

Multimodal Machine Translation: Leveraging Images for Enhanced Language Understanding

Sai Vamsee Bandaru
The City College of New York
New York, United States of America
sbandar000@citymail.cuny.edu

Prof. Jia Xu
Stevens Institute of Technology
New Jersey, United States of America
xujia@gmail.com

Dr. Abdul Rafae Khan
Stevens Institute of Technology
New Jersey, United States of America
rafae015@gmail.com

Abstract

Text-only machine translation systems often fail to resolve ambiguity when textual context is insufficient. This limitation is particularly evident in e-commerce product titles, which are short, noisy, and multilingual. The project proposes a multimodal machine translation framework that integrates visual context with textual input to generate context-aware translations. The system combines a pretrained SigLIP Vision Transformer with an mBART language model using a fusion layer and parameter-efficient fine-tuning via Low-Rank Adaptation (LoRA). Experiments conducted on Multi30K, Image-Guided Translation, and real-world e-commerce datasets across English, German, and French demonstrate that multimodal individual training improves BLEU scores from 41.79 to 43.51, consistently outperforming text-only baselines.

Keywords

Multimodal Machine Translation, Vision-Language Models, SigLIP, mBART, LoRA, E-commerce Translation

ACM Reference Format:

Sai Vamsee Bandaru, Prof. Jia Xu, and Dr. Abdul Rafae Khan. 2025. Multimodal Machine Translation: Leveraging Images for Enhanced Language Understanding. In . ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/nnnnnnnn.nnnnnnnn>

1 Introduction

Machine Translation (MT) has experienced rapid progress over the past decade with the adoption of deep learning and Transformer-based architectures, such as sequence-to-sequence models with self-attention. These advances have significantly improved translation fluency and grammatical correctness across many language pairs. Despite these improvements, most state-of-the-art MT systems rely exclusively on textual input, implicitly assuming that language alone is sufficient to capture meaning.

In many real-world applications, however, textual information is often incomplete or ambiguous. Words and phrases may have multiple meanings that cannot be resolved without additional context. This limitation is particularly shown in domains such as

e-commerce, where product titles are typically short, noisy, and frequently contain mixed-language terms, abbreviations, or incomplete descriptions. For example, visually grounded terms such as “boots,” “jacket,” or “pitch” may require image context to determine their correct semantic interpretation. As a result, text-only MT models often produce inconsistent or incorrect translations, negatively impacting user experience.

Multimodal Machine Translation (MMT) has emerged as a promising direction to address these challenges by incorporating visual information alongside text. By leveraging images associated with source sentences, multimodal systems can access complementary contextual cues that help disambiguate meaning and improve translation robustness. Visual signals are especially valuable for short text inputs, where linguistic context alone is insufficient.

In this project, we investigate a multimodal translation framework that fuses visual and textual representations to generate context-aware translations. The proposed approach integrates a pretrained vision-language encoder with a Transformer-based language model and employs parameter-efficient fine-tuning techniques to ensure scalability. Through experiments on benchmark datasets and real-world e-commerce data, this work demonstrates that incorporating visual context consistently improves translation quality over text-only baselines.

2 Problem Statement and Objectives

Text-only machine translation models fail to capture visual context, leading to ambiguous translations. In the e-commerce domain, inconsistent multilingual translations negatively impact user experience. So, these are the objectives the project would address and solve.

The objectives of this project are:

- To design a multimodal translation architecture combining image and text inputs
- To extract visual features using a pretrained SigLIP Vision Transformer
- To fine-tune language models using parameter-efficient LoRA techniques
- To evaluate performance using BLEU scores and compare against text-only baselines

3 Related Work

Recent surveys [1] on multimodal machine translation (MMT) highlight that the field has expanded beyond traditional image-caption translation tasks to address diverse real-world applications. Section 3 categorizes emerging MMT paradigms, including

Unpublished working draft. Not for distribution.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted by ACM, provided that the copies are not made for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
Conference'17, Washington, DC, USA

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-XXXX-X/2018/06
<https://doi.org/10.1145/nnnnnnnn.nnnnnnnn>

e-commerce product translation, text–image translation involving embedded visual text, video-guided translation using temporal context, multimodal simultaneous translation for real-time settings, and multimodal conversational translation systems. These approaches demonstrate that incorporating visual information can significantly improve translation quality in scenarios where textual input alone is ambiguous or incomplete. This broader perspective reinforces the motivation for applying multimodal translation techniques to short, noisy e-commerce product titles, as explored in this project.

Introduce the ConECT dataset, [2] a large-scale multimodal benchmark designed to address data scarcity in context-aware e-commerce machine translation. The dataset focuses on real-world product titles and integrates visual context to improve translation quality in scenarios where textual information alone is insufficient. The authors demonstrate that conventional text-only translation models struggle with short, ambiguous, and domain-specific product descriptions, while multimodal approaches leveraging images achieve more consistent and accurate translations across languages. Their work highlights the importance of domain-specific multimodal datasets and motivates the use of image–text fusion techniques for improving machine translation performance in e-commerce settings, which directly aligns with the objectives of this project.

The attention-based neural machine translation framework [3], which jointly learns to align and translate source and target sequences. Unlike earlier encoder–decoder models that compressed the entire source sentence into a fixed-length vector, their approach dynamically attends to relevant parts of the source sentence during decoding. This attention mechanism significantly improved translation quality, especially for longer and more complex sentences, and became a foundational component of modern neural machine translation systems. The concept of learned alignment proposed in this work directly influences contemporary multimodal translation architectures, where cross-attention mechanisms are used to fuse textual and visual representations.

[4] proposed a multimodal neural machine translation approach that incorporates visual information through an auxiliary learning objective, encouraging the model to ground textual representations in visual features. Instead of directly injecting image features at inference time, their method trains the translation model to jointly learn translation and visual imagination, improving semantic representations of the source text. Experiments on multimodal benchmarks demonstrate that visually grounded training can lead to improved translation quality, particularly in scenarios with limited textual context. This work provides early evidence that visual grounding enhances translation performance and motivates later fusion-based and cross-attention multimodal architectures, such as those explored in this project.

4 Dataset Overview

This project utilizes multiple datasets to evaluate the effectiveness of multimodal machine translation across different domains and language pairs. The selected datasets include both benchmark

multimodal translation corpora and real-world e-commerce data, enabling a comprehensive analysis of model performance under controlled and practical conditions.

4.1 Languages

The experiments were conducted across three languages:

- English (EN)
- German (DE)
- French (FR)

These languages were chosen due to their availability in existing multimodal datasets and their relevance to global e-commerce applications.

4.2 Multi30K Dataset

The Multi30K [6] dataset is a widely used benchmark for multimodal machine translation and image-captioning tasks. It consists of images paired with descriptive captions in multiple languages. In this project, approximately 15,000 training samples were used from the Multi30K dataset, covering all six translation directions (EN↔DE, EN↔FR, DE↔FR). The dataset provides high-quality aligned image–text pairs, making it suitable for learning visual grounding in translation models.

4.3 Image-Guided Translation Dataset

To further evaluate multimodal performance [5], an Image-Guided Translation dataset was employed. This dataset also contains paired images and multilingual textual descriptions, enabling direct comparison between text-only and image-enhanced translation models. Approximately 15,000 samples were used from this dataset to fine-tune both text-only and multimodal variants of the model for (EN↔DE) and 7500 samples for (EN↔FR).

4.4 Data Splits and Usage

All datasets were split into training and evaluation sets following standard practices. Models were trained under two settings: (i) individual language-pair training and (ii) combined training across all six language directions. This setup allowed for analysis of both specialized and generalized multimodal learning behavior.

Overall, the combination of benchmark and real-world datasets enables a robust evaluation of multimodal machine translation, demonstrating the strengths and limitations of visual context integration across different data distributions.

5 Methodology

This section describes the overall methodology adopted to design, train, and evaluate the proposed multimodal machine translation system. The approach integrates visual and textual information using a vision-language architecture and evaluates its effectiveness through systematic experiments across multiple datasets and language pairs.

Overall Framework : The proposed framework follows a multimodal translation pipeline that processes both textual and visual inputs. Given a source sentence and its corresponding image, the

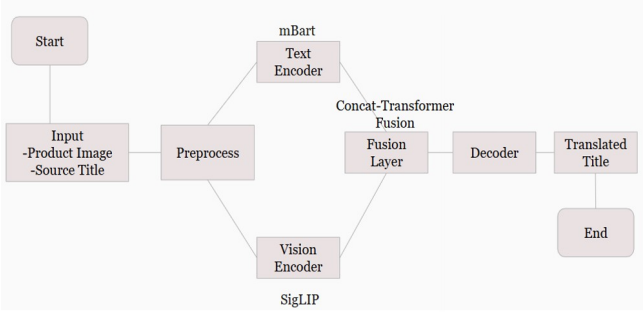


Figure 1: The image shows the flow of the architecture

system extracts visual features using a pretrained vision encoder and textual representations using a Transformer-based language model. These representations are fused through a cross-attention mechanism and decoded into the target language. The framework supports both text-only and multimodal training configurations, enabling a direct comparison between unimodal and multimodal translation performance.

