Towards Improving Multimodal Machine Translation with LLMs: A Focus on Indic Languages

Amulya Ratna Dash, Chirag Wadhwa and Yashvardhan Sharma
Birla Institute of Technology and Science, Pilani, Rajasthan, India
p20200105@pilani.bits-pilani.ac.in, chiragwadhwa.55555@gmail.com, yash@pilani.bits-pilani.ac.in

Abstract

Recent advances in Multimodal Machine Translation (MMT) have attempted to address ambiguity and polysemy in text alone by enabling models to draw additional contextual cues from paired images, thereby improving disambiguation and translation accuracy. Datasets such as Multi30K and Visual Genome have significantly advanced this line of research. However, these datasets do not always compel models to rely on visual information. The CoMMuTE dataset takes a stronger step in this direction by serving as an evaluation benchmark specifically designed around ambiguous English sentences that can only be correctly interpreted with their accompanying images. In this work, we extend CoMMuTE to two Indic languages, introducing IndicCoMMuTE — an evaluation dataset for assessing MMT systems on low-resource Indic languages. We benchmark a range of open-source multimodal Large Language Models (< 15B parameters) and a strong text-only baseline across eight languages. We fine-tune one of these LLMs on two Indic languages. Our findings provide insights into the strengths and limitations of LLMs and establish IndicCoMMuTE as a valuable benchmark for future research on Multimodal Machine Translation in Indic languages.

Keywords: Multimodal Machine Translation, Large Language Models, Low Resource Languages, Parameter Efficient Fine-Tuning

1. Introduction

Over the past few decades, the field of machine translation has witnessed multiple breakthroughs. Neural Machine Translation (NMT) revolutionized automatic translation by introducing end-to-end sequence-to-sequence models (Sutskever et al., 2014) that outperformed the previously used statistical approaches. Attention mechanism proposed by (Bahdanau et al., 2016) allowed the models to selectively focus on the most relevant parts of the input to predict the next token, thereby improving the translation quality. The subsequent introduction of the Transformer model by (Vaswani et al., 2017), leveraging multi-head self-attention, further boosted scalability and fluency, rapidly becoming the state-of-the-art across Machine Translation and numerous Natural Language Processing (NLP) tasks. Previous works, such as mBART (Liu et al., 2020) and M2M-100 (Fan et al., 2020), have extended these improvements to low-resource languages by facilitating effective cross-lingual transfer and fine-tuning on limited parallel data.

Recent years have witnessed an increase in focus on Multimodal Machine Translation (MMT), where model inputs are accompanied by other data sources (images or videos) along with the source text. Inherent ambiguities, particularly in the form of polysemous words, can pose challenges for text-only models to produce accurate translations. Additional contextual cues from images can aid the model in producing more accurate translations. For example, the English word *"glass"* may refer to a drinking glass or eyeglasses; an accompanying image can clarify the intended meaning.

Despite this promise, real-world benefits from images have been modest: the visual modality is often ignored by sufficiently strong text-only models, and progress is hampered by limited datasets and evaluation frameworks that truly test visual understanding (for instance, cases where the accompanying image is necessary for disambiguation) (Futeral et al., 2023). The CoMMuTE evaluation set (Futeral et al., 2023) was introduced to address this problem and provide a standardized benchmark for MMT models. It contains a set of ambiguous English sentences and their possible translations in six languages (French, German, Czech, Arabic, Russian, and Chinese), accompanied by disambiguating images corresponding to each translation.

This situation is worse for low-resource languages such as Indic languages, as most of the previous works in MMT have focused on high-resource languages. The scarcity of high-quality multimodal datasets for Indic languages and a lack of strong evaluation benchmarks make it difficult to assess the effectiveness of MMT models.

