

# A Collaborative Approach to Multimodal Machine Translation: VLM and LLM

Amulya Ratna Dash and Yashvardhan Sharma

*Birla Institute of Technology and Science, Pilani, Rajasthan, India*  
{p20200105, yash}@pilani.bits-pilani.ac.in

**Keywords:** Natural Language Processing, Multimodal Machine Translation, Large Language Models, Image Captioning.

**Abstract:** With advancements in Large Language Models (LLMs) and Vision Language Pretrained Models (VLMs), there is a growing need to evaluate their capabilities and research methods to use them together for vision language tasks. This study focuses on using VLM and LLM collaboratively for Multimodal Machine Translation (MMT). We finetune LLaMA-3 to use provided image captions from VLMs to disambiguate and generate accurate translations for MMT tasks. We evaluate our novel approach using the German, French and Hindi languages, and observe consistent translation quality improvements. The final model shows an improvement of +3 BLEU score against the baseline and +4 BLEU score against the state-of-the-art model.

## 1 INTRODUCTION

Machine Translation (MT) is a Natural Language Processing (NLP) task in which text from a source language is translated to another target language while preserving the semantics and required terminology. Classical machine translation systems only perceive textual information, ignoring useful information from visual modalities such as images and videos. Multimodal Machine Translation (MMT) is the process where other modalities are used to improve the quality of the language translation. The most popular modality in MMT is a visual clue or image. The visual context helps disambiguate a few words by providing additional information along with source textual context. MMT can be extensively adopted for diverse practical applications, including subtitle translations considering image/video along with original source language movie script, and the translation of product descriptions and reviews in e-commerce platforms from original product listing language to preferred language of customer. Consequently, there is an increasing focus on research in MMT.

In recent years, Large Language Models (LLMs) have become increasingly widespread. Models like GPTs (Brown et al., 2020), LLaMA (Touvron et al., 2023), and others have shown remarkable performance in various NLP tasks such as text generation, question answering, and summarization. Vision Language Models (VLMs) understand and generate language in the context of visual information. These

models leverage large datasets of paired images and text to jointly learn the visual and linguistic relationship. VLMs have shown steady improvement in tasks such as image captioning and visual question answering (VQA).

In this work, we propose a collaborative approach to Multimodal Machine Translation, using Vision Language Pretrained Models (VLMs) and Large Language Models (LLMs). By integrating VLMs into our methodology, we propose to utilize their ability to offer enhanced contextual insights, thus boosting the effectiveness of Multimodal Machine Translation using LLMs. Our findings show that fine-tuning LLMs with extremely limited training data and computational resource can also help improve the translation accuracy for both high resource and low resource languages.

The rest of the paper is organized as follows. Section 2 presents the review of related works. The methodology and dataset are briefly described in Section 3. Sections 4 and 5 detail the experiments and results, followed by the conclusion and future scope in Section 6.

## 2 RELATED WORK

### 2.1 Multimodal Machine Translation

Most of the multimodal machine translation (MMT) work are based on Multi30K (Elliott et al., 2016)

dataset. Initial approaches focused on using visual features in sequence-to-sequence models based on Recurrent Neural Networks (RNNs) and Transformers (Vaswani, 2017) architecture. (Calixto et al., 2017) proposed a doubly-attentive model for both source text and global image features. Gated Fusion showed a mechanism to use image features in cross-attention of decoder. (Libovický and Helcl, 2017) used flat and hierarchical approach to combine the attention mechanism for multimodal translation. Most of the research was benchmarked on English to German and French evaluation datasets.

Parida et al., 2019 proposed Hindi Visual Genome dataset, a subset of the Visual Genome (Krishna et al., 2017) dataset for multimodal translation between English and Hindi, allowing researchers to experiment and benchmark using a new language. An adversarial study of MMT showed replacing/removing the associated image does not hamper the MMT output, which suggests that source text is sufficient to generate target text semantically close to reference evaluation text (Elliott, 2018). CoMMuTe (Futeral et al., 2022) dataset helps to better evaluate MMT systems, as it contains lexically ambiguous sentences and only the associated image helps disambiguate. In our work, we evaluate our MMT approach using CoMMuTe and HVG datasets.

## 2.2 Pretrained Language Models and MMT

mBART (Chipman et al., 2022), a denoising auto-encoder pre-trained on monolingual corpora in multiple languages shows its effectiveness for machine translation(MT). Zhu et al., 2020 used BERT to improve neural machine translation, by fusing sentence representations from BERT with each encoder and decoder layer. CLIP (Bordes et al., 2024) model learns to perform multiple computer vision tasks like image captioning during pretraining, using contrastive language image pretraining approach. BLIP (Li et al., 2022) model uses the vision language pretraining approach which improves both vision language understanding and generation tasks. BLIP 2 (Li et al., 2023) further improves the accuracy on vision-language tasks by reducing the image and text modality gap with a Querying Transformer. Florence-2 (Yuan et al., 2021) is a vision-language model which takes text instructions and generates textual results based on the image. The performance of Florence-2-L model(0.77B params) on captioning and visual question answering tasks are comparable to other much larger models like BLIP-2 and Flamingo (Alayrac et al., 2022).

Futeral et al., 2022 propose a multimodal MT model VGAMT based on mBART. VGAMT adds bottleneck adapters (Houlsby et al., 2019), and linear visual projection layers to the finetuned mBART, uses MDETR (Misra et al., 2021) and CLIP for local and global image features, respectively. Our work uses BLIP-2 and Florence to generate image captions, followed by adapting a large language model to condition upon the generated image caption while performing the translation task.

## 2.3 Large Language Models and MT

The machine translation capability of LLMs has been studied by multiple researchers. Xu et al. (2023) proposed a paradigm shift in machine translation which adapts LLaMA for translation using two steps, continued pretraining on monolingual data and finetuning using bilingual data. Tower LLM (Alves et al., 2024) is a multilingual LLM for translation-related tasks which also adapts LLaMA using two steps, uses monolingual and parallel data in first step and finetunes on instructions for multiple tasks like translation, automatic post editing (APE), named entity recognition (NER) and grammatical error correction(GEC) in the second step. In our work, we focus on adapting LLaMA for multimodal machine translation via single-step fine-tuning using extremely limited amount of finetuning data tuples(image caption and bilingual text data).

# 3 METHODOLOGY

## 3.1 Models

To investigate the collaborative approach to Multimodal Machine Translation, we require two categories of models: an Image-to-Text Model and a Large Language Model. The Image-to-Text model is used to generate the caption of the image associated with the source text. The Large Language Model is used to translate the source language text into target language text, using the description of the image generated by Image-to-Text model as context. BLIP-2 and Florence-2 are used as Image-to-Text models and LLaMa 3 as the Large Language Model. BLIP 2 offers multiple model variants based on underlying backbone model and model size, for image encoder (ViT-L or ViT-g) and for text generation (OPT 2.7B, OPT 6.7B, FlanT5 XL, FlanT5 XXL). LLaMa 3 offers two model variants based on model size, a 2 billion and 8 billion parameters model. The smallest model variant was selected so that with limited

resource we can perform efficient finetuning and inference.

We use BLIP-2 (ViT-L OPT-2.7B) model, which uses OPT-2.7B model along with a CLIP-like image encoder and a Querying Transformer to generate text conditioned on image and optional text. LLaMa 3 8B<sup>1</sup> model was shortlisted as it focuses on being multilingual with high quality non-English data in the training dataset, resulting in impressive zero-shot translation performance. LLaMA 3 8B model was finetuned using the LoRA (Hu et al., 2021) parameter efficient fine-tuning technique (PEFT).

### 3.2 Languages and Evaluation Data

To evaluate our approach on Multimodal machine translation, we consider 3 languages: German, French and Hindi. CoMMuTE, a Contrastive Multilingual Multimodal Translation Evaluation dataset of lexically ambiguous sentences whose correct translation requires the context from associated image. Multimodal machine translation test sets like Test2016, Test2017, MSCOCO, HVG are some of the other available evaluation datasets. CoMMuTE (300 segments) was considered as the evaluation dataset for German and French tasks, as other evaluation datasets have relatively lower percentage of ambiguous examples, thus diluting the evaluation score. HVG Challenge set (1400 segments) was used as an evaluation dataset for the Hindi task.

### 3.3 Evaluation Metrics

ChrF (Popović, 2015), COMET-22 (Rei et al., 2022) and BLEU (Papineni et al., 2002) score were used as evaluation metrics. ChrF evaluates translations at character level which make it more robust as compared to token level metrics. Crosslingual Optimized Metric for Evaluation of Translation 2022 (COMET 22) is an embedding-based evaluation metric, which considers context and semantics of the text while assigning scores. BLEU score is calculated by comparing the n-gram overlap between the system and the reference texts.

<sup>1</sup><https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct>

## 4 EXPERIMENTS

### 4.1 Tasks

We consider three multimodal translation tasks: (i) English → German (ii) English → French (iii) English → Hindi. First, we create a text only translation baseline using LLaMa 3 model. Secondly, we perform multimodal machine translation using our cascaded approach, where we generate image description using BLIP-2 model and use it as context for translation using LLaMa 3 model. Finally, we use our finetuned LLaMA 3 model along with image captions generated from BLIP-2 and Florence-2 models to perform multimodal machine translation.

### 4.2 Data

The finetuning dataset consists of German, French and Hindi portions. We use the *val* split of Multi30K dataset for German and French, and the *dev* split of the HindiVisualGenome (HVG) dataset for Hindi. The context (in English) part of our training dataset is generated using the BLIP-2 model for the image associated with each record. Then we process the context phrase and bilingual text data into a chat instruction template as required for finetuning LLaMa 3 model. The template used for training data is available in Figure 1. The distribution of processed dataset is shown in Table 1.

Table 1: Composition of dataset used for finetuning LLM.

Language Pair	Train	Val
English - German	900	100
English - French	900	100
English - Hindi	900	100
Total	2700	300

### 4.3 Training Setup

The model was finetuned for multilingual multimodal machine translation using 2700 training records and 300 validation records. The target modules for LoRA finetuning are Query, Key, Value, Output, Gate, Up and Down projections. The hyperparameters are detailed in Table 2.

Role	Instruction and Response Template
Human	<p>Don't provide any justification or extra output, just the translated text.</p> <p>Translate the following sentence using the given context from <math>\langle \text{Source language} \rangle</math> to <math>\langle \text{Target language} \rangle</math>:</p> <p>Context: <math>\langle \text{Image Caption generated by VLM} \rangle</math>  <math>\langle \text{Text in Source language} \rangle</math></p>
Assistant (LLM)	$\langle \text{Text in Target language} \rangle$

Figure 1: Template of training dataset.

Table 2: Training hyperparameters.

Parameters	Values
Learning rate	0.0003
Train batch size	8
Eval batch size	4
Seed	3407
Gradient accumulation	8
Total train batch size	64
Optimizer	Adam(Kingma, 2014)
Adam $\beta_1$	0.9
Adam $\beta_2$	0.999
Adam $\epsilon$	1e-08
LR scheduler type	cosine
LR warmup steps	0.01
Num of Epochs	2
Num of GPU	1 (A100 40GB)
LoRA Rank	16
LoRA $\alpha$	16
LoRA Bias	None

## 5 RESULTS & ANALYSIS

### 5.1 Results

We evaluated vanilla LLaMA 3 for Text only translation and used the scores as baseline result. The comparative scores from cascaded approach of VLM and LLM (vanilla and finetuned) can be found in Table 3 for  $English \rightarrow \{German, French\}$  and in Table 4 for  $English \rightarrow Hindi$ . We compared the state-of-the-art BLEU scores reported by the authors of VGAMT on CoMMuTE dataset with our best model as shown in Table 5.

- CoMMuTe Eval set: Our model MMT-LLaMa-3 when augmented with caption from Florence model during inference received **5** best scores out

of 3 metrics each for  $English \rightarrow German$  and  $English \rightarrow French$  multimodal translations. COMET score improved by **+4** and **+3.4** points, and ChrF score boosted by **+3.76** and **+4.27** absolute points for  $English \rightarrow German$  and  $English \rightarrow French$  respectively.

MMT-LLaMa-3 + BLIP collaborative method outperformed existing methods by a margin of **+4** for both  $English \rightarrow \{German, French\}$  multimodal translations, with only 2700 training data and less than 15 minutes of GPU training time.

- HVG Challenge Eval set: Our model MMT-LLaMa-3 + BLIP collaborative approach outperformed the MMT-LLaMa-3 + Florence-2 collaborative approach by **+3.25** ChrF points. However, the clear winner on the multimodal  $English \rightarrow Hindi$  translation was vanilla LLaMA 3 (text translation without any hint from the associated image) based on average scores.

### 5.2 Analysis

- German and French are from the Germanic and Romance language family, and belong to Indo-European languages. We observe almost similar gains of +4 points for both German and French. This can be attributed to the high resourcefulness of both languages and maybe to usage of an equal number (around 900) of training data segments during the fine-tuning process. In addition, the English source segments from the CoMMuTE datasets overlap almost completely.
- The performance of the finetuned model on the HVG Challenge set for  $English \rightarrow Hindi$  multimodal translation was not aligned with our hypothesis. The observation from analyzing few individual translations are that majority of the source segments in HVG Challenge set are of shorter length relatively, which makes it easier to

Table 3: MultiModal Translation performance for *English*  $\rightarrow$  {*German, French*} on CoMMuTE dataset.

Model	German			French		
	BLEU	COMET	ChrF	BLEU	COMET	ChrF
LLaMa 3	30.52	78.49	50.83	33.02	77.63	53.25
LLaMa 3 + BLIP	32.32	81.95	53.96	35.76	80.73	57.46
MMT-LLaMa-3 (Ours) + BLIP	<b>33.3</b>	82.36	54.31	36.28	80.89	56.82
MMT-LLaMa-3 (Ours) + Florence	32.79	<b>82.47</b>	<b>54.59</b>	<b>36.69</b>	<b>81.04</b>	<b>57.52</b>

Table 4: MultiModal Translation performance for *English*  $\rightarrow$  *Hindi* on HVG Challenge dataset.

Model	BLEU	COMET	ChrF
LLaMa 3	<b>23.27</b>	<b>73.59</b>	<b>48.79</b>
LLaMa 3 + BLIP	16.83	69.08	43.7
MMT-LLaMa-3 (Ours) + BLIP	11.5	72.5	43.31
MMT-LLaMa-3 (Ours) + Florence	9.31	71.14	40.06

Table 5: Comparison of BLEU score on *English*  $\rightarrow$  {*German, French*} on CoMMuTE dataset.

Model	German	French
VGAMT	29.3	32.2
MMT-LLaMa-3 + BLIP (Ours)	<b>33.3</b>	<b>36.28</b>

translate directly. BLIP generated very short captions as compared to Florence, thus the shorter context helped get better scores for translation as the attention span reduced due to short and to-the-point caption used as context. Few possible reasons which we have not validated extensively may be related to less image dependency and ambiguity coverage in the evaluation dataset.

- Sample inference examples are shown in Table 6 and 7.
- In Table 6, the words “bow” are disambiguated with the help of the image. The image shows a red bow sitting on a table and a young boy holding a bow and arrow. This context helps clarify that the first “bow” refers to a decorative bow, while the second “bow” refers to a weapon used for archery. The image context facilitated the German translations, with “Schleier” for the decorative bow and “Bogen” for the archery bow. In the first image, if the VLM would have generated the caption with word ‘red veil’ instead of ‘red bow’, the German translation using our finetuned model would be ‘Schleife’.
- In Table 7, the words “chopper” are disambiguated with the help of the image. The image

shows a blue motorcycle parked in a parking lot next to other motorcycles and a helicopter flying in the sky over a house. This context helps clarify that the first “chopper” refers to a motorcycle, while the second “chopper” refers to a helicopter. The French translation accurately reflects this distinction, with “chopper” for the motorcycle and “hélicoptère” for the helicopter.

## 6 CONCLUSION AND FUTURE WORK

The multilingual ability of LLMs are increasing rapidly with the release of new LLMs. These LLMs being trained on enormous amounts of web scale data and using extremely powerful GPU clusters; they should be utilized for novel use cases. Most of the times, in-context learning or adapter based finetuning of LLMs helps to get acceptable accuracy for some NLP problems like machine translation with limited data and GPU power.

Our study concludes that using a collaborative approach of using a pre-trained small vision language foundation models (VLM) and multilingual large language models (LLM) jointly can provide a resource-efficient solution to multimodal machine translation of both low resource and high resource languages. An improvement of +4 in the ChrF score and +3 points for BLEU and COMET score was achieved compared to the baseline score for the German and French translation tasks.

In the future, we would experiment the multimodal machine translation problem with additional language pairs and perform a comparative evaluation of multiple large language models on diverse evaluation dataset to develop a generalized highly accurate MMT system.

Table 6: Sample multimodal translation of English → German, where Caption is used as context by the finetuned model as visual hint.





<b>Image</b>		
<b>Caption:</b>	A red bow sitting on top of a table.	A young boy holding a bow and arrow in his hand.
<b>Source Text:</b>	Hand me that bow.	Hand me that bow.
<b>Translated Text:</b>	Gib mir den Schleier.	Gib mir jenen Bogen.
<b>Reference Translation:</b>	Gib mir die Schleife.	Gib mir den Bogen.

Table 7: Sample multimodal translation of English → French, where Caption is used as context by the finetuned model as visual hint.

<b>Image</b>		
<b>Caption:</b>	A blue motorcycle parked in a parking lot next to other motorcycles.	A helicopter flying in the sky over a house.
<b>Source text:</b>	My husband bought a chopper recently.	My husband bought a chopper recently.
<b>Translated text:</b>	Mon mari a acheté un chopper récemment.	Mon mari a acheté un hélicoptère récemment.
<b>Reference Translation:</b>	Mon mari a acheté une moto récemment.	Mon mari a acheté un hélicoptère récemment.

## REFERENCES

- Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. (2022). Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.
- Alves, D. M., Pombal, J., Guerreiro, N. M., Martins, P. H., Alves, J., Farajian, A., Peters, B., Rei, R., Fernandes, P., Agrawal, S., et al. (2024). Tower: An open multilingual large language model for translation-related tasks. *arXiv preprint arXiv:2402.17733*.
- Bordes, F., Pang, R. Y., Ajay, A., Li, A. C., Bardes, A., Petryk, S., Mañas, O., Lin, Z., Mahmoud, A., Jayaraman, B., et al. (2024). An introduction to vision-language modeling. *arXiv preprint arXiv:2405.17247*.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Calixto, I., Liu, Q., and Campbell, N. (2017). Doubly-attentive decoder for multi-modal neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1913–1924.
- Chipman, H. A., George, E. I., McCulloch, R. E., and Shively, T. S. (2022). mbart: multidimensional monotone bart. *Bayesian Analysis*, 17(2):515–544.
- Elliott, D. (2018). Adversarial evaluation of multimodal machine translation. In Riloff, E., Chiang, D., Hockenmaier, J., and Tsujii, J., editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2974–2978, Brussels,

- Belgium. Association for Computational Linguistics.
- 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.
- Futeral, M., Schmid, C., Laptev, I., Sagot, B., and Bawden, R. (2022). Tackling ambiguity with images: Improved multimodal machine translation and contrastive evaluation. *arXiv preprint arXiv:2212.10140*.
- Houlsby, N., Giurghi, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., and Gelly, S. (2019). Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR.
- Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2021). Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.
- Kingma, D. P. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., et al. (2017). Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73.
- Li, J., Li, D., Savarese, S., and Hoi, S. (2023). Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.
- Li, J., Li, D., Xiong, C., and Hoi, S. (2022). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.
- Libovický, J. and Helcl, J. (2017). Attention strategies for multi-source sequence-to-sequence learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 196–202.
- Misra, I., Girdhar, R., and Joulin, A. (2021). An end-to-end transformer model for 3d object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2906–2917.
- Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In Isabelle, P., Charniak, E., and Lin, D., editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Parida, S., Bojar, O., and Dash, S. R. (2019). Hindi visual genome: A dataset for multi-modal english to hindi machine translation. *Computación y Sistemas*, 23(4):1499–1505.
- 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.
- Rei, R., De Souza, J. G., Alves, D., Zerva, C., Farinha, A. C., Glushkova, T., Lavie, A., Coheur, L., and Martins, A. F. (2022). Comet-22: Unbabel-ist 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585.
- Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Vaswani, A. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.
- Xu, H., Kim, Y. J., Sharaf, A., and Awadalla, H. H. (2023). A paradigm shift in machine translation: Boosting translation performance of large language models. *arXiv preprint arXiv:2309.11674*.
- Yuan, L., Chen, D., Chen, Y.-L., Codella, N., Dai, X., Gao, J., Hu, H., Huang, X., Li, B., Li, C., et al. (2021). Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*.
- Zhu, J., Xia, Y., Wu, L., He, D., Qin, T., Zhou, W., Li, H., and Liu, T.-Y. (2020). Incorporating bert into neural machine translation. *arXiv preprint arXiv:2002.06823*.