Does Vision Still Help? Multimodal Translation with CLIP-Based Image Selection

Deepak Kumar[*], **Baban Gain**[*], **Kshetrimayum Boynao Singh** and **Asif Ekbal**
Dept. of Computer Science and Engineering, Indian Institute of Technology Patna, India
{deepakkumar1538, gainbaban, boynfrancis}@gmail.com, asif@iitp.ac.in

Abstract

Multimodal Machine Translation aims to enhance conventional text-only translation systems by incorporating visual context, typically in the form of images paired with captions. In this work, we present our submission to the WAT 2025 Multimodal Translation Shared Task, which explores the role of visual information in translating English captions into four Indic languages: Hindi, Bengali, Malayalam, and Odia. Our system builds upon the strong multilingual text translation backbone *IndicTrans*, augmented with a CLIP-based selective visual grounding mechanism. Specifically, we compute cosine similarities between text and image embeddings (both full and cropped regions) and automatically select the most semantically aligned image representation to integrate into the translation model. We observe that overall contribution of visual features is questionable. Our findings reaffirm recent evidence that large multilingual translation models can perform competitively without explicit visual grounding.

1 Introduction

Multimodal Machine Translation (MMT) extends traditional text-only translation by incorporating auxiliary visual information typically an image paired with the source sentence. The motivation behind this integration is that images can provide crucial contextual clues that help resolve linguistic ambiguities and improve translation accuracy. For example, consider the English sentence "The man is standing near the court." Without additional context, the word "court" could refer to a sports court (e.g., tennis or basketball), or a legal court. A text-only translation model may incorrectly choose one sense based solely on linguistic priors. However, if the corresponding image depicts a tennis court, the visual cue instantly clarifies the intended meaning, guiding the model toward the correct translation in the target language. This exemplifies how visual grounding can disambiguate polysemous words that textual context alone may not fully resolve.

Although several studies have shown that incorporating image information improves translation performance, most prior work trains their MMT models from scratch, learning both textual and visual representations jointly. These models often report improvements over text-only Neural Machine Translation (NMT) systems trained under similar conditions. However, while the relative gains appear significant, the absolute translation scores remain low compared to strong pretrained text-only baselines. Moreover, in many benchmark datasets, intra-sentence textual context is already sufficient to produce correct translations, reducing the actual necessity of visual input. Consequently, it remains unclear whether the observed improvements truly arise from visual grounding or from differences in model training setups.

Another source of debate in MMT lies in the choice of visual input. Given an image, its caption, and a cropped version of the image focused specifically on the captioned region, should the model use the full image or only the cropped area? The full image may offer richer contextual information but might also introduce irrelevant details. Conversely, the cropped image may better correspond to the caption but risk losing broader scene semantics.

To address this challenge, we propose a selective visual alignment approach that automatically chooses the most relevant visual representation for translation. Specifically, we extract CLIP embeddings from both the full and cropped versions of each image and compute their cosine similarity with the corresponding text embedding. The image version that exhibits higher textual similarity is selected and passed to the translation system. Our MMT model integrates these CLIP-based features through a Selective Attention mechanism, which performs cross-attention between the image and text representations, allowing the model to focus on visually aligned information.

We use IndicTrans as our base model a strong pretrained multilingual translation system covering multiple Indic languages such as Hindi, Bengali, Malayalam, and Odia. Interestingly, while our ap-

[*] Equal contribution.

Proceedings of the Twelfth Workshop on Asian Translation (WAT 2025), pages 115–123
December 24, 2025 ©2025 Association for Computational Linguistics

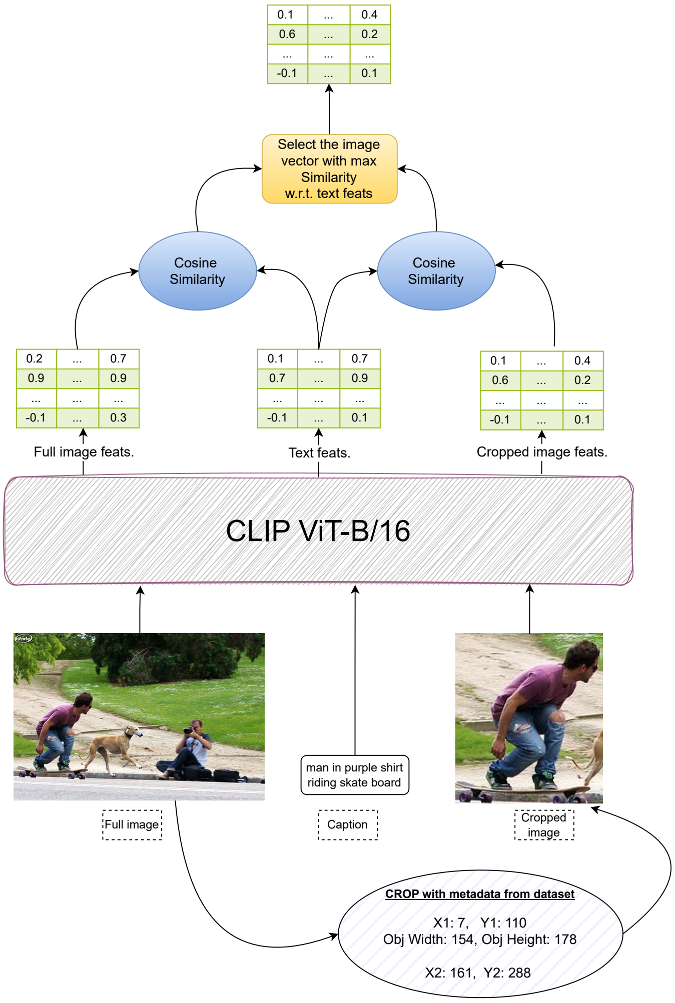

Figure 1: Flow diagram illustrating the image–text similarity computation using CLIP ViT-B/16. The model compares full-image and cropped-image features with text embeddings via cosine similarity, selecting the image vector with the highest similarity score; it is further forwarded to the translation model

proach achieves high absolute translation quality, we find that incorporating visual features does not consistently improve results compared to the text-only baseline. This observation is consistent with previous findings in the literature. Prior work (Li et al., 2022) has also questioned the real contribution of visual information in multimodal translation systems. In this paper, we present the observations in the WAT 2025 Multimodal Translation Task, aiming to further investigate this phenomenon in a competitive shared-task setting. Our system builds upon the Selective Attention architecture (Li et al., 2022; Gain et al., 2025), which effectively integrates visual features into a Transformer-based translation framework. We extend this system in two key ways: (i) we retrain the model to additionally support the *Odia* language, thereby expanding its coverage, and (ii) we incorporate an image selection mechanism that compares the cosine similarity between text embeddings and CLIP features extracted from both the full and cropped versions of each image, forwarding the representation with higher textual alignment. This selective image integration allows the model to better exploit visual cues when relevant, while avoiding unnecessary noise from less informative image regions.

2 Related Works

Multimodal Neural Machine Translation (MMT) seeks to integrate both textual and visual modalities in order to improve translation quality-particularly by helping to disambiguate linguistic phenomena or provide grounding beyond the source text. Early pioneering studies investigated the use of image features (often extracted from convolutional neural networks) alongside an encoder–decoder architecture with attention over both text and image features (Elliott et al., 2016),(Calixto et al., 2017). It was also been observed thaty that MMT systems could leverage visual input under conditions of degraded textual context, but that gains were modest when textual input alone was sufficient (Caglayan et al., 2019).

Subsequent research questioned the actual utility of the visual modality in standard benchmarks, noting that when images were replaced by mismatched or random images, model performance often did not degrade significantly. The authors in (Li et al., 2021) highlighted that existing MMT datasets and architectures might encourage models to ignore the image input altogether. Related work also explored the integration of visual features via fused or hierarchical attention mechanisms (Yao and Wan, 2020) and in low-resource scenarios where the textual signal is weaker.

Multimodal translation in Indian languages has been underexplored, with most studies focusing on the English–Hindi pair. The majority of these works are adaptations of architectures originally designed for high-resource settings.

The earliest work on integrating visual information into Indian language translation can be traced to the approach proposed in (Laskar et al., 2020), which utilized a doubly attentive decoder capable of simultaneously attending to both textual and visual modalities. This model was later refined in (Laskar et al., 2021) through additional text-only pre-training on the IITB parallel corpus (Kunchukuttan et al., 2018) and data augmentation using phrase pairs generated with the Giza++ tool (Marchisio et al., 2022). The visual representations were obtained using a pre-trained VGG19 network (Simonyan and Zisserman, 2015). The same framework was subsequently extended to the English–Bengali language pair in (Laskar et al., 2022), where the model achieved BLEU scores of 43.90 and 28.70 on the Test and Challenge sets, respectively.

Following these early studies, the work presented in (Gupta et al., 2021) introduced an alternative strategy that enriched textual input with object-level visual cues. An object detection model was employed to identify entities within the image, and their class labels were appended to the source sentence to provide additional semantic context. The system, built upon mBART (Liu et al., 2020), achieved state-of-the-art performance for English–Hindi translation; however, the improvement was primarily attributed to large-scale pretraining rather than genuine multimodal fusion. Specifically, the model exhibited a modest gain of +0.52 BLEU on the standard test set while showing a slight decline of 0.06 BLEU on the Challenge set. A subsequent extension of this framework to English–Bengali and English–Malayalam translation was reported in (Parida et al., 2022), yielding comparable trends.

More recent work in (Gain et al., 2021) explored a multimodal transformer architecture for English–Hindi translation. The study demonstrated that focusing on cropped regions of the image corresponding to the textual referents produced more accurate translations than utilizing full-image features. Later, the methodology was revisited in (Shi and Yu, 2022), which introduced refined preprocessing steps such as the removal of duplicate and grayscale images. By employing ResNet50-based features (He et al., 2015) and optimized hyperparameters, this system achieved BLEU scores of 42.29 and 42.70 on the Test and Challenge sets, respectively, highlighting the significant role of data quality and preprocessing in multimodal translation performance.

Overall, while these studies represent significant steps toward integrating visual information in Indian language translation, they collectively indicate that the performance gains from multimodality remain limited. Most improvements appear to arise from better pre-training and data curation rather than from truly leveraging visual grounding.

3 Methodology

3.1 Datasets

The dataset used in this work is part of the WAT 2025 Multimodal Translation shared task (Parida et al., 2024) and is designed to facilitate research on multimodal translation between English and multiple Indic languages. Each data instance comprises an **English caption** paired with its **reference translations** in four target languages: *Hindi* (Parida and Bojar, 2020) , *Bengali* (Sen et al., 2022), *Malayalam* (Parida and Bojar, 2021), and *Odia* (Parida et al., 2025).

In addition to the text pairs, each example is associated with an **image** that visually represents the described scene. To support fine-grained visual

Subset	Sentences	Avg. Src (en)	Avg. Tgt Words			
			hi	bn	ml	or
train	28930	4.95	5.03	3.94	3.70	4.90
test	1595	4.92	4.92	4.02	3.57	4.85
valid	998	4.93	4.99	3.94	3.63	4.92
challenge	1400	5.85	6.17	4.76	4.32	5.79

Table 1: Statistics of the multilingual parallel datasets showing the number of sentences and average word counts for English source and four target languages.

grounding, the dataset also provides **bounding box coordinates** corresponding to the region of interest (ROI) within the image that the caption explicitly refers to.

The multimodal nature of this dataset allows translation models to learn both linguistic mappings and visual alignments, thereby grounding the translation process in contextual image information. Table 1 presents the detailed statistics of the dataset, including the number of sentence pairs and the average word counts for the English source and the four target languages. The corpus contains approximately **29K training examples**, along with dedicated **validation**, **test**, and **challenge** subsets to facilitate comprehensive evaluation.

This multimodal setup provides a valuable benchmark for assessing whether and how visual information contributes to disambiguating textual input during translation, particularly in resource-constrained Indic language settings.

3.2 Experimental Setup

For the multimodal experiments, we enrich the textual input with visual representations extracted from the images paired with the parallel data. To obtain these representations, we use CLIP ViT-B/16 encoder to compute fixed-dimensional embeddings for every image. The encoder outputs a 512-dimensional feature vector that captures high-level semantic attributes relevant for translation. These features are pre-computed offline to avoid additional computational overhead during model training. All models are trained using the Fairseq (Ott et al., 2019) framework and adapted from (Gain et al., 2025) [1], following a consistent configuration across both text-only and multimodal settings to facilitate controlled comparison. Training is carried out using the inverse square root learning rate schedule with 4,000 warm-up steps, Adam optimization with $\beta_1 = 0.9$ and $\beta_2 = 0.98$, and label-smoothed cross-entropy with a smoothing factor of 0.1. We constrain the maximum source and target lengths to 210 tokens and train for up to 20,000 updates with mixed-precision (FP16). A dropout rate of 0.3 is applied uniformly across the model, and early stopping is triggered based on validation loss with a patience of 5 epochs. For the multimodal system, the base Transformer architecture is augmented with an image-aware fusion module, where the pre-extracted CLIP features are injected into the encoder using a shallow self-attention fusion layer. Additional dropout is applied to the image features and the fusion attention to enhance robustness. Apart from this cross-modal extension, all hyperparameters remain consistent with the text-only baseline. Decoding is performed using beam search with a beam width of 5 and a batch size of 64. All predictions undergo standard post-processing, including the removal of subword segmentation markers.

3.3 Visual Feature Extraction

We employ CLIP (Radford et al., 2021), a vision–language model developed by OpenAI, to integrate visual grounding into our translation framework. CLIP learns joint representations of images and text through contrastive learning on large-scale image–caption pairs, mapping both modalities into a shared embedding space where semantic similarity can be effectively measured. This allows the model to capture fine-grained correspondences between visual and linguistic concepts.

In our setup, each data instance contains both a full image and a cropped region specified by bounding box coordinates provided in the dataset. We use CLIP to extract embeddings for both these variants one representing the global context and the other focusing on the region of interest. The text caption accompanying the image is also encoded using CLIP's text encoder, resulting in a dense semantic representation. To determine which visual

[1] https://github.com/babangain/indicMMT/

Model Name	Image Used	Eval. Dataset	Bengali	Hindi	Malayalam	Odia	Average
Textual finetune	✗	Eval Set	49.50	45.40	51.20	64.30	52.60
Multimodal finetune	✓	Eval Set	48.70 (-0.80)	44.90 (-0.50)	50.70 (-0.50)	63.50 (-0.80)	51.95 (-0.65)
Textual finetune	✗	Challenge Set	47.50	56.10	40.30	55.40	49.83
Multimodal finetune	✓	Challenge Set	47.00 (-0.50)	56.60 (+0.50)	38.90 (-1.40)	55.20 (-0.20)	49.43 (-0.40)

Table 2: BLEU Score of our models on different Indic languages from WAT evaluations.

Model Name	Image Used	Eval. Dataset	Bengali	Hindi	Malayalam	Odia	Average
Textual finetune	✗	Eval Set	80.17	83.50	76.08	90.65	82.60
Multimodal finetune	✓	Eval Set	79.97	83.08	76.55	90.36	82.49
Textual finetune	✗	Challenge Set	81.97	87.09	75.73	91.68	84.12
Multimodal finetune	✓	Challenge Set	81.54	87.22	74.94	91.60	83.83

Table 3: RIBES Score of our models on different Indic languages from WAT evaluations.

variant best aligns with the caption, we compute two cosine similarity scores: (a) between the text embedding and the full-image embedding, and (b) between the text embedding and the cropped-image embedding. Since it is not known a priori which of the two visual representations (global or localized) provides more relevant contextual cues, we adopt a simple yet effective heuristic selecting the image feature that yields the higher similarity score with the text. This strategy allows the system to automatically adapt to the most semantically aligned visual cue for each instance, ensuring that the translation model attends to the most meaningful image content while ignoring irrelevant background noise. The overall CLIP-based image selection process is illustrated in Figure 1.

3.4 IndicTrans

IndicTrans (Ramesh et al., 2023) is a multilingual neural machine translation model designed for translation between English and multiple Indic languages. It is trained on large-scale parallel corpora and optimized for high-quality translation across diverse language pairs such as Hindi, Bengali, Malayalam, and Odia. Leveraging a transformer-based architecture and multilingual pretraining, IndicTrans achieves strong performance even in low-resource scenarios, making it a robust baseline for multilingual and multimodal translation research. In this work, we adopt IndicTrans as the underlying translation backbone due to its strong pretraining across Indic languages and its ability to provide robust sentence-level representations and cross-lingual transfer capabilities, making it a suitable foundation for multimodal extensions.

3.5 Model Architecture

We use the Selective Attention architecture (Li et al., 2022) for incorporating visual information into our multimodal translation framework. The model combines the visual features extracted as described in Section 3.3 with the pretrained *IndicTrans* model detailed in Section 3.4. This architecture enables fine-grained alignment between image regions and text tokens through a combination of gated fusion and selective attention mechanisms.

Formally, let the textual input sequence be X_{text} and the corresponding image (either full or cropped, selected via the CLIP-based mechanism) be X_{img}. The *IndicTrans* encoder processes the source text to obtain the hidden representation:

$$H_{\text{text}} = \text{TransformerEncoder}(X_{\text{text}}), \qquad (1)$$

while the visual encoder (e.g., ViT) produces image representations:

$$H_{\text{img}} = W \cdot \text{ViT}(X_{\text{img}}), \qquad (2)$$

where W is a projection matrix that matches the dimensionality of image and text features.

Following Li et al. (2022), a gated fusion mechanism is used to control the relative contribution of the two modalities:

$$\lambda = \sigma(UH_{\text{text}} + VH_{\text{img}}), \qquad (3)$$

$$H_{\text{out}} = (1 - \lambda) \odot H_{\text{text}} + \lambda \odot H_{\text{img}}, \qquad (4)$$

where U and V are trainable parameters and σ is the sigmoid activation. The gating variable λ regulates the degree to which visual information influences the textual representation, allowing adaptive fusion based on semantic relevance.

Image			
Source	date when taken in yellow	knife block sitting on counter with knives in it	player running on court
Ground Truth	তারিখ যখন হলুদ নেওয়া হয়	चाकू ब्लॉक में चाकू लेकर काउंटर पर बैठे	कोर्ट पर दौड़ता हुआ खिलाड़ी
Unimodal	তারিখ যখন হলুদ নেওয়া হয়	चाकू ब्लॉक में चाकू लेकर काउंटर पर बैठे	खिलाड़ी कोर्ट पर चल रहा है
Multimodal	হলুদ রঙের সময় তারিখ	इसमें चाकू के साथ काउंटर पर बैठे चाकू ब्लॉक	कोर्ट पर दौड़ता हुआ खिलाड़ी
Explanation	The original text conveys that the date when the photo was taken is depicted in yellow. However, the reference is not reflecting this. Although the “text+image” translation meaning is somewhat accurate, the real improvement is not captured as due to reference.	Unimodal: Awkward and incorrect. Multimodal: Clearer and closer to ground truth, just slightly verbose.	Incorrect verb, says “walking” instead of “running” in case of unimodal

Figure 2: Examples of outputs from unimodal and multimodal model. The major improvements are generally from grammatical issues.

To capture localized visual textual correspondences, the model further applies a Selective Attention layer that correlates textual queries with image patches:

$$H_{\text{img}}^{\text{attn}} = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{5}$$

where Q is derived from H_{text}, and K and V are obtained from H_{img}. The attention weighted visual representation $H_{\text{img}}^{\text{attn}}$ is subsequently used in the gated fusion equations above, ensuring that the model focuses on semantically relevant visual regions while down-weighting background noise.

The resulting fused representation H_{out} is then provided to the *IndicTrans* decoder for translation generation. In summary, our framework directly employs the Selective Attention architecture (Li et al., 2022), integrating it with *IndicTrans* and CLIP-based image selection to ground translation in visually relevant content.

4 Results and Analysis

We evaluated our proposed CLIP-based multimodal translation approach on the English→Indic Multimodal Translation Task using four target languages: Bengali, Hindi, Malayalam and Odia. The results are reported in terms of BLEU and RIBES scores on both the Eval and *Challenge* sets, as shown in Table 2 and Table 3. The *Textual finetune* models correspond to the IndicTrans baseline trained purely on text, while the *Multimodal finetune* models integrate visual features selected through our CLIP-based image–text similarity mechanism.

4.1 Quantitative Evaluation

The BLEU results (Table 2) show that the text-only IndicTrans baseline achieves strong performance across all languages, with average BLEU scores of 52.60 on the *Eval Set* and 49.83 on the *Challenge Set*. Incorporating visual information through CLIP-based multimodal fine-tuning yields small but consistent variations across languages. On the *Eval Set*, multimodal finetuning slightly decreases the average BLEU by 0.65 points, while on the *Challenge Set*, it results in a marginal average reduction of 0.40 points. Interestingly, Hindi demonstrates a minor improvement (+0.50 BLEU) under noisy or out-of-domain conditions, suggesting that visual grounding may be helpful when textual cues are ambiguous or degraded.

For other languages, the observed differences remain within ±1 BLEU, which aligns with prior findings that visual information contributes weakly to translation quality when the text provides sufficient context. Malayalam and Odia, in particular, show small declines, possibly due to the limited correlation between the visual content and sentence semantics in the dataset, leading to minor noise introduction during fusion.

The RIBES results (Table 3) mirror these trends. The textual baseline achieves an average RIBES of 82.60 and 84.12 on the *Eval* and *Challenge* sets, respectively. The multimodal variants record comparable averages of 82.49 and 83.83, indicating no statistically significant degradation. These consistent RIBES values suggest that the inclusion of visual embeddings does not disrupt sentence-level reordering or fluency, even though it provides limited benefits to lexical adequacy.

4.2 Cross-Language Observations

Among all Indic languages, Hindi exhibits the most stable and slightly positive response to multimodal cues, showing improvements in both BLEU (+0.50) and RIBES (+0.13) on the Challenge Set. This is likely due to Hindi's richer contextual grounding in the shared training corpus and its relatively better alignment with English sentence structures. In contrast, Malayalam shows the largest negative shift, consistent with its morphological complexity and looser syntactic alignment, which may hinder effective multimodal fusion.

5 Conclusion

This work presented a systematic investigation of the impact of visual information in multilingual MMT for Indic languages. Building upon the strong text-only IndicTrans model, we proposed a CLIP-based selective visual grounding mechanism that dynamically identifies the most semantically aligned image representation between the full and cropped variants. We observed that visual grounding offers limited gains in translation quality compared to strong text-only baselines. While the absolute BLEU and RIBES scores remain competitive across all languages, the improvements from multimodal finetuning are modest and often lower than text-only model. These findings are consistent with recent studies questioning the necessity of visual input in multimodal translation, particularly when models are pretrained on large-scale textual corpora.

Acknowledgement

The authors express gratitude to the COIL-D Project under Bhashini, funded by MeitY, Govt. of India, for the support and resources that facilitated the successful execution of this research.

References

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.

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.

Baban Gain, Dibyanayan Bandyopadhyay, and Asif Ekbal. 2021. Experiences of adapting multimodal machine translation techniques for Hindi. In *Proceedings of the First Workshop on Multimodal Machine Translation for Low Resource Languages (MMTLRL 2021)*, pages 40–44, Online (Virtual Mode). INCOMA Ltd.

Baban Gain, Dibyanayan Bandyopadhyay, Samrat Mukherjee, Chandranath Adak, and Asif Ekbal. 2025. Impact of visual context on noisy multimodal nmt: An empirical study for english to indian languages. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 24(8).

Kshitij Gupta, Devansh Gautam, and Radhika Mamidi. 2021. ViTA: Visual-linguistic translation by aligning object tags. In *Proceedings of the 8th Workshop* *on Asian Translation (WAT2021)*, pages 166–173, Online. Association for Computational Linguistics.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*.

Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. 2018. The IIT Bombay English-Hindi parallel corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Sahinur Rahman Laskar, Pankaj Dadure, Riyanka Manna, Partha Pakray, and Sivaji Bandyopadhyay. 2022. English to Bengali multimodal neural machine translation using transliteration-based phrase pairs augmentation. In *Proceedings of the 9th Workshop on Asian Translation*, pages 111–116, Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Sahinur Rahman Laskar, Abdullah Faiz Ur Rahman Khilji, Darsh Kaushik, Partha Pakray, and Sivaji Bandyopadhyay. 2021. Improved English to Hindi multimodal neural machine translation. In *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*, pages 155–160, Online. Association for Computational Linguistics.

Sahinur Rahman Laskar, Abdullah Faiz Ur Rahman Khilji, Partha Pakray, and Sivaji Bandyopadhyay. 2020. Multimodal neural machine translation for English to Hindi. In *Proceedings of the 7th Workshop on Asian Translation*, pages 109–113, Suzhou, China. Association for Computational Linguistics.

Bei Li, Chuanhao Lv, Zefan Zhou, Tao Zhou, Tong Xiao, Anxiang Ma, and JingBo Zhu. 2022. On vision features in multimodal machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6327–6337, Dublin, Ireland. Association for Computational Linguistics.

Jiaoda Li, Duygu Ataman, and Rico Sennrich. 2021. Vision matters when it should: Sanity checking multimodal machine translation models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8556–8562, Online and Punta Cana, Dominican Republic. 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.

Kelly Marchisio, Conghao Xiong, and Philipp Koehn. 2022. Embedding-enhanced giza++: Improving alignment in low- and high- resource scenarios using embedding space geometry. *Preprint*, arXiv:2104.08721.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Shantipriya Parida and Ondřej Bojar. 2020. Hindi visual genome 1.1. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).

Shantipriya Parida and Ondřej Bojar. 2021. Malayalam visual genome 1.0. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).

Shantipriya Parida, Ondřej Bojar, Idris Abdulmumin, Shamsuddeen Hassan Muhammad, and Ibrahim Said Ahmad. 2024. Findings of WMT2024 English-to-low resource multimodal translation task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 677–683, Miami, Florida, USA. Association for Computational Linguistics.

Shantipriya Parida, Subhadarshi Panda, Stig-Arne Grönroos, Mark Granroth-Wilding, and Mika Koistinen. 2022. Silo NLP‘s participation at WAT2022. In *Proceedings of the 9th Workshop on Asian Translation*, pages 99–105, Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Shantipriya Parida, Shashikanta Sahoo, Kalyanamalini Sahoo, Ondřej Bojar, and Satya Ranjan Dash. 2025. Odia visual genome. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2023. Samanantar: The largest publicly available parallel corpora collection for 11 indic languages. *Preprint*, arXiv:2104.05596.

Arghyadeep Sen, Shantipriya Parida, Ketan Kotwal, Subhadarshi Panda, Ondřej Bojar, and Satya Ranjan Dash. 2022. Bengali visual genome: A multimodal dataset for machine translation and image captioning. In *Intelligent Data Engineering and Analytics*, pages 63–70, Singapore. Springer Nature Singapore.

Xiayang Shi and Zhenqiang Yu. 2022. Adding visual information to improve multimodal machine translation for low-resource language. *Mathematical Problems in Engineering*, 2022.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. *Preprint*, arXiv:1409.1556.

Shaowei Yao and Xiaojun Wan. 2020. Multimodal transformer for multimodal machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4346–4350, Online. Association for Computational Linguistics.