Vision Encoder : Visual features [7] are extracted using a pretrained SigLIP Vision Transformer. The encoder processes product images and generates dense visual embeddings that capture semantic information relevant to the source text. SigLIP was selected due to its strong performance in vision-language representation learning and its compatibility with Transformer-based fusion mechanisms.

Text Encoder and Decoder : For textual processing [8], the multilingual BART (mBART) model is used as the backbone language model. The text encoder converts source-language tokens into contextual embeddings, while the decoder generates translations in the target language. mBART supports multilingual translation and provides a strong initialization for both low-resource and high-resource language pairs.

Multimodal Fusion : To integrate visual and textual information, a fusion layer based on cross-attention is employed. The textual embeddings attend to the visual embeddings, allowing the model to selectively incorporate image-based cues during translation. This fusion strategy enables the system to resolve ambiguities in visually grounded terms while preserving linguistic coherence.

Parameter-Efficient Fine-Tuning : To reduce computational cost and improve scalability, Low-Rank Adaptation (LoRA) [9] is applied during fine-tuning. LoRA adapters are inserted into the attention layers of the language model, allowing efficient adaptation to new datasets and language pairs without updating the full model parameters. This approach significantly reduces training overhead while maintaining competitive performance.

5.1 Training Strategy

Training is conducted under two settings:

- **Individual Training:** Separate models are trained for each language pair.
- **Combined Training:** A single model is trained across all six translation directions.
- **Pre-trained model + E-commerce data:** The pre-trained model is fine-tuned on e-commerce data.

Both text-only and multimodal variants are trained under each setting to evaluate the impact of visual context.

5.2 Implementation Details and Model Configuration

This subsection describes the implementation settings, model architecture choices, and hyperparameters used for training and evaluating the proposed multimodal machine translation system.

5.2.1 *Model Architecture.* The proposed system is built using a vision-language architecture that integrates pretrained models for both visual and textual processing:

- **Vision Encoder:** SigLIP Vision Transformer, pretrained on large-scale image-text data
- **Text Encoder-Decoder:** Multilingual BART (mBART)
- **Multimodal Fusion:** Concat-Transformer fusion layer with cross-attention
- **Fine-Tuning Strategy:** Low-Rank Adaptation (LoRA)

5.2.2 *Training Hyperparameters.* The following hyperparameters were used consistently across all experiments unless stated otherwise:

- **Batch Size:** 2
- **Maximum Sequence Length:** 64 tokens
- **Learning Rate:** 2×10^{-4}
- **Number of Epochs:** 6
- **Optimizer:** AdamW
- **Loss Function:** Cross-entropy loss with teacher forcing

5.2.3 *LoRA Configuration.* To enable parameter-efficient fine-tuning, LoRA adapters were applied to the attention layers of the language model:

- **LoRA Rank (r):** 8
- **Scaling Factor (α):** 16
- **Dropout Rate:** 0.1
- **Target Modules:** Attention projection layers in both the encoder and decoder

5.2.4 *Computational Environment.* All experiments were conducted in a GPU-enabled environment using google colab:

- **Hardware:** GPU-based training environment
- **Execution Platform:** Google Colab

5.3 Evaluation Metrics

Model performance is evaluated using the BLEU score, a standard metric for machine translation quality. BLEU scores are computed for each translation direction and averaged across language pairs. Comparative analysis is performed between text-only and multimodal models to quantify the contribution of visual information.

6 Results

After evaluating all individual models using performance metrics such as BLEU. These are the results for the test datasets. For multi30k dataset - test-2016-flickr and test-2017-flickr. For E-commerce the test dataset is present in the dataset.

6.1 Multimodal Training with Individual Language-Pair Fine-Tuning

Direction	Text-only	Multimodal	Δ Gain
EN→DE	36.09	37.42	+1.33
EN→FR	48.75	51.30	+2.55
DE→EN	45.79	45.85	+0.06
DE→FR	34.24	36.03	+1.79
FR→EN	50.30	50.47	+0.17
FR→DE	27.90	29.33	+1.43

Table 1: BLEU Scores on Flickr 2017 Test Set (Individual Language-Pair Training)

As shown in Table 1, multimodal translation consistently improves BLEU scores across all language pairs compared to text-only models. The largest gains are observed for EN→FR and DE→FR, indicating that visual context is most effective for resolving ambiguity in these translation directions.

Direction	Text-only	Multimodal	Δ Gain
EN→DE	38.55	40.38	+1.83
EN→FR	54.02	57.29	+3.27
DE→EN	45.04	46.17	+1.13
DE→FR	37.01	39.51	+2.50
FR→EN	51.67	54.33	+2.66
FR→DE	32.17	33.87	+1.69

Table 2: BLEU Scores on Flickr 2016 Test Set (Individual Language-Pair Training)

Table 2 demonstrates that multimodal machine translation achieves consistent BLEU improvements across all language pairs on the Flickr 2016 test set. The largest gains are observed for EN→FR (+3.27) and FR→EN (+2.66), highlighting the strong impact of visual context in improving translation quality for ambiguous and context-dependent language directions.

6.2 Multimodal Training with Multilingual Fine-Tuning

As shown in Table 3, multi modal translation improves BLEU scores across all language pairs by +0.66 to +2.78 points compared to text-only models. The highest gains are observed for EN→FR (+2.78) and DE→FR (+2.57), demonstrating the effectiveness of visual context

Direction	Text-only	Multimodal	Δ Gain
EN→DE	34.50	35.61	+1.11
EN→FR	44.47	47.25	+2.78
DE→EN	43.97	44.63	+0.66
DE→FR	31.50	34.07	+2.57
FR→EN	48.20	49.47	+1.27
FR→DE	26.32	28.40	+2.08

Table 3: BLEU Scores on Flickr 2017 Test Set (Multilingual Language-Pair Training)

Direction	Text-only	Multimodal	Δ Gain
EN→DE	37.74	38.30	+0.56
EN→FR	50.71	53.65	+2.94
DE→EN	44.38	45.03	+0.65
DE→FR	37.67	40.80	+3.13
FR→EN	52.08	55.88	+3.80
FR→DE	32.82	35.24	+2.42

Table 4: BLEU Scores on Flickr 2016 Test Set (Multilingual Language-Pair Training)

in resolving ambiguity in these translation directions.

Table 4 shows that multimodal machine translation consistently improves performance over text-only models on the Flickr 2016 test set, with BLEU gains ranging from +0.56 to +3.80. The largest improvements are observed for FR→EN (+3.80) and DE→FR (+3.13), while all other language pairs also show positive gains, confirming the effectiveness of visual context in enhancing translation quality across diverse translation directions.

After completing the evaluation of the both processes, we could infer that the individual model training is working better than the multilingual model training. While the multilingual model training loses the multi modality. So we can train the models individually for getting better results.

6.3 Multimodal Training with Individual Language-Pair Fine-Tuning for E-commerce data

Now, we need to check how the model works only for the e-commerce dataset. The same architecture and parameters were used to check the results in the same environment.

Table 5 indicates that when the model is trained only on the e-commerce dataset, multi modal translation does not provide consistent benefits. As shown, BLEU scores slightly decrease for both EN→DE (0.44) and DE→EN (0.27) compared to text-only models, suggesting that limited and noisy domain-specific data is insufficient for effective image-text fusion.

Direction	Text-only	Multimodal	Δ Gain
EN→DE	18.70	18.26	-0.44
DE→EN	20.21	19.94	-0.27

Table 5: BLEU Scores on E-Commerce Dataset (Text-Only vs Multimodal Training)

6.4 Individual Language-Pair Fine-Tuning with pre-trained model and E-commerce data

Now the same data is fine-tuned on the individual models to get better scores in the e-commerce data

Direction	Text-only	Multimodal	Gain
EN→DE	35.99	37.32	+1.33
DE→EN	44.25	45.38	+1.13
EN→FR	17.79	32.09	+14.30
FR→EN	20.97	31.08	+10.11

Table 6: BLEU Scores on Pretrained model + E-commerce dataset

Table 6 show that combining pretraining with e-commerce data used Table 5 and fine-tuning significantly enhances multimodal translation performance. The domain-specific fine-tuning on e-commerce data is critical for achieving strong multimodal translation gains in short, ambiguous product-title settings. In our experiments, the fine-tuning stage uses an e-commerce dataset derived from the Image-Guided Translation setup (product image + product title/description pairs across languages). For the EN→DE and DE→EN e-commerce fine-tuning runs, the model is fine-tuned on 15,000 training samples with an additional 2,000 samples reserved for validation.

For the EN→FR and FR→EN e-commerce fine-tuning runs, training is performed on 6,000 samples per direction. Large BLEU improvements are observed for EN→FR (+14.30) and FR→EN (+10.11), indicating that visual context becomes highly effective when supported by a strong pretrained model. Smaller but consistent gains for EN→DE (+1.33) and DE→EN (+1.13) further confirm that pre-training enables multimodal models to better leverage image-text alignment across language pairs.

7 Conclusion

On the Multi30K dataset, multimodal machine translation consistently outperforms text-only models. The average BLEU score for text-only individual training is 41.79, which increases to 43.51 with multimodal individual training, resulting in a gain of +1.72 BLEU. In the combined training setting, where the model learns all six language directions simultaneously, the average BLEU score improves from 40.36 (text-only) to 40.59 (multimodal). Although the improvement is smaller in this setting, it reflects the increased

generalization challenge of multi-direction training while still benefiting from visual context.

When trained exclusively on e-commerce data, multimodal learning provides limited benefit, with the average BLEU score slightly decreasing from 19.45 (text-only) to 19.01 (multimodal). However, when the model is pretrained on a large-scale multimodal dataset and subsequently fine-tuned on e-commerce data, performance improves significantly. The average BLEU score increases from 19.75 (text-only) to 36.91 (multimodal), representing nearly a 50 percent improvement. This demonstrates that multimodal pretraining is essential for effectively transferring visual knowledge to real-world, domain-specific translation tasks.

8 Future Scope

Future extensions of this work include expanding the multimodal machine translation framework to support additional languages, particularly low-resource and morphologically complex languages. Evaluating the model across a wider set of language pairs would help assess its scalability and robustness while improving translation quality in diverse linguistic settings. Additionally, building larger and more diverse multimodal datasets, especially for real-world domains such as e-commerce, can significantly enhance model performance by reducing noise and improving generalization through richer image-text alignments.

Another promising direction is improving the quality of image-text integration within the translation pipeline. Rather than relying solely on global image embeddings, future work can incorporate object detection or semantic segmentation techniques to extract fine-grained visual features. Leveraging region-level visual cues and stronger alignment strategies may allow the model to better associate visual elements with specific words or phrases, leading to more precise and context-aware translations.

9 References

- (1) Shen, H., Shao, L., Li, W., Lan, Z., Liu, Z., & Su, J. *A Survey on Multi-modal Machine Translation: Tasks, Methods and Challenges*. arXiv:2405.12669 (2024). <https://arxiv.org/abs/2405.12669>
- (2) Pokrywka, M., Kusa, W., Rutkowski, M., & Koszowski, M. *ConECT Dataset: Overcoming Data Scarcity in Context-Aware E-Commerce Machine Translation*. arXiv:2506.04929 (2025). <https://arxiv.org/abs/2506.04929>
- (3) Bahdanau, D., Cho, K., & Bengio, Y. *Neural Machine Translation by Jointly Learning to Align and Translate*. arXiv:1409.0473 (2016). <https://arxiv.org/abs/1409.0473>
- (4) Elliott, D., Frank, S., Sima'an, K., & Specia, L. *Imagination Improves Multimodal Translation*. arXiv:1702.01287 (2017). <https://arxiv.org/abs/1702.01287>
- (5) eBay Inc. *Image-Guided Machine Translation Dataset for E-Commerce*. (2018). <https://github.com/eBay/ImageGuidedTranslationDataset>
- (6) Sankesara, H. *Flickr Image Dataset*. Kaggle (2018). <https://www.kaggle.com/datasets/hsankesara/flickr-image-dataset>
- (7) Hugging Face. *SigLIP Model Documentation*. (2023). https://huggingface.co/docs/transformers/en/model_doc/siglip

- (8) Hugging Face. *mBART Model Documentation*. (2023). https://huggingface.co/docs/transformers/en/model_doc/mbart
- (9) Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. *LoRA: Low-Rank Adaptation of Large*

Language Models. arXiv:2106.09685 (2021). <https://arxiv.org/abs/2106.09685>