

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

xujjia@gmail.com

Dr. Abdul Rafee 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 Rafee 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 part or all of this work for personal or
internal use is granted by the copyright holders for the author(s) and/or editor(s), to be
distributed for profit or commercial advantage and that copies bear this notice and the full citation
on the first page. Copyright for components of this work owned by others than the
author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or
republishto 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>

117 e-commerce product translation, text–image translation involving
 118 embedded visual text, video-guided translation using temporal context,
 119 multimodal simultaneous translation for real-time settings, and
 120 multimodal conversational translation systems. These approaches
 121 demonstrate that incorporating visual information can significantly
 122 improve translation quality in scenarios where textual input alone
 123 is ambiguous or incomplete. This broader perspective reinforces
 124 the motivation for applying multimodal translation techniques to
 125 short, noisy e-commerce product titles, as explored in this project.
 126

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

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

153 [4] proposed a multimodal neural machine translation approach
 154 that incorporates visual information through an auxiliary learning
 155 objective, encouraging the model to ground textual representa-
 156 tions in visual features. Instead of directly injecting image features
 157 at inference time, their method trains the translation model to
 158 jointly learn translation and visual imagination, improving seman-
 159 tic representations of the source text. Experiments on multimodal
 160 benchmarks demonstrate that visually grounded training can lead
 161 to improved translation quality, particularly in scenarios with lim-
 162 ited textual context. This work provides early evidence that visual
 163 grounding enhances translation performance and motivates later
 164 fusion-based and cross-attention multimodal architectures, such as
 165 those explored in this project.

166 4 Dataset Overview

167 This project utilizes multiple datasets to evaluate the effective-
 168 ness of multimodal machine translation across different domains
 169 and language pairs. The selected datasets include both benchmark

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

173 4.1 Languages

174 The experiments were conducted across three languages:

- 175 • English (EN)
- 176 • German (DE)
- 177 • French (FR)

178 These languages were chosen due to their availability in existing
 179 multimodal datasets and their relevance to global e-commerce ap-
 180 plications.

181 4.2 Multi30K Dataset

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

190 4.3 Image-Guided Translation Dataset

191 To further evaluate multimodal performance [5], an Image-Guided
 192 Translation dataset was employed. This dataset also contains paired
 193 images and multilingual textual descriptions, enabling direct com-
 194 parison between text-only and image-enhanced translation mod-
 195 els. Approximately 15,000 samples were used from this dataset to
 196 fine-tune both text-only and multimodal variants of the model for
 197 ($EN \leftrightarrow DE$) and 7500 samples for ($EN \leftrightarrow FR$).

198 4.4 Data Splits and Usage

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

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

208 5 Methodology

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

215 **Overall Framework :** The proposed framework follows a mul-
 216 timodal translation pipeline that processes both textual and visual
 217 inputs. Given a source sentence and its corresponding image, the

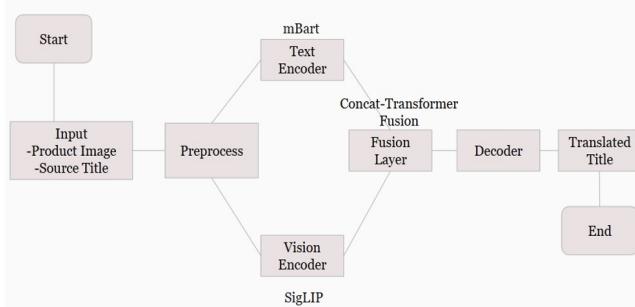


Figure 1: The image shows the flow of the architecture

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 pre-trained 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:

2025-12-18 05:09. Page 3 of 1–6.

- **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.

291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348

349 6 Results

350 After evaluating all individual models using performance met-
 351 rics such as BLEU. These are the results for the test datasets. For
 352 multi30k dataset - test-2016-flickr and test-2017-flickr. For E-commerce
 353 the test dataset is present in the dataset.
 354

355 6.1 Multimodal Training with Individual 356 Language-Pair Fine-Tuning

359 Direction	360 Text-only	361 Multimodal	362 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

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

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

378 Direction	379 Text-only	380 Multimodal	381 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

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

391 Table 2 demonstrates that multimodal machine translation achieves
 392 consistent BLEU improvements across all language pairs on the
 393 Flickr 2016 test set. The largest gains are observed for EN→FR
 394 (+3.27) and FR→EN (+2.66), highlighting the strong impact of vi-
 395 sual context in improving translation quality for ambiguous and
 396 context-dependent language directions.

399 6.2 Multimodal Training with Multilingual 400 Fine-Tuning

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

407 Direction	408 Text-only	409 Multimodal	410 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

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

418 Direction	419 Text-only	420 Multimodal	421 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

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

433 in resolving ambiguity in these translation directions.

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

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

449 6.3 Multimodal Training with Individual 450 Language-Pair Fine-Tuning for E-commerce 451 data

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

455 Table 5 indicate that when the model is trained only on the
 456 e-commerce dataset, multi modal translation does not provide con-
 457 sistent benefits. As shown, BLEU scores slightly decrease for both
 458 EN→DE (0.44) and DE→EN (0.27) compared to text-only models,
 459 suggesting that limited and noisy domain-specific data is insuffi-
 460 cient 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

581	(8) Hugging Face. <i>mBART Model Documentation</i> . (2023). https://huggingface.co/docs/transformers/en/model_doc/mbart	639
582		640
583	(9) Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S.,	641
584	Wang, L., & Chen, W. <i>LORA: Low-Rank Adaptation of Large</i>	642
585		643
586		644
587		645
588		646
589		647
590		648
591		649
592		650
593		651
594		652
595		653
596		654
597		655
598		656
599		657
600		658
601		659
602		660
603		661
604		662
605		663
606		664
607		665
608		666
609		667
610		668
611		669
612		670
613		671
614		672
615		673
616		674
617		675
618		676
619		677
620		678
621		679
622		680
623		681
624		682
625		683
626		684
627		685
628		686
629		687
630		688
631		689
632		690
633		691
634		692
635		693
636		694
637		695
638	2025-12-18 05:09. Page 6 of 1–6.	696