With the introduction of Large Language Models (LLMs), the research in MT has been fundamentally transformed, with models like GPT-4, PaLM, and open source variants like Llama that demonstrate strong zero-shot and few-shot performance

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 8872–8882
11–16 May 2026.

through in-context learning, even without explicit bilingual supervision. As shown by (Zhu et al., 2024), LLMs show impressive translation capabilities, and models like GPT-4 even beat supervised models such as NLLB (Team, 2024) in a number of different languages, although a significant gap remains in translation quality for low-resource languages. Support for multimodality in open-source LLMs offers promising avenues for MMT. Prior research on LLMs' capabilities in MMT for low-resource Indic languages (Khan et al., 2025;Dash and Sharma, 2025) has shown promising results. Building on this, we conduct a systematic study of state-of-the-art open-source LLMs for MMT of low-resource Indic languages, assessing their ability to leverage visual cues for contextual disambiguation.

In this paper, we focus on Multimodal Machine Translation of English → Hindi and English → Odia languages in a contrastive evaluation setting. Our contributions are as follows.

- We introduce an extended multimodal evaluation framework for two Indic languages based on CoMMuTE datasets.
- We benchmark five open-source LLMs and a text-only baseline in contrastive multimodal and text-only settings for eight languages.
- We fine-tune a multimodal LLM, demonstrating significant improvements in translation quality for low-resource Indic languages with limited compute and data resources.
- We demonstrate that fine-tuning on Indic languages also improves multimodal translation performance on unrelated language pairs.

The rest of the paper is organized as follows. Section 2 reviews related works. Section 3 describes the IndicCoMMuTE evaluation dataset, benchmarking and fine-tuning process. Section 4 presents the experimental results and analysis. Finally, Section 5 provides the conclusion and outlines the future directions.

2. Related Works

2.1. Multimodal Machine Translation (MMT)

The aim of Multimodal Machine Translation is to leverage additional context present in non-textual data, primarily images, that can aid the models in resolving the inherent ambiguities present in language. One of the foundational datasets used by multiple studies in this field is the Multi30K dataset (Elliott et al., 2016). It is an extension to Flickr30K (Young et al., 2014) and includes images along with multilingual captions in English and German. Moreover, the WMT shared tasks (Specia et al., 2016; Elliott et al., 2017; Barrault et al., 2018) have helped in increasing multilingual coverage and task complexity, encouraging deeper research in the field. (Huang et al., 2016) integrated transformed global and regional visual features with the textual input by concatenating them into attendable sequences for the encoder. (Calixto et al., 2017) used a doubly-attentive decoder that attends to both textual and visual representations simultaneously. (Elliott and Kádár, 2017) introduced Imagination, a model that separates multimodal translation into two objectives of learning a translation model, and learning visually grounded representations. As a result of this design, the model can be trained on external datasets containing either parallel text or image descriptions, consequently leveraging a wider range of existing resources.

(a) *English*: He had a nice cold beer out on the deck.
German: Er hatte ein schönes kaltes Bier auf dem Deck.

(b) *English*: He had a nice cold beer out on the deck.
German: Er hatte ein schönes kaltes Bier auf der Veranda.

Figure 1: An example from the CoMMuTE dataset

While multiple papers indicated that the visual modality was either unnecessary or only slightly beneficial, (Caglayan et al., 2019) argued the simplicity, completeness, and repetitiveness of the sentences in the only dataset used for the task (Multi30K) was the main cause. This results in the source text alone being sufficient for context. In order to tackle this, (Futeral et al., 2023) introduced CoMMuTE, a contrastive evaluation set of ambiguous English sentences, each paired with two images and two possible translations in 3 languages, which was later extended to six languages. An example of one data point is presented in Figure 1, where the word *"deck"* can mean 2 different things, but the ambiguity disappears once the images are considered.

There has been some notable work recently that focuses on using multimodal data more strategically, rather than relying on brute-force model training. For instance, (Futeral et al., 2025) proposes the ZeroMMT method, which adapts a strong text-only MT model into a multimodal one without re-

quiring any supervised multimodal training data. This is achieved by leveraging visually conditioned masked language modeling and KL-divergence losses. Furthermore, (Liu et al., 2025) proposes the DeDiT framework, which explicitly detects ambiguity in the source text and triggers visual reasoning only when ambiguity is present. They demonstrate consistent gains over standard MMT models on the CoMMuTE benchmark.

