Journal of Universal Computer Science, vol. 30, no. 5 (2024), 694-717
submitted: 12/7/2023, accepted: 29/12/2023, appeared: 28/5/2024 CC BY-ND 4.0

Multimodal Machine Translation Approaches for Indian Languages: A Comprehensive Survey

Binnu Paul
(Department of Computer Science and Information Technology, SHUATS, Prayagraj, Uttar Pradesh, India.
Department of CSE, National Institute of Technology, Agartala, Tripura, India
https://orcid.org/0000-0002-5982-1220, binnu.paul@shiats.edu.in)

Dwijen Rudrapal
(Department of CSE, National Institute of Technology, Agartala, Tripura, India
https://orcid.org/0000-0002-9729-277X, dwijen.rudrapal@gmail.com)

Kunal Chakma
(Department of CSE, National Institute of Technology, Agartala, Tripura, India
https://orcid.org/0000-0002-5648-0918, kchakma.cse@nita.ac.in)

Anupam Jamatia
(Department of CSE, National Institute of Technology, Agartala, Tripura, India
https://orcid.org/0000-0001-6244-8626, anupamjamatia@gmail.com)

Abstract: Multimodal machine translation (MMT) is a challenging task in the linguistically diverse Indian landscape. Machine translation refers to the task of automatically converting content from one language to another without human involvement. Within the realm of natural language processing, a significant challenge arises from the inherent ambiguity present in human language. Translation ambiguity is a cross-lingual phenomenon that can manifest itself for various reasons, including lexical ambiguity, the occasional need to impute missing words, the presence of gender ambiguity, and word-sense ambiguities. These factors can lead to a decrease in translation accuracy. The integration of multiple modalities, such as images, videos, and audio, in addition to text, plays a pivotal role in improving the robustness and precision of translation systems. Over the past five years, extensive research has been dedicated to incorporating secondary modalities alongside text to improve language translation and comprehension. In this comprehensive study, our objective was to identify and explore promising MMT approaches, available corpora, evaluation metrics, research challenges, and the future direction of research specifically for Indian languages. We evaluated 81 papers, including MMT models, MMT dataset in Indian languages, survey on MMT approach, and the effects of multiple modalities in machine translation. The performance of the different proposed approaches has also been briefly analyzed on the basis of the claimed results and comparative evaluations. Finally, the challenges associated with the MMT task for India and some possible directions for future research in this domain are highlighted.

Keywords: Multimodal Machine Translation, Multimodal data, Machine Translation, Indian languages, Machine Learning
Categories: H.3.1, H.3.3, H.5.1, I.1.3, I.2.1, I.2.7
DOI: 10.3897/jucs.109227

1 Introduction

Language serves as a medium for the expression of views and thoughts among human beings. In the world, there exist thousands of languages with varying degrees of similarity and diverse structures and constituents. Dissimilarities between two languages create proportional obstacles for communicators. Machine Translation (MT) is an IT solution aimed at overcoming communication barriers between communicators. In other words, MT is a tool that automatically translates content from one language to another. The linguistic aspects of a source language, such as syntax, semantics, and pragmatics, are comprehended by MT tools and generate fluent and adequate content in another language. Robust MT approaches can fluently translate most of the content. However, in many cases, translating content while preserving its meaning in the target language is not achievable through textual features alone due to the multimodal nature of human expressions. For example, MT encounters complexities when translating words with multiple senses or gender information from a gender-neutral language to a gender-dependent language. For language pairs with very few resources, training with Neural Machine Translation (NMT), a memory-based approach, often results in poor performance in machine translation systems [Singh and Singh, 2020, Singh and Hujon, 2020, Novotný et al., 2022]. To simplify the translation process, various modalities of information, such as images, videos, or audio related to an event, contribute to better understanding and translation.

Table 1 provides some MT examples from English to the Hindi language. In the first two examples, English sentences are not translated into exact Hindi sentences due to the presence of homonyms like 'lying' and 'bank'. Although the third example appears to be correct, considering the gender of the tennis player, two possible Hindi translations are presented in Table 2. Hindi is a gender-specific language and, as such, the translated Hindi word 'रहा' is used when the subject is male, and 'रही' is used when the subject is female. When only the text is considered for translation, there is a high possibility of incorrect translation. However, when an image is taken into account, as shown in the same example, the extracted features clarify that the tennis player in the image is female. Thus, for the above example, the mentioned translation ambiguity can be resolved correctly by adding a modality, i.e., the associated image, to the text. Likewise, alternative information modalities such as video and audio can improve the precision of the source text during translation into the target language. For example, consider translating a written transcript of a motivational speech from English to Hindi. In such cases, video and audio recordings become essential, as they capture tone, rhythm, and emotions, all of which significantly influence the speaker's message[Honegger et al., 2021, D'Andrea et al., 2021].

SN	English Input	Hindi Output
1	The sheep's are lying.	भेड़ें झूठ बोल रही हैं।
2	I stand at the bank.	मैं बैंक में खड़ा हूं।
3	A tennis Player is playing	एक टेनिस खिलाड़ी खेल रहा है

Table 1: Example of English-Hindi machine translation

Integration of properties of multiple modalities of information with the textual characteristics of the content from the source language to be translated into the target language

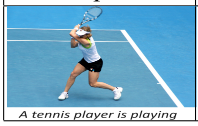
Input	Wrong Output	Correct Output
 A tennis player is playing	एक टेनिस खिलाड़ी खेल रहा है	एक टेनिस खिलाड़ी खेल रही है

Table 2: Example of multimodal English-Hindi machine translation

is carried out by multimodal machine translation (MMT) [Shah et al., 2016]. Applications of this task include caption translation, video content translation, sign language translation [Albahri et al., 2023, Noyan et al., 2022], and spoken language translation. In the last decade, research works in MMT have been proposed for different languages spoken throughout the world.

Research on MMT for Indian languages is also considered an important research domain due to the diversity of regional languages in the country and the interpretation and understanding of communications. Significant progress has been made in recent years for a few spoken Indian languages such as Hindi, Bengali, Malayalam, Assamese, and Mizo, although several limitations and challenges remain. These limitations and challenges are discussed separately in Section 6.1 in detail. The translation of languages is crucial in India to make any application or utility accessible to all regions. Therefore, it is necessary to extend existing research and develop new approaches for mostly spoken languages. In this comprehensive study, efforts were made to identify and explore promising MMT approaches, available corpora, evaluation metrics, research challenges, and the future direction of research specifically for Indian languages. Although existing survey work has attempted to provide an overview of the task, research trends, and progress in this domain, this current survey work is the first attempt, to our knowledge, to review MMT research progress for Indian languages.

In this current survey work, our main focus is on covering the state-of-the-art MMT approaches developed for Indian languages. We considered 127 papers during our literature search and eventually selected 81 papers, which included multimodal machine translation models for Indian languages, MMT dataset in Indian languages, the survey on MMT approaches and the effects of multiple modalities in machine translation. Our search was restricted to various available digital repositories such as Springer, ACL, IEEE, Web of Science, etc. and grasp, prior and derivative works related to the subject in focus using tools available in Connected Papers [1].

The aims of this comparative study are as follows:

- To study and analyze the methodologies and datasets adopted in previous research works to address MMT tasks for different Indian languages.

- To assess the current progress of proposed translation approaches by comparing the evaluation methodologies adopted in promising research works.

- To make an in-depth analysis of various challenges related to multimodal machine translation (MMT) tasks specifically for Indian languages and to outline possible future directions of research in this domain.

The article is structured as follows: Section 2 describes the background of MMT research and promising approaches proposed for widely spoken global languages. The

[1] https://www.connectedpapers.com/

MMT data sets developed in different Indian languages and the approaches proposed in these data sets are discussed in Section 3 and Section 4, respectively. An analysis of the performance of the proposed approaches is briefly discussed in Section 5 followed by the research challenges and the future direction of the research in Section 6. Finally, the article concludes the survey work in Section 7.

2 Research Progress of Multimodal Machine Translation task

Multimodal machine translation [Sulubacak et al., 2020] involves information from multiple modalities that contain useful associated views of input data and make interpretation of the text in the source language into a target language. The information from each of the modalities plays a crucial role in the proper interpretation and understanding of the meaning in the source language. The goal of multimodal machine translation is to learn a mapping function F that can generate the target language sentence Y given the source language sentence X and the multimodal data $D_1, D_2, ..., D_m$. The mathematical formulation for MMT can be written as:

$$Y = F(X, D_1, D_2, ..., D_m) \tag{1}$$

In general, the mathematical formulation for MMT involves learning a mapping function that can generate the target language sentence by considering information from multiple modalities. The training objective is to minimize the deviation of the meaning of the sentence between the source and target language.

The goal of MMT is to understand the meaning of a sentence in one language and then translate it into another. This understanding becomes more unambiguous when a representation comes from multiple associated modalities of information, such as text, visuals, audio, and video, rather than from one modality alone. Consequently, humans also develop the ability to integrate multiple modalities for an effective understanding [Stein et al., 2009] of language, in addition to depending on a single modality in an isolated way. Initial research work in this direction by [Silberer and Lapata, 2012] showed that the combination of linguistic and perceptual information correlates better with human judgments than the representation of individual modality. Since the introduction of the multi-modality concepts for efficient perception, research communities contributed many promising datasets on various languages like English-German [Elliott et al., 2016], English–German and Spanish [Grubinger et al., 2006], English–Chinese [Wang et al., 2020], Japanese–English [Tamura et al., 2020] and many prominent MMT approaches for these languages. Another dataset 'Flickr30k dataset [Young et al., 2014] of English image descriptions has been translated into different languages by the approaches proposed in [Elliott et al., 2017, Barrault et al., 2018].

Although the various approaches proposed for MMT are diverse, they generally exhibit modality-specific characteristics and share several common factors. The crucial role of the composition of the modality, as indicated by [Sulubacak et al., 2020] in their study on multimodal translation, determines the suitability of data, methodologies, and analysis for various MMT tasks. We can categorize the existing MMT approaches into three classes, namely:

1. Retrieval-based Models
2. Generative Models

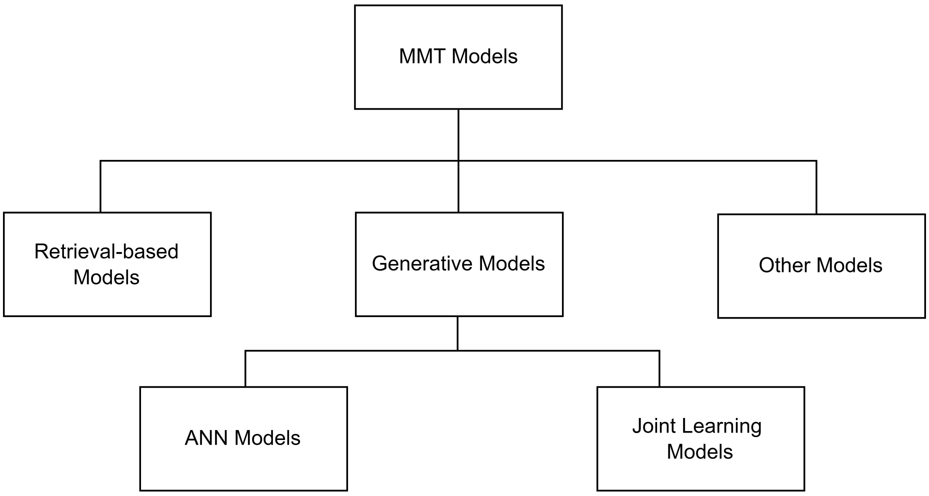


Figure 1: Classification of MMT Models

3. Other Models

The categorization of the MMT models depicted in Figure 1 is predominantly based on their approach to incorporating visual information into the translation process. Generative models belong to the category of conventional MMT systems, which rely on direct alignments between text and images. In contrast, retrieval-based models dynamically retrieve pertinent images to offer additional context for translation.

2.1 Retrieval-based Models:

Retrieval-based MMT models retrieve relevant images from an image corpus to provide additional context for translation. The research of [Zhang et al., 2019] integrates a retrieval component with MMT that utilizes TF-IDF to create a token-to-image lookup table, allowing retrieval of images with similar topics to generate training instