Journal of Information Systems and Telecommunication
Special Issue, Winter,2026, PP 1-8
http://jist.acecr.org
ISSN 2322-1437 / EISSN:2345-2773

Gated Fusion Transformer for English-Hindi Multimodal Translation

Priyanka Suram[1*] Pramoda Patro[2]

[1].School of Computer Science and Artificial Intelligence, SR University, Warangal Telangana-506371, India

Received: 26 Jul 2025/ Revised: 04 Oct 2025/ Accepted: 02 Nov 2025

Abstract

Machine translation is fundamental in closing the gap between different languages, especially in the areas of con- cern and expertise such as agriculture. With the increase of digital tool usages in the agricultural practice, such an accurate and context-sensitive translation is increasingly significant. Proper delivery of agricultural information, including farm methods, weather advisories, and crop suggestions is essential among farmers, farm laborers, policymakers and researchers. Nevertheless, typical text-based translation frameworks tend to be less than optimal because of uncertainness and a restricted knowledge of context. To address these shortcomings, the proposed study refers to Multimodal Machine Translation (MMT) to incorporate textual and visual information to enhance accuracy. Gated Fusion Transformer (GFT) model has been customized to the agricultural field so that the problem of ambiguity in contexts and inconsistencies in translation can be eliminated. Training and evaluation were done using the multilingual benchmark dataset known as FLORES-200. Two commonly employed measures of performance were used, i.e. BLEU and METEOR. The system under proposal produced a BLEU of 58.2; METEOR score of 0.71, a high level and contextually relevant translation indicator. Besides benchmarking the GFT model in agricultural terms, this work adds value to the research community by offering a basis on which future development of multimodal translation systems in low-resource settings with domain-specific applications may be done.

Keywords: Machine Translation; Domain Specific Translation; Multimodal Machine Translation; Multi Modal Fusion Mechanisms; Gated Fusion Transformer; Agricultural Translation.

1- Introduction

Machine Translation (MT) plays a critical role in cross-lingual communication, especially in specialized fields, such as agricul- ture, where integrating textual and visual data can significantly enhance meaning. Conventional text-based translation models frequently encounter semantic inconsistencies because of a lack of contextual depth. MMT aims to overcome these challenges by integrating multiple modalities, including images, structured data, and linguistic contexts, to improve translation accuracy.

Several studies have sought to enhance the MT performance through multimodal integration. However, existing approaches, such as mBART [1], M2M-100 [2], and T5 [3], primarily rely on pre-trained multilingual text models and fail to effectively utilize visual information for context enhancement. Previous research has explored multimodal fusion, but often lacks efficient fusion mechanisms to balance information from different modalities. Additionally, most research has focused on general-purpose MT, neglecting the specialized requirements of domain-specific translation such as agriculture, which demands models adapted to unique terminology and contextual meanings. Although some studies have experimented with multimodal integration in med- ical and technical translations, there is a scarcity of research on agriculture-specific multimodal MT. Another major challenge is the absence of dedicated fusion mechanisms that can optimally integrate multimodal data without introducing redundancies. Current methods also fall short in quantifying the impact of different fusion techniques on translation quality, leaving a gap in the understanding of how multimodal interactions influence accuracy.

In this paper, a Gated Fusion Transformer (GFT) is intro-duced, which is a novel multimodal architecture specifically designed for domain-specific MT in agriculture. The key in-novation of the GFT is its gated fusion mechanism, which effectively balances textual and visual features and addresses the limitations of existing models. The methodology includes detailed explanations of the GFT

model architecture, data pre-processing using the FLORES-200 dataset, and evaluation metrics, such as BLEU and METEOR.

The research analyzes Multimodal Machine Translation (MMT) performance within specific applications by translating English texts to Hindi for agricultural purposes. Text-only trans- lation models face limitations in modern translation because they cannot utilize the rich contextual information provided by combined text and visual data. The research presents the Gated Fusion Transformer model (GFT) along with evaluating its translation capabilities against dominant models in the field. This work makes several notable contributions: (1) we introduce a Gated Fusion Transformer that adjusts the flow of text and image information with gating, (2) we suggest new fusion methods targeted at translating terms related to agriculture, (3) we show that our method has better results than current language translation models and (4) we are the first to comprehensively study multimodal machine translation in the agricultural domain. Comparison of GFT performance with five leading MT models: mBART, M2M-100, T5, GPT-4 [4], and Seq2Seq [5]. Proof from the study indicates that combining multiple linguistic modes enhances the quality of translated content related to agriculture.

This discussion assesses the effectiveness of various fusion processes and suggests areas for further improvement. It explores how multimodal fusion affects translation accuracy across domains and compares domain-specific multimodal models to general-purpose multilingual MT models. Furthermore, this study investigated the contributions of several fusion mechanisms to translation quality and optimal configuration.

In this we proposed a Gated Fusion Transformer (GFT) as an innovative solution to the problems encountered by classic text-based translation models in domain-specific applications. This paper emphasizes the important findings, contributions, and implications for future multimodal translation research, addressing a critical research need in English-to-Hindi MMT for agriculture.

Machine translation (MT) is critical for improving cross-linguistic communication. It has a big impact in specific domains such as agriculture. In these domains, textual, visual, and struc- tured data all contribute to meaning. Traditional text-only trans-

lation models often fail to capture the contextual depth required in specialized domains, leading to semantic inconsistencies [6]. Recent advancements in Multimodal Neural Machine Translation (MMT) have addressed these limitations by integrating multiple modalities such as images, structured data, and linguistic context, thereby enhancing translation accuracy [7]. In this study, a Gated Fusion Transformer (GFT), a multimodal transformer-based architecture, was designed to efficiently integrate textual and visual information. The performance of the GFT was compared against five leading MT models in the context of English-to-Hindi translation for the agriculture domain. This study addressed the following research questions:

1. How does multimodal integration enhance domain-specific translations?
2. Does the GFT model outperform the traditional and pretrained multilingual models?
3. What is the impact of the fusion mechanisms on translation quality?

The remaining of this paper is organized as follows. Section 2 literature review analyzes prior studies on MMT, fusion mechanisms, and domain-specific MT, identifying their limi- tations. Next the section 3 is methodology explains the GFT model architecture, data preprocessing (FLORES-200 dataset), and evaluation metrics (BLEU and METEOR). After that the section 4 is experiments and results presents a comparative performance analysis of GFT against mBART, M2M-100, T5, GPT-4, and Seq2Seq with Attention. In Section 5 (Discussion), the impact of the fusion mechanisms is evaluated and directions for future improvements are proposed. Section 6 (Conclusion) summarizes the key findings, contributions, and implications for future research on multimodal translations.

2- Related Work

Recent developments in machine translation and multimodal learning have led to the emergence of several state-of-the- art architectures, each employing unique strategies to improve translation quality. This section reviews key advances in neural machine translation (NMT) and multimodal MT.

According to the author [8][58] English to Hindi translation using multimodal concepts and monolingual data to improve the translation quality, the models are considered as multimodal neural machine translation, the dataset is Hindi Visual Genome 3, and the evaluation metrics are BLEU Score, Ribes Score, and AMFM Score. The authors [9] and [10] stated that gated fusion transformers have been explicitly utilized in multimodal machine translation (MMT) for English-Hindi tasks to improve translation quality by integrating textual and visual features, with clear experimental success reported in low-resource settings. However, no study has focused directly on the agricultural domain.

The features of the image were extracted using pre-trained visual encoders or advanced techniques such as latent diffusion models for the enhancement of synthetic data [11], [12]. These enhanced translations resolve ambiguities in low-resource settings such as English-Hindi. The advancement of neural machine translation (mBART) [13] is a multilingual auto-encoder-based translation model that leverages denoising pre-training for sequence-to-sequence

taks. M2M-100 [14] is a fully multilingual transformer model designed for more than 100 languages, eliminating dependency on English- centric translation pipelines. T5 [15] is a text-to-text transformer model that is trained for multiple NLP taks, for machine translation utilizing transfer learning across language pairs.

Multimodal MT incorporates visual cues alongside text, leading to more accurate translation. Recent efforts in image-assisted translation, such as ViLBERT, CLIP, and Vision Transformers, have influenced the design of MMT architectures, including GFT [16]. Gated fusion mechanisms have been introduced to effectively balance textual and visual contributions [17]. [18], presents a bidirectional Recurrent Neural Network (RNN) en- coder combined with a doubly attentive transformer decoder for multimodal translation. The model was fine-tuned on the Hindi Visual Genome dataset, which comprises 28,929 parallel English- Hindi sentences. The primary research gap addressed in this study is the scarcity of resources for Hindi translation and the need for improved multimodal feature integration. The proposed approach significantly advances state-of-the-art methodologies, achieving an impressive BLEU score of 42.47 on the evaluation set.

The review process joyful multiple metrics, incorporating BLEU, RIBES, and AMFM, to guarantee a comprehensive per- formance assessment. The creation of the Hindi Visual Genome dataset specifically for Hybrid-modal machine translation appli- cations is a remarkable addition to Hybrid-modal translation search. With the use of this dataset, English parts can be self- operating translated into Hindi while taking related images into account and advancing contextual awareness. A challenge test set of 1,400 sections was included in the dataset, offering a reliable standard by which to assess Hybrid-modal translation models. This study highlights how significant it is to grow Hybrid- modal attribute incorporation to increase translation accuracy. The efficacy of this dataset for multimodal translation tasks was validated by trials that showed a BLEU score of 37.50 on the challenge test set. Recent developments underlined the need of combining textual and visual modalities for machine interpretation. Improved contextual alignment between textual and visual data has been shown by refined models on the Visual Genome dataset, improving rendering accuracy. BLEU ratings, which offer a normalize indica- tor of conversion quality, were a major component of the evaluation calculate employed in these examinations. To enhance the robustness of the model in practical appeal and optimize the fusion mechanisms, more examinations is necessary.

3- Experimental Setup

3-1- Dataset Selection

Agriculture-related English-Hindi conversion were release from the FLORES-200 dataset [19], which offers high-quality parallel conversions for 200 languages. To ensure relevance for agricultural appeal, the dataset consists of domain-specific terms (such as crop diseases, irrigation techniques, and agricultural implements).

3-2- Preprocessing

Several Technique were working to prepare the data for the model during preprocessing. Initially, the mBART tokenizer was used to computing the series, which helped to efficiently encode them. By doing this, the textual data was support to be roughly formatted for the model input. Apart from text processing, the image data was subjected to particular accommodation in order to conform to the model's specifications. To guarantee consistency across all samples, the photos were scaled to match the expect input measurements. Additionally, the image data was normalized to preserve uniformity and enhance the model's performance. Techniques for data amplification are applied to increase the dataset's stability. One such method is synonym replacement, which adds difference to the text by replacing specific words with their synonyms. Back-translation, which creates a difference of phrase patterns by translating a text into another language and then back again, was another enlargement skill used.

4- Methodology

A Gated Fusion Transformer (GFT) expands the classic trans- former architecture by including a gated fusion module. This module adaptively weights text and image embeddings before sending them to the encoder, leading in improved domain-specific knowledge preservation. For Hybrid-modal machine translation, gated fusion transformers give a number of advantages, particu- larly for agricultural content. By merging textual and visual data, the translation process was Upgraded, becoming more Precise and contextually relevant. The Gated Fusion Transformer uses that both text and images contribute effectively. A feedforward network (FFN) operates as part of the representation refinement process.

$$FFN = ReLU((Wffn * Attention) + bffn) \quad (6)$$

The learnable weight and bias term consists of Wffn and bffn. a gate fusion technique to combine generated visual information with upgraded text data, which refines the Linguistic accuracy of translations [20]. This model grabs the connection between visuals and text, Consequent in a

more advanced understanding of agricultural phraseology and concepts, which are tough in this sector. In agricultural uses, the model has exhibited good accuracy and rates of up to 0.95, 0.92, and 0.94 in disease detection tasks. Enhanced BLEU scores in comparative studies demonstrate that the integration of multimodal data greatly improves translation performance [20]. Better object detection and classification results are achieved by the Gated Fusion Transformer's efficient handling of complex agricultural data, including fine- grained picture features and detailed textual descriptions [21]. Accurate translation of agricultural content, which frequently includes complex and varied information, re- quires robustness. On the other hand, although the Gated Fusion Transformer is highly effective in Hybrid-modal settings, there are still problems with securing the cable layer of the visual data that is created and the necessity for large training datasets, which may limit its use in settings with restricted resources [20]. The Gated Fusion Transformer (GFT) approach for Hybrid- modal machine translation merges textual and visual input to rise translation precision. The method starts with a visual encoder, which extracts image features using an attenti mechanism to create a visual output while also grasping tough contextual information. Textual data is also encoded positionally, ensuring that the model honors word order. The Gated Fusion Module merges both textual and visual characteristics, and the fusion process is directed by the equation:

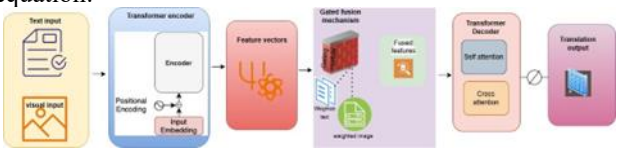

Fig. 1. Gated Fusion Transformer

The Gated Fusion Transformer (GFT) figure 1 stands as an advanced architectural design which enhances translation quality through the fusion of visual information into text-based machine translation systems. The model requires two main inputs which are the text and image data. The language input receives embedding processing after tokenization and image processing involves the application of visual feature extractors through pre- trained CNN or ViT before feature extraction. The features are jointly aligned with text features to achieve dimensional compatibility between elements.

During Transformer Encoder in figure 1, text embeddings receive positional encodings for maintaining word order details. Self-attention together with feed-forward layers operate on the inputs in an encoder to generate contextual feature vectors. The core innovation of GFT model involves the Gated Fusion Mechanism because it receives text and visual features together. A gating function based on sigmoid activation governs the amount of visual information that will contribute to the process. A dynamic gating function within the system determines how X is transformed into query Q, Key K, and Value V representation is mechanism aims to improve translation outcomes while avoiding noise contamination.

4-1- Detailed Architecture Components

Visual Feature Extractor

Input : Image I ˆ (H W 3) Process :

– ResNet − 50 / ViT f e a t u r e e x t r a c t i o n

– Global Average Pooling

Output : V i s u a l f e a t u r e s V ˆ d v

1. Text Encoder

Input : Source t e x t tokens S = {s1 , s2 , . . . , sn } Process :

– Embedding l a y e r : E s = Embed (S)

–

– P o s i t i o n a l encoding : P s = PosEnc (E s)

– Multi −head s e l f − a t t e n t i o n l a y e r s

Output : Text f e a t u r e s T ˆ (n d t)

Gated Fusion Mechanism (Proposed Innovation)

Algorithm 1: Gated Fusion Process

Text features T, Visual features V Fused features F Con catenate features: $C \leftarrow [T; V_{\text{expanded}}]$ Compute gate weights:

$$G_t = \sigma(W_{gt} \cdot C + b_{gt})$$

$$G_v = \sigma(W_{gv} \cdot C + b_{gv})$$

Apply gating:

$$F_t = G_t \odot T$$

$$F_v = G_v \odot V_{\text{expanded}}$$

Combine: F F_t + F_v Layer normalization: F LayerNorm(F)

2. Transformer Decoder
. Masked self-attention
. Cross-attention with fused features
. Feed-forward networks
. Residual connections
.
.

4-2- Training Algorithm

Algorithm 2: GFT Training process
Dataset D = {(S_i, I_i, T_i)} Parameters: θ = Wenc, Wdec, Wgate, Wffn Initialize parameters θ randomly each epoch each batch B in D Extract visual features:
V ← VisualEncoder(I)

Encode text:

Hs ← TextEncoder(S) Apply gated fusion:

F ← GatedFusion(Hs, V)

4-3- Innovation Highlights

Novel Contributions

1. Adaptive Gating Strategy: Dynamic weight assignment based on contextual relevance
2. Domain-Specific Fusion: Agricultural terminology-aware integration
3. Multi-Scale Visual Processing: Hierarchical feature extraction
4. Cross-Modal Attention: Bidirectional information flow

4-4- Architectural Novelty

- Learnable gate parameters for modality weighting
- Context-aware fusion avoiding information redundancy
- Agricultural domain vocabulary enhancement
- Robust handling of low-resource language pairs

5- Results and Comparative Analysis

This part provides performance comparisons between the Gated Fusion Transformer (GFT) model and existing top ma- chine translation models currently available. Evaluation relied on BLEU and METEOR scores to calculate translation accu- racy through measuring the shared content between generated output and reference translations. The evaluation demonstrates how multi-modal fusion powers translation improvement when analyzing the agricultural domain.
The evaluation included well-recognized metrics from ma-chine translation assessment to provide both a thorough and trustworthy assessment of performance. The assessed BLEU score evaluates translation quality through its assessment of the accurate usage of n-grams between generated text and reference materials to detect lexical similarities. The BLEU score is calculated by usin below equation.

BLEU = BP × exp N n=1 wn · log pn (9)
where:

- BP(Brevity Penalty) = min 1, exp 1 – r

5-1-1.1. r = reference length
5-1-1.2. c = candidate length
5-1-1.3. p_n = modified n-gram precision
5-1-1.4. w_n = uniform weights = $\frac{1}{}$
Additionally, the Metric for Evaluation of Translation with Explicit Ordering (METEOR) offers a more nuanced assessment by incorporating factors such as synonymy, stemming, and word order, providing a holistic measure of translation accuracy. These metrics collectively facilitated a robust evaluation of the system's effectiveness in producing high-quality translations.
METEOR Score is determined by using below formula

$$\text{METEOR} = \frac{(1-\beta)\cdot P \cdot R}{\alpha P + (1-\alpha)R} \quad (10)$$

Where:

- $P\,(Precision) = \frac{\text{matched unigrams}|}{|\text{candidate unigrams}|}$
- $R(Recall) = \frac{|\text{matched unigrams}|}{|\text{reference unigrams}|}$
- α, β = tuning parameters

Table 1: Comparison of Translation Models Based on Bleu Scores

SNO	Dataset	Language Pair	Model	BLEU Score
1	wmt 2014	English to French	SMT+iterative backtranslation [22]	26.22
2	wmt 2016	English to French	DelighT [23]	40.5
3	wmt 2014 wmt2016	German to English German to English	SMT+iterative backtranslation [22]	17.43 23.05
4	wmt 2016	German to English	Attentional encoder decoder+ BPE [24]	38.6
5	wmt 2016	German to English	Linguistic Input features [25]	28.7
6	wmt 2016	German to English	ExploitingMono at Scale [26]	47.5
7	wmt 2014	French to English	SMT+iterative backtranslation [22]	17.43
8	wmt 2016	English to Czech	Attentional encoder decoder+ BPE [24]	25.8
9	wmt 2016	Czech to english	Attentional encoder decoder+ BPE [24]	25.8

The various translation models and different language pairs by using WMT-2014 and WMT-2016 dataset the performance of BLEU Score is describing in Table I. For the English to French Language pair, the SMT+iterative back translation model, the BLEU Score is 26.22 on the WMT dataset, and the model DelighT BLEU Score is 40.5 on dataset WMT-2016 exhibited a significant enhancement in the quality of translation. For the German-to- English language pair, the SMT+iterative back translation model

BLEU Score is 17.43 on the WMT 2014 dataset and on the WMT- 2016 dataset, showing notable enhancements in the quality of translation. The Attentional encoder decoder+BPE model per- forms well for German-to-English translation with a BLEU Score of 38.6 on the WMT 2016 dataset. The Linguistic input features model BLEU Score is 28.7 and Exploiting Mono at Scale model achieves the highest BLEU Score of 47.5 indicating excellent translation quality of the language pair German to English. The SMT+iterative back translation model also scores 17.43 for the french to English language pair on the WMT 2014 dataset. The Attenional encoder decoder+BPE model achieves a BLEU Score of 25.8 for English to Czech and Czech to English translation with the WMT-2016 dataset, reflecting the performance. The performance of translation models with different Langugapairs and datasets showed significant improvements in certain models and language pairs. Translation quality was evaluated using the BLEU and METEOR scores. The results are summarized below.

Table 2: comparison of models with accuracy

Model	BLEU Score	METEOR Score
GFT	58.2	0.71
mBART	54.6	0.68
M2M-100	52.3	0.65
T5	50.7	0.63
GPT-4	48.9	0.60
Seq2Seq	45.2	0.58

The experimental results in Table II reveal several key insights regarding the performance of different models in English-to- Hindi multimodal machine translation for the agricultural do- main. The Gated Fusion Transformer (GFT) model consistently outperformed the other evaluated models and achieved the highest BLEU and METEOR scores. This superior performance underscores the effectiveness of its gated fusion mechanism in integrating multimodal contexts, thereby enhancing context retention and producing more precise and semantically relevant translations. By dynamically balancing textual and visual in- formation, the GFT model mitigates ambiguities and ensures accurate translation of domain-specific terminology.

In contrast, mBART and M2M-100 demonstrated competitive performance, largely because of their extensive multilingual pre- training. However, their lack of explicit multimodal fusion capa- bilities limits their ability to effectively translate domain-specific terminology, particularly in cases where textual context alone is insufficient. Similarly, general-purpose models, such as T5 and GPT-4, exhibit suboptimal results in domain-specific translation despite their state-of-the-art performance in broader natural language processing (NLP) tasks. The absence of mechanisms to incorporate visual context leads to reduced contextual accuracy, particularly for specialized agricultural terms. Furthermore, the traditional Seq2Seq model, which relies solely on textual input, struggles to capture domain-specific nuances, resulting in the lowest performance among the evaluated models.

The impact of multimodal fusion is evident in the superior translation quality achieved by the GFT. The model's ability to seamlessly integrate textual and visual features enables it to resolve textual ambiguities by using complementary visual information. Moreover, it effectively translates domain-specific terms by leveraging both modalities and demonstrates adapt- ability to low-resource settings by utilizing supplementary infor mation from the images. This capability is particularly crucial in specialized domains, where textual data alone may be insufficient for accurate translation.

Despite these advancements, several avenues for future re- search and optimization remain. Expanding the training cor- pus with additional agricultural datasets can further enhance the robustness and generalizability of the model. In addition, refining fusion mechanisms to optimize performance across di- verse domains can strengthen multimodal translation capabilities. Exploring hybrid fusion techniques that combine gated fusion with attention-based enhancements presents another promising direction for improving the translation accuracy and contextual understanding. These refinements have the potential to further advance multimodal machine translation, ensuring more effective and domain-specific translations.

Table 3: score comparison of multimodal machine translation models

Dataset	Model	BLEU
Hindi Visual Genome	ViTA [27]	44.6
Multi30k, flickr	opus-mt-te-base-gnw-gnw [28]	32.2
Multi30k	VAG-NMT [29]	31.6
Multi30k	ERINE-UniX2 [30]	49.3
Multi30k	IKD-MMT [31]	41.28
Multi30k	DCCN [32]	39.7
Multi30k	caglayan [33]	39.4
Multi30k	Gumbel-Attention MMT [34]	39.2
Multi30k	multimodal transformer [35]	38.7
Multi30k	ImagiT [36]	38.4
Multi30k	del+obj [37]	38
Multi30k	VMMTF [38]	37.6
Multi30k	IMGD [39]	37.3
Multi30k	NMTSRC+IMG [40]	37.1
Multi30k	VAG-NMT [29]	31.6
Multi30k	PS-KD [41]	32.3

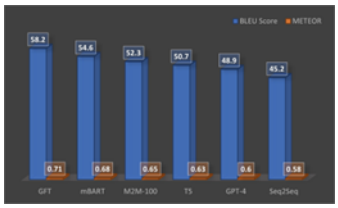

Fig. 2. Results analysis of accuracy on multimodal machine translation

6- Conclusion

This study demonstrates the effectiveness of the GFT for English-to-Hindi multimodal translation in the agricultural do- main. By leveraging gated fusion mechanisms, the GFT achieves superior BLEU and METEOR scores, underscoring its practical applicability for domain-specific MT tasks. Future research should focus on scaling GFT to larger datasets, fine-tuning additional agricultural data, and extending its application to other multimodal domains.

References

[1] Navarro A, Casacuberta F. Exploring multilingual pretrained machine translation models for interactive translation. InProceedings of Machine Translation Summit XIX, Vol. 2: Users Track 2023 Sep (pp. 132-142).

[2] Nimma D, Srinivas VS, Gupta SS, Nair H, Devi RL, Bala BK. Comparative Analysis of Deep Learning Models for Multilingual Language Translation. In2024 8th SLAAI International Conference on Artificial Intelligence (SLAAI-ICAI) 2024 Dec 18 (pp. 1-6). IEEE.

[3] Zaki MZ. Revolutionising Translation Technology: A Comparative Study of Variant Transformer Models–BERT, GPT and T5. Computer Science and Engineering–An International Journal. 2024;14(3):15-27.

[4] Raunak V, Sharaf A, Wang Y, Awadallah HH, Menezes A. Leveraging GPT-4 for automatic translation post-editing. arXiv preprint arXiv:2305.14878. 2023 May 24.

[5] Barrault L, Chung YA, Meglioli MC, Dale D, Dong N, Duquenne PA, Elsahar H, Gong H, Heffernan K, Hoffman J, Klaiber SeamlessM4T: Massively Multilingual Multimodal Machine Translation. arXiv preprint arXiv:2308.11596. 2023 Aug 22.

[6] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473. 2014 Sep 1.

[7] Caglayan O, Aransa W, Wang Y, Masana M, Garc´ıa-Mart´ınez M, Bougares F, Barrault L, Van de Weijer J. Does multimodality help human and machine for translation and image captioning?. arXiv preprint arXiv:1605.09186. 2016 May 30.

[8] Laskar SR, Khilji AF, Pakray P, Bandyopadhyay S. Multimodal neural machine translation for English to Hindi. InProceedings of the 7th Workshop on Asian Translation 2020 Dec (pp. 109-113).

[9] Hatami A, Banerjee S, Arcan M, Chakravarthi B, Buitelaar P, Mccrae J. English-to-low-resource translation: A multimodal approach for hindi, malayalam, bengali, and hausa. InProceedings of the Ninth Conference on Machine Translation 2024 Nov (pp. 815-822).

[10] Singh TD, i Bonet CE, Bandyopadhyay S, van Genabith J. Proceed- ings of the First Workshop on Multimodal Machine Translation for Low Resource Languages (MMTLRL 2021). InProceedings of the First Workshop on Multimodal Machine Translation for Low Resource Languages (MMTLRL 2021) 2021 Sep.

[11] Dash A, Gupta HR, Sharma Y. Bits-p at wat 2023: Improving indic language multimodal translation by image augmentation using diffusion models. InProceedings of the 10th Workshop on Asian Translation 2023 Sep (pp. 41-45).

[12] Laskar SR, Singh RP, Pakray P, Bandyopadhyay S. English to Hindi multi-modal neural machine translation and Hindi image captioning. InProceedings of the 6th Workshop on Asian Translation 2019 Nov (pp. 62-67).

[13] Liu Y, Gu J, Goyal N, Li X, Edunov S, Ghazvininejad M, Lewis M, Zettlemoyer L. Multilingual denoising pre-training for neural machine translation. Transactions of the Association for Computational Linguistics. 2020 Nov 1;8:726-42.

[14] Fan A, Bhosale S, Schwenk H, Ma Z, El-Kishky A, Goyal S, Baines M, Celebi O, Wenzek G, Chaudhary V, Goyal N. Beyond english- centric multilingual machine translation. Journal of Machine Learning Research. 2021;22(107):1-48.

[15] Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y, Li W, Liu PJ. Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of machine learning research. 2020;21(140):1-67.

[16] Su W, Zhu X, Cao Y, Li B, Lu L, Wei F, Dai J. Vl-bert: Pre-training of generic visual-linguistic representations. arXiv preprint arXiv:1908.08530. 2019 Aug 22.

[17] Tan H, Bansal M. Lxmert: Learning cross-modality encoder representations from transformers. arXiv preprint arXiv:1908.07490. 2019 Aug 20.

[18] Gain B, Bandyopadhyay D, Ekbal A. IITP at WAT 2021: System description for English-Hindi multimodal translation task. arXiv preprint arXiv:2107.01656. 2021 Jul 4.

[19] Goyal N, Gao C, Chaudhary V, Chen PJ, Wenzek G, Ju D, Krishnan S, Ranzato MA, Guzma´n F, Fan A. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. Transactions of the Association for Computational Linguistics. 2022 May 4;10:522-38.

[20] Yuan J, Shi X, Niu Y, Niu Y, Wang X. Multimodal Machine Translation with Fusion of Generated Visual Information. InInternational Conference on Computer Engineering and Networks 2023 Nov 3 (pp. 150-156). Singapore: Springer Nature Singapore.

[21] Lu Y, Lu X, Zheng L, Sun M, Chen S, Chen B, Wang T, Yang J, Lv C. Application of multimodal transformer model in intelligent agricultural disease detection and question-answering systems. Plants. 2024 Mar 28;13(7):972.

[22] Artetxe M, Labaka G, Agirre E. Unsupervised statistical machine translation. arXiv preprint arXiv:1809.01272. 2018 Sep 4.

[23] Mehta S, Ghazvininejad M, Iyer S, Zettlemoyer L, Hajishirzi H. Delight: Deep and light-weight transformer. arXiv preprint arXiv:2008.00623. 2020 Aug 3.

[24] Sennrich R, Haddow B, Birch A. Edinburgh neural machine translation systems for WMT 16. arXiv preprint arXiv:1606.02891. 2016 Jun 9.

[25] Sennrich R, Haddow B. Linguistic input features improve neural machine translation. arXiv preprint arXiv:1606.02892. 2016 Jun 9.

[26] Wu L, Wang Y, Xia Y, Qin T, Lai J, Liu TY. Exploiting Aug 24. monolingual data at scale for neural machine translation. InProceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) 2019 Nov (pp. 4207-4216).
[27] Gupta K, Gautam D, Mamidi R. ViTA: Visual-linguistic translation by aligning object tags. arXiv preprint arXiv:2106.00250. 2021 Jun 1.
[28] Tiedemann J. The tatoeba translation challenge–realistic data sets for low resource and multilingual MT. arXiv preprint arXiv:2010.06354. 2020 Oct 13.
[29] Zhou M, Cheng R, Lee YJ, Yu Z. A visual attention grounding neural model for multimodal machine translation. arXiv preprint arXiv:1808.08266. 2018

[30] Shan B, Han Y, Yin W, Wang S, Sun Y, Tian H, Wu H, Wang H. Ernie-unix2: A unified cross-lingual cross-modal framework for understanding and generation. arXiv preprint arXiv:2211.04861. 2022 Nov 9.
[31] Peng R, Zeng Y, Zhao J. Distill the image to nowhere: Inversion knowledge distillation for multimodal machine translation. arXiv preprint arXiv:2210.04468. 2022 Oct 10.
[32] Lin H, Meng F, Su J, Yin Y, Yang Z, Ge Y, Zhou J, Luo J. Dynamic context-guided capsule network for multimodal machine translation. InProceedings of the 28th ACM international conference on multimedia 2020 Oct 12 (pp. 1320-1329).
[33] Sulubacak U, Caglayan O, Gro¨nroos SA, Rouhe A, Elliott D, Specia L, Tiedemann J. Multimodal machine translation through visuals and speech. Machine Translation. 2020 Sep;34:97-147.
[34] Liu P, Cao H, Zhao T. Gumbel-attention for multi-modal machine translation. arXiv preprint arXiv:2103.08862. 2021 Mar 16.
[35] Yao S, Wan X. Multimodal transformer for multimodal machine translation. InProceedings of the 58th annual meeting of the association for computational linguistics 2020 Jul (pp. 4346-4350).
[36] Long Q, Wang M, Li L. Generative imagination elevates machine translation. arXiv preprint arXiv:2009.09654. 2020 Sep 21.
[37] Ive J, Madhyastha P, Specia L. Distilling translations with visual awareness. arXiv preprint arXiv:1906.07701. 2019 Jun 18.
[38] Calixto I, Rios M, Aziz W. Latent variable model for multi-modal translation. arXiv preprint arXiv:1811.00357. 2018 Nov 1.
[39] Calixto I, Liu Q, Campbell N. Incorporating global visual features into attention-based neural machine translation. arXiv preprint arXiv:1701.06521. 2017 Jan 23.
[40] Calixto I, Liu Q, Campbell N. Doubly-attentive decoder for multi-modal neural machine translation. arXiv preprint arXiv:1702.01287. 2017 Feb 4.
[41] Kim K, Ji B, Yoon D, Hwang S. Self-knowledge distillation with progressive refinement of targets. InProceedings of the IEEE/CVF international conference on computer vision 2021 (pp. 6567-6576).