2.2. Vision-Language Models & Multimodal LLMs

With recent advances in vision-language models (VLMs) and multimodal LLMs, new possibilities have emerged for the field. For instance, BLIP-3, introduced by (Xue et al., 2025), presents a new framework for combining the powers of pre-trained vision encoders and LLMs into an end-to-end multimodal system, achieving competitive performance on multimodal tasks. Likewise, the Llama 3(Touvron et al., 2023) family now includes vision capabilities following the release of version 3.2. Similarly, Google's Gemma 3 (Team et al., 2024) series has introduced multimodal capabilities in lightweight, open models optimized for both research and deployment.

Despite their success in multimodal tasks, the use of such LLMs for low-resource languages remains underexplored. Previous works, such as (Dash and Sharma, 2025), have investigated this direction by combining a text-only LLM with BLIP-style vision modules for multimodal translation tasks. Our work builds on these efforts by introducing a structured way to benchmark and improve such LLMs for Indic languages.

2.3. Multimodal Translation in Low-Resource and Indic Contexts

Research in MMT for low-resource or Indic languages has been relatively limited. Hindi Visual Genome (HVG) (Parida et al., 2019) was the first dataset for multimodal English-Hindi machine translation, which was built over the previously available Visual Genome dataset (Krishna et al., 2016). WAT (Workshop on Asian Translation) has frequently used this dataset for its multimodal translation tasks.

In recent years, Indic languages have been getting more focus at WAT and WMT workshop shared tasks, with the inclusion of languages like Hindi, Bengali, Malayalam, etc. in the English-to-Lowres Multimodal Translation Task. One of the notable recent submissions (Rajpoot et al., 2024), proposed a chain-of-thought guided MMT approach using LLMs, which achieved competitive results for Hindi and Bengali. Moreover, (Khan et al., 2025) introduced Chitranuvad, a multimodal model that fuses a multilingual LLM with a vision encoder, achieving strong results for Hindi language.

Some studies have tried to overcome the lack of multimodal data in low-resource languages by synthetic multimodal data generation and cross-modal alignment. For example, in the work (Xiao et al., 2025), the authors propose converting text-only corpora to visually enriched datasets using image-generation models. The generated output is then validated by an LLM to make sure visual content is consistent with textual translations.

2.4. Evaluation & Metrics in MMT

Metrics used to evaluate translation quality include BLEU, ChrF (Popović, 2015), and neural metrics such as COMET (Rei et al., 2020). However, in the multimodal context, these metrics do not account for the contribution of visual information. The CoMMuTE benchmark (Futeral et al., 2023) addresses this limitation by introducing a contrastive scoring framework. Each source sentence in the CoMMuTE evaluation set is paired with two images and their corresponding translations. The model is then asked to rank the likelihood of the two translations separately for each image. A final accuracy score is computed based on the proportion of correctly ranked translation pairs over the total number of examples in the CoMMuTE evaluation set.

3. Methodology

3.1. IndicCoMMuTE

The CoMMuTE (Contrastive Multilingual Multimodal Translation Evaluation) benchmark, introduced by (Futeral et al., 2023), provides a standardized framework to evaluate a system's ability to utilize visual cues during text translation. CoMMuTE focuses on ambiguous sentences, where an accompanied image is required for understanding the context (example in Figure 1). The evaluation set is comprised of English sentences, each of which is grouped with two images, where the meaning of the sentence changes in the context of each image, along with the corresponding translations in different languages. The goal of CoMMuTE is to use the MMT models to rank each of the two translations for sentence-image pairs individually.

In the original work (Futeral et al., 2023), the translations were provided for 3 languages: French, German, and Czech. Later, the work (Futeral et al., 2025) extended to 3 additional languages, Arabic, Russian, and Chinese.

In this paper, we extend the CoMMuTE benchmark with two additional low resource languages - Hindi and Odia, representing Indo-Aryan linguistic

Language	Script	# Samples
German	Latin	300
French	Latin	308
Czech	Latin	308
Russian	Cyrillic	324
Chinese	Han	324
Arabic	Arabic	324
Hindi	Devanagari	300
Odia	Odia	300

Table 1: Language coverage and dataset size for CoMMuTE and IndicCoMMuTE.

families. We refer to this extended version as IndicCoMMuTE [1], with an aim to broaden the scope of multimodal machine translation research beyond high-resource languages. We kept the existing structure of CoMMuTE evaluation dataset intact, where each English sentence is complemented with two translations in Hindi and Odia, respectively, aligning with the contexts represented in the associated images. The English text was translated and reviewed by native speakers of Hindi and Odia languages.

Together, CoMMuTE and IndicCoMMuTE now cover eight languages as shown in Table 1, enabling more comprehensive multilingual evaluation of multimodal translation systems.

Along with expanding the coverage of CoMMuTE benchmark, the introduction of IndicCoMMuTE also provides a systematic way to examine the generalization capabilities of various multimodal systems across typologically distant languages. We intend to publicly release the IndicCoMMuTE dataset in order to facilitate reproducible research and future work on MMT in low-resource languages.

3.2. Models

We choose a representative set of 5 open-source multimodal LLMs with fewer than 15B parameters to benchmark and investigate the translation capabilities of state-of-the-art LLMs in multimodal settings. Each of these models has been tested across eight target languages (German, French, Czech, Russian, Chinese, Arabic, Hindi, and Odia) in both text-only and multimodal settings.

While selecting the models for evaluation, emphasis was placed on their open availability, architectural diversity, and size. To maintain computational feasibility and reproducibility, we refrained from selecting very large proprietary systems and therefore restricted our selection to models with fewer than 15 billion parameters. This threshold strikes an adequate balance between having enough parameters to effectively capture multimodal context across a variety of languages and remaining accessible for resource-constrained inference settings. For a comprehensive comparative analysis, we also present translation scores from a text-only model, which serve as baselines for assessing the multimodal ability of LLMs to leverage visual grounding. Table 2 presents the shortlisted models and their sizes in terms of the number of parameters.

[1] https://github.com/chirag-wadhwa/IndicCoMMuTE

Model	# Parameters
Gemma-3-12b-it	12B
Pixtral-12B-2409	12B
Llama-3.2-11B-Vision	11B
Qwen3-VL-8B-Instruct	8B
Phi-4-multimodal-instruct	5.6B
Nllb-200-3.3B	3.3B

Table 2: Benchmarked models and their parameter sizes.

Parmeter	Value
Training Data Samples	Hindi: 28930 Odia: 28930
LoRA Rank	8
LoRA Alpha	8
Steps	100
Batch Size	64
GPU	1 (Nvidia A100 40 GB)
Training Duration	120 mins

Table 3: LoRA Finetuning parameters.

3.3. Evaluation Metrics

We evaluate the models in both text-only and multimodal settings. The prompts used are as follows:

- **Text-only** — *"Translate the following English sentence to* `{target_language}`*. Only provide the translated text. English sentence:* `{source_sentence}`*."*
- **Multimodal** — *"Translate the following English sentence to* `{target_language}`*, using the provided image for context. Ensure the translation accurately reflects the content and context of the image. Only provide the translated text. English sentence:* `{source_sentence}`*."*

We make use of two complementary metrics for the quantitative assessment of translations by all the selected models: ChrF and COMET. Focusing on multiple aspects of translation performance, this setup provides a holistic understanding of character-level overlap and semantic competence of the translated text.

(a) Wir haben den Trainer gewechselt.

(b) Wir haben die Waggons gewechselt.

Figure 2: Pixtral's **text-only and multimodal German** translations for the source English sentence: ***We changed coaches.*** The model correctly differentiated between the two meanings of coach—a sports trainer and a railway carriage—using visual context.

(a) एक व्यक्ति चट्टान को देख रहा है।

(b) एक व्यक्ति पत्थर को देख रहा है।

Figure 3: Gemma-3-12b's **text-only and multimodal Hindi** translations for the source English sentence: ***A person looking at a rock.*** The model correctly differentiated between the two meanings of rock—a large rock formation and a small stone—using visual context.

CHRF (Character n-gram F-score) (Popović, 2015) is an automatic evaluation metric for machine translation that measures the overlap of character n-grams between system outputs and reference translations, combining precision and recall via the F-score.

COMET (Crosslingual Optimized Metric for Evaluation of Translation) (Rei et al., 2020) is a neural network–based evaluation metric that predicts human judgment scores by comparing source, hypothesis, and reference embeddings from pretrained multilingual models. For the purpose of this study, we use the *wmt22-comet-da* model.

We evaluate the statistical significance of fine-tuning gains using paired bootstrap resampling. For each language, we compute sentence-level paired differences (COMET score) between the fine-tuned and baseline systems. These paired differences are resampled with replacement for 1,000 bootstrap iterations to construct an empirical distribution of mean score differences. From this distribution, we estimate the mean improvement and derive a 95% percentile confidence interval (CI). Each evaluation consists of 300 sentences(same size as IndicCoMMuTe), with sentence-level pairing ensuring that each test instance contributes a single paired difference. Improvements are considered statistically significant when the corresponding 95% CI excludes zero. By avoiding distributional assumptions, this non-parametric approach provides reliable estimates of uncertainty around the observed gains.

3.4. Parameter Efficient Fine-tuning

We fine-tuned the Llama-3.2-11B-Vision model using Low-Rank Adaptation (LoRA)(Hu et al., 2021) to enhance its multimodal translation performance. We specifically focused on the two Indic languages, Hindi and Odia. LoRA fine-tuning was performed to improve the model's multimodal capabilities without changing the base model weights and avoiding overfitting by choosing smaller rank and scaling values. The datasets chosen for fine-tuning were Hindi Visual Genome (HVG) and Odia

(a) Before fine-tuning: क्या आप लाल रंग को नहीं देख सकते हैं?
After fine-tuning: क्या आप लाल बत्ती नहीं देख सकते?

(b) Before fine-tuning: क्या आप लाल रंग को देख सकते हैं?
After fine-tuning: क्या आप लाल रंग की रोशनी नहीं देख सकते?

Figure 4: Llama-3.2-11B-Vision's **before and after fine-tuning Hindi** translations for the source English sentence: ***Can you not see the red light?*** While the model initially failed to distinguish between the two meanings, fine-tuning enabled it to correctly identify the visual differences between the images.

Visual Genome (OVG). The detailed hyperparameters are provided in Table 3.

4. Results and Analysis

4.1. Overall Performance Trends

Table 4 and Table 5 show the highest ChrF and COMET scores for each language were achieved by the multimodal models, as compared to the text-only baseline. Figure 2 illustrates an example of how *Pixtral-12B-2409* is able to successfully capture the visual context in its translation for the source English sentence *"We changed coaches."* Similarly, Figure 3 illustrates an example of *Gemma-3-12b-it* successfully translating the source sentence *"A person looking at a rock."* using the image contexts. While some models showed superior performance in European languages such as German and French, others achieved better results in morphologically distinct languages like Arabic and Russian. Notably, *Pixtral-12B-2409* and *Gemma-3-12B-it* performed competitively across most languages. One major reason for these variations could be imbalances in the size and quality of the language-specific datasets used during model training.

4.2. Performance on Indic Languages

For the newly introduced low-resource Indic languages, Hindi and Odia, included as part of the IndicCoMMuTE, the multimodal models were able to leverage the visual context to disambiguate the source text, as reflected by the higher scores. Among them, *Gemma-3-12B-it* and *Qwen3-VL-8B-Instruct* achieved the best scores. However, overall scores remained lower, particularly for Odia, than those for the other languages. This is partly due to the limited availability of training data for these models and the presence of hallucinated outputs observed in the qualitative analysis. These hallucinations, where the model generated semantically irrelevant or extraneous content, contributed to reduced CHRF and COMET scores.

