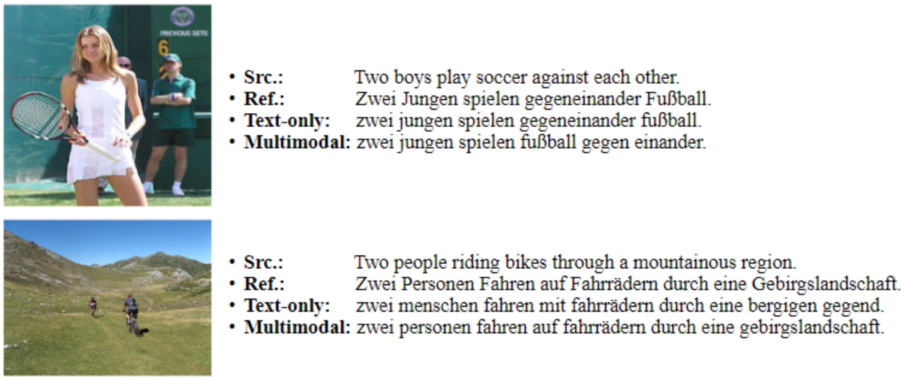
(d) En → Fr, Test 2017

Figure 5: The charts display BLEU scores across various Hybrid models in English to German and French for the 2016 and 2017 test sets. Solid lines represent BLEU scores for S-BERT and LSA, while dashed lines indicate the overall performance for multimodal MT models.

cally significant improvement over the baseline. Hybrid (100) and Hybrid (150) consistently outperform the baseline, with results showing statistical significance.

In English-to-French translation, the Hybrid models show slight improvement over the baseline multimodal model. In Test 2016, the baseline model has a higher BLEU score compared with the Hybrid models. The Hybrid model of 50 slightly improves the chrF2 and TER scores, but they are not statistically significant. Similar to Test 2016, Test 2017 does not represent notable improvements regarding BLEU, chrF2, and TER scores. This indicates that the idea of using ambiguity scores to evaluate the importance of using visual information is less effective for English-to-French translation.

To better analyze the role of a sentence ambiguity score in the proposed Hybrid models, the BLEU scores for LSA and S-BERT for *G-Mean* are presented in Figure 5. In each subgraph, the red dashed line shows the overall BLEU scores for the baseline multimodal model for each language pair in the 2016 and 2017 test sets. The orange and blue lines show the BLEU scores in different Hybrid models. For both language combinations, LSA and S-BERT follow the same pattern. In English-to-German translation, by increasing the number of sentences in the Hybrid model, the BLEU scores started from 40.9 and 32.6 for Test 2016 and Test 2017, respectively, and reached the baseline multimodal models. This indicates that visual information is useful in translating around 150 sentences with higher ambiguity scores. However, using visual information for the remaining sentences with lower ambiguity ranking sharply drops translation performance. In contrast, for English-to-French translation, we do not see the same pattern. In Test 2016, all Hybrid models have BLEU scores lower than the baseline multimodal model, showing the effectiveness of using visual information in most sentences. In Test



- **Src.:** Two boys play soccer against each other.
- **Ref.:** Zwei Jungen spielen gegeneinander Fußball.
- **Text-only:** zwei jungen spielen gegeneinander fußball.
- **Multimodal:** zwei jungen spielen fußball gegen einander.

- **Src.:** Two people riding bikes through a mountainous region.
- **Ref.:** Zwei Personen Fahren auf Fahrrädern durch eine Gebirgslandschaft.
- **Text-only:** zwei menschen fahren mit fahrrädern durch eine bergigen gegend.
- **Multimodal:** zwei personen fahren auf fahrrädern durch eine gebirgslandschaft.

Figure 6: Examples from Multi30k illustrate the effectiveness of using images based on the ambiguity level of the source sentence. The top image shows a source sentence with a low ambiguity score (1.46), which was translated more accurately using the Text-only model. The bottom image shows a source sentence with a high ambiguity score (1.81), where the Multimodal model provided a better translation.

2017, there are consistent fluctuations by changing the number of sentences, but it remains above the baseline model except in a few cases.

Figure 6 shows examples from the Multi30k dataset to illustrate the impact of sentence ambiguity on the effectiveness of translation models. The top image presents a source sentence with a low ambiguity score of 1.46, where the Text-only model outperformed the Multimodal model according to automatic evaluation metrics like the BLEU score. However, interestingly, human analysis revealed that the translation provided by the Multimodal model not only better explained the image but was also more readable than even the reference sentence. Conversely, the bottom image presents a source sentence with a higher ambiguity score of 1.81, where the Multimodal model produced a superior translation compared to the Text-only model. While various factors can influence the performance of multimodal translation models, these findings suggest that the sentence ambiguity score can serve as a valuable parameter in determining when visual information enhances translation quality.

6 Conclusion

This study contributes to the ongoing discussion on the effective utilisation of visual cues in translation tasks and provides insights into optimizing multimodal translation systems. In this paper, we investigate the impact of integrating visual elements into the translation process on overall translation quality. Through an analysis of the relationship between sentence ambiguity and translation quality, we aimed to determine the circumstances under which visual information enhances translation quality. By establishing ambiguity scores for individual sentences using semantic diversity within sentence vector embedding spaces, we investigated how visual information influences translation quality across different ranges of sentence ambiguity scores. Our research highlights the importance of discerning the contextual relevance of visual information in multimodal tasks, suggesting semantic diversity as a valuable metric for determining the significance of visual cues in multimodal machine translation models. We plan to look at clustering approaches to cluster meanings or usages of words based on their semantic similarities. This can be used to assign ambiguity scores to each word based on the number of clusters.

Acknowledgements

We thank the anonymous reviewers for their insights on this work. This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289_P2 (Insight), co-funded by the European Regional Development Fund. 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

Aston, G. and Burnard, L. (1998). *The BNC Handbook: Exploring the British National Corpus with SARA*. Edinburgh textbooks in empirical linguistics. Edinburgh University Press.

Bowen, B., Vijayan, V., Grigsby, S., Anderson, T., and Gwinnup, J. (2024). Detecting concrete visual tokens for multimodal machine translation.

Caglayan, O., Aransa, W., Wang, Y., Masana, M., García-Martínez, M., Bougares, F., Barrault, L., and van de Weijer, J. (2016). Does multimodality help human and machine for translation and image captioning? In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 627–633, Berlin, Germany. Association for Computational Linguistics.

Caglayan, O., Madhyastha, P., Specia, L., and Barrault, L. (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.

Calixto, I., Rios, M., and Aziz, W. (2019). Latent variable model for multi-modal translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6392–6405, Florence, Italy. Association for Computational Linguistics.

Cho, K., van Merriënboer, B., Bahdanau, D., and Bengio, Y. (2014). On the properties of neural machine translation: Encoder–decoder approaches. In Wu, D., Carpuat, M., Carreras, X., and Vecchi, E. M., editors, *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar. Association for Computational Linguistics.

Elliott, D., Frank, S., Barrault, L., Bougares, F., and Specia, L. (2017). Findings of the Second Shared Task on Multimodal Machine Translation and Multilingual Image Description. In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark.

Elliott, D., Frank, S., Sima'an, K., and Specia, L. (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.

Futeral, M., Schmid, C., Laptev, I., Sagot, B., and Bawden, R. (2023). Tackling ambiguity with images: Improved multimodal machine translation and contrastive evaluation. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *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.

Gonzales, A. R., Mascarell, L., and Sennrich, R. (2017). Improving word sense disambiguation in neural machine translation with sense embeddings. In *Conference on Machine Translation*.

Harris, Z. S. (1954). Distributional structure. *WORD*, 10(2-3):146–162.

Hatami, A., Buitelaar, P., and Arcan, M. (2022). Analysing the correlation between lexical ambiguity and translation quality in a multimodal setting using WordNet. In Ippolito, D., Li, L. H., Pacheco, M. L., Chen, D., and Xue, N., editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 89–95, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Hatami, A., Buitelaar, P., and Arcan, M. (2023). A filtering approach to object region detection in multimodal machine translation. In Utiyama, M. and Wang, R., editors, *Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track*, pages 393–405, Macau SAR, China. Asia-Pacific Association for Machine Translation.

Hoffman, P., Ralph, M., and Rogers, T. (2012). Semantic diversity: A measure of semantic ambiguity based on variability in the contextual usage of words. *Behavior research methods*, 45.

Ive, J., Madhyastha, P., and Specia, L. (2019). Distilling translations with visual awareness. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6525–6538, Florence, Italy. Association for Computational Linguistics.

Klein, G., Kim, Y., Deng, Y., Nguyen, V., Senellart, J., and Rush, A. (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.

Kudo, T. and Richardson, J. (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.

Lala, C., Madhyastha, P. S., Scarton, C., and Specia, L. (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.

Lala, C. and Specia, L. (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).

Landauer, T. K. and Dumais, S. T. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104:211–240.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (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.

Popović, M. (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.

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (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.

Tang, G., Sennrich, R., and Nivre, J. (2018). An analysis of attention mechanisms: The case of word sense disambiguation in neural machine translation. In Bojar, O., Chatterjee, R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Monz, C., Negri, M., Névéol, A., Neves, M., Post, M., Specia, L., Turchi, M., and Verspoor, K., editors, *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 26–35, Brussels, Belgium. Association for Computational Linguistics.

Vaswani, A., Bengio, S., Brevdo, E., Chollet, F., Gomez, A. N., Gouws, S., Jones, L., Kaiser, L., Kalchbrenner, N., Parmar, N., Sepassi, R., Shazeer, N., and Uszkoreit, J. (2018). Tensor2tensor for neural machine translation. *Computing Research Repository (CoRR)*, abs/1803.07416.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *Computing Research Repository (CoRR)*, abs/1706.03762.

Vijayan, V., Bowen, B., Grigsby, S., Anderson, T., and Gwinnup, J. (2024). Adding multimodal capabilities to a text-only translation model.

Wang, D. and Xiong, D. (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.

Wu, Z., Kong, L., Bi, W., Li, X., and Kao, B. (2021). Good for misconceived reasons: An empirical revisiting on the need for visual context in multimodal machine translation. *CoRR*, abs/2105.14462.

Yao, S. and Wan, X. (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.

Yin, Y., Meng, F., Su, J., Zhou, C., Yang, Z., Zhou, J., and Luo, J. (2020). A novel graph-based multi-modal fusion encoder for neural machine translation. In *Annual Meeting of the Association for Computational Linguistics*.

Zhao, Y., Komachi, M., Kajiwara, T., and Chu, C. (2022). Region-attentive multimodal neural machine translation. *Neurocomputing*, 476:1–13.