4.3. Fine-tuning for Low-resource Languages

Llama-3.2-11B-Vision, the model with the lowest average baseline score, was selected for LoRA-based fine-tuning to improve performance. The focus was on translation performance for Indic languages; therefore, the selected datasets were training subset of Hindi Visual Genome and Odia Visual Genome. This led to significant gains in both ChrF and COMET metrics for Hindi and Odia languages, highlighting that lightweight, parameter-efficient adaptation can substantially enhance multimodal translation quality with minimal compute and data requirements. Figure 4 illustrates how the model was better able to leverage visual context after fine-tuning for the source sentence *"Can you not see the red light?"*

Model	Arabic	Czech	German	French	Russian	Chinese	Hindi	Odia
Text only								
Nllb-200-3.3B	43.37	51.80	58.18	59.65	43.86	20.03	48.67	51.23
Pixtral-12B-2409	25.16	36.61	55.97	57.40	40.98	25.35	34.56	13.57
Qwen3-VL-8B-Instruct	36.31	41.28	54.70	57.33	**43.11**	**27.99**	39.22	**27.65**
Gemma-3-12b-it	**40.94**	**49.19**	**58.64**	**59.15**	42.74	23.75	**47.37**	25.64
Phi-4-multimodal-instruct	30.19	32.24	49.32	54.11	38.66	23.72	30.36	1.69
Llama-3.2-11B-Vision	33.19	40.46	52.03	54.22	37.67	19.49	39.30	11.40
Multimodal (Zero-shot Prompt)								
Pixtral-12B-2409	31.96	40.62	**61.29**	**66.31**	**48.09**	29.22	37.30	13.47
Qwen3-VL-8B-Instruct	36.53	44.31	58.77	61.22	46.53	**29.26**	40.35	**27.83**
Gemma-3-12b-it	**41.62**	**48.56**	58.89	61.51	47.95	20.93	**42.18**	25.21
Phi-4-multimodal-instruct	22.95	29.76	48.09	54.79	37.49	21.20	28.94	0.85
Llama-3.2-11B-Vision	17.97	28.87	42.30	50.18	28.41	13.05	23.24	0.20

Table 4: ChrF scores for En→X translation models across multiple languages. Best scores in each column are highlighted.

Model	Arabic	Czech	German	French	Russian	Chinese	Hindi	Odia
Text only								
Nllb-200-3.3B	80.17	84.91	81.33	81.41	82.23	76.89	73.76	82.03
Pixtral-12B-2409	67.06	73.27	81.19	78.84	79.84	79.24	66.78	44.85
Qwen3-VL-8B-Instruct	76.27	80.58	79.80	79.37	80.81	**80.97**	70.36	**65.12**
Gemma-3-12b-it	**78.10**	**83.49**	**81.72**	**80.22**	**81.63**	79.67	**77.09**	62.46
Phi-4-multimodal-instruct	69.49	72.68	77.49	78.36	77.19	79.15	63.57	41.61
Llama-3.2-11B-Vision	72.43	77.25	78.93	78.44	77.73	78.38	71.60	53.77
Multimodal (Zero-shot Prompt)								
Pixtral-12B-2409	72.57	77.01	**85.28**	**85.47**	**85.53**	85.12	69.00	44.39
Qwen3-VL-8B-Instruct	77.65	82.62	83.80	82.76	84.76	**86.40**	72.03	**70.40**
Gemma-3-12b-it	**80.33**	**85.80**	84.20	84.46	85.44	83.21	**73.10**	65.89
Phi-4-multimodal-instruct	67.17	67.24	78.63	79.27	78.68	82.23	61.73	40.28
Llama-3.2-11B-Vision	61.10	69.36	73.28	77.14	72.57	76.89	57.08	37.87

Table 5: COMET scores for En→X translation models across multiple languages. Best scores in each column are highlighted.

	Llama-3.2-11B-Vision	
Language	Baseline	Fine-tuned
Hindi	57.08	69.88
Odia	37.87	61.68
German	73.28	79.97

Table 6: COMET Scores for vanilla and finetuned model.

Table 6 presents the COMET scores for *Llama-3.2-11B-Vision*, before and after fine-tuning. The fine-tuning improvements for Hindi and Odia are statistically significant under paired bootstrap resampling, with 95% confidence intervals excluding zero.

4.4. Human Evaluation

4.4.1. Qualitative Analysis of Multimodal Translation Outputs

To complement automatic evaluation metrics, we performed a qualitative analysis of translations generated by several multimodal LLMs alongside a text-only NMT baseline. The analysis focused on aspects that are difficult to capture through automatic metrics, such as visual disambiguation, hallucinations, script correctness, and translation fluency. Overall, multimodal models demonstrate the ability to use visual context to resolve certain lexical ambiguities that text-only systems typically fail to distinguish. For instance, different image contexts often led to different translations for ambiguous words such as bank, spring, and glasses, indicating that visual grounding can guide translation decisions. At the same time, the analysis

reveals several limitations. Performance varies considerably across models and languages, with translations into high-resource languages generally appearing more stable and fluent. In contrast, translations into low-resource Indic languages are more prone to issues such as script inconsistencies, transliteration of English words, and occasional hallucinated or malformed outputs. These observations highlight both the potential of multimodal grounding for translation and the challenges that remain in supporting linguistically diverse settings.

4.4.2. Impact of Parameter-Efficient Fine-Tuning

We also examined the effect of lightweight LoRA fine-tuning on the *Llama-3.2-11B-Vision* model using multimodal training data in Hindi and Odia. The qualitative analysis indicates that fine-tuning substantially improves the reliability and formatting of the model's outputs, particularly for low-resource Indic languages. After fine-tuning, translations appear more consistently in the correct script and exhibit far fewer generation artifacts such as verbose explanations, image descriptions, or repetitive output patterns. The model also shows improved stability when generating translations in these languages, producing outputs that are generally more coherent and usable. In addition, modest improvements in visually grounded translation behavior can be observed in some cases, suggesting that the fine-tuning process helps the model better align visual context with translation decisions. Interestingly, some improvements are also visible in languages that were not included in the fine-tuning data, indicating that the adaptation may influence the model's broader multimodal generation behavior. Overall, these observations suggest that parameter-efficient fine-tuning can meaningfully improve the usability of multimodal translation systems, particularly for languages that are underrepresented in the original training data.

4.5. Cross-lingual Generalization

We also investigated whether fine-tuning on the Indic datasets led to any cross-lingual transfer and improved instruction following. Notably, when the fine-tuned *Llama-3.2-11B-Vision* model was evaluated on English–German multimodal translation, it achieved higher ChrF and COMET scores compared to the baseline, indicating that the fine-tuning process improved the model's multimodal instruction following and general visual–semantic understanding rather than language-specific representations. Table 6 presents the COMET scores for *Llama-3.2-11B-Vision* for German, before and after fine-tuning. IndicCoMMuTE currently includes two Indic languages and 300 contrastive items per language, limiting broad generalization claims across the diverse Indic language families.

5. Conclusion and Future Work

Through this work, we explored the impact of visual cues on multimodal machine translation across eight languages using the CoMMuTE and IndicCoMMuTE evaluation sets. We compared a baseline text-only model with 5 LLM-based multimodal architectures, analyzed behavior for two low-resource Indic languages, and examined cross-lingual generalization effects. Based on our findings, we summarize the key conclusions as follows:

- The multimodal models consistently outperformed the text-only baselines, demonstrating the influence of visual context in disambiguating ambiguous texts.
- Translation performance varied across languages, largely impacted by the amount and quality of training data available for each language.
- For the low-resource Indic languages introduced in this study - Hindi and Odia, Gemma 3 (12B) and Qwen 3 (8B) achieved the highest scores.
- With the LoRA-based fine-tuning of the *Llama-3.2-11B-Vision* model on the HVG and OVG datasets, a substantial gain in translation accuracy was observed for both the Indic languages, demonstrating that even lightweight tuning can enhance instruction following and multimodal understanding.
- *Llama-3.2-11B-Vision*, fine-tuned for Hindi and Odia, also showed improved performance on English-German multimodal translation, suggesting that fine-tuning aids cross-lingual transfer of visual-semantic knowledge.
- Overall, incorporating visual context combined with efficient fine-tuning can lead to improved translation, particularly for low-resource languages.

Building on the findings of this study, we identify several promising directions for future research:

- We plan to fine-tune other multimodal models, such as Pixtral, Gemma, and Qwen-VL, to evaluate whether the observed gains from LoRA-based adaptation generalize across architectures and model scales.

- We plan to extend the IndicCoMMuTE benchmark to include additional Indic languages, particularly Dravidian languages, and other multimodal LLMs with the objective of broadening coverage and advancing research in multimodal translation for underrepresented linguistic domains.

- We plan to focus on mitigating hallucinations through improved image–text alignment objectives and increasing the volume of fine-tuning dataset.

6. Bibliographical References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. Neural machine translation by jointly learning to align and translate.

Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. 2018. Findings of the third shared task on multimodal machine translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 304–323, Belgium, Brussels. Association for Computational Linguistics.

Ozan Caglayan, Pranava Madhyastha, Lucia Specia, and Loïc Barrault. 2019. Probing the need for visual context in multimodal machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4159–4170, Minneapolis, Minnesota. Association for Computational Linguistics.

Iacer Calixto, Qun Liu, and Nick Campbell. 2017. Doubly-attentive decoder for multi-modal neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1913–1924, Vancouver, Canada. Association for Computational Linguistics.

Amulya Dash and Yashvardhan Sharma. 2025. A collaborative approach to multimodal machine translation: Vlm and llm. In *Proceedings of the 17th International Conference on Agents and Artificial Intelligence - Volume 3: ICAART*, pages 1412–1418. INSTICC, SciTePress.

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

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

Desmond Elliott and Ákos Kádár. 2017. Imagination improves multimodal translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 130–141, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation.

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

Matthieu Futeral, Cordelia Schmid, Benoît Sagot, and Rachel Bawden. 2025. Towards zero-shot multimodal machine translation.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Po-Yao Huang, Frederick Liu, Sz-Rung Shiang, Jean Oh, and Chris Dyer. 2016. Attention-based multimodal neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 639–645, Berlin, Germany. Association for Computational Linguistics.

Shaharukh Khan, Ayush Tarun, Ali Faraz, Palash Kamble, Vivek Dahiya, Praveen Pokala, Ashish Kulkarni, Chandra Khatri, Abhinav Ravi, and Shubham Agarwal. 2025. Chitranuvad: Adapting multi-lingual llms for multimodal translation.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations.

Danyang Liu, Fanjie Kong, Xiaohang Sun, Dhruva Patil, Avijit Vajpayee, Zhu Liu, Vimal Bhat, and Najmeh Sadoughi. 2025. Detect, disambiguate, and translate: On-demand visual reasoning for multimodal machine translation with large vision-language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1559–1570, Albuquerque, New Mexico. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Shantipriya Parida, Ondřej Bojar, and Satya Ranjan Dash. 2019. Hindi visual genome: A dataset for multimodal english-to-hindi machine translation.

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

Pawan Rajpoot, Nagaraj Bhat, and Ashish Shrivastava. 2024. Multimodal machine translation for low-resource Indic languages: A chain-of-thought approach using large language models. In *Proceedings of the Ninth Conference on Machine Translation*, pages 833–838, Miami, Florida, USA. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Lucia Specia, Stella Frank, Khalil Sima'an, and Desmond Elliott. 2016. A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 543–553, Berlin, Germany. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open models based on gemini research and technology.

NLLB Team. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Bushi Xiao, Qian Shen, and Daisy Zhe Wang. 2025. From text to multi-modal: Advancing low-resource-language translation through synthetic data generation and cross-modal alignments. In *Proceedings of the Eighth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2025)*, pages 24–35, Albuquerque, New Mexico, U.S.A. Association for Computational Linguistics.

Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S Ryoo, Shrikant Kendre, Jieyu Zhang, Shaoyen Tseng, Gustavo A Lujan-Moreno, Matthew L Olson, Musashi Hinck, David Cobbley, Vasudev Lal, Can Qin, Shu Zhang, Chia-Chih Chen, Ning Yu, Juntao Tan, Tulika Manoj Awalgaonkar, Shelby Heinecke, Huan Wang, Yejin Choi, Ludwig Schmidt, Zeyuan Chen, Silvio Savarese, Juan Carlos Niebles, Caiming Xiong, and Ran Xu. 2025. xgen-mm (blip-3): A family of open large multimodal models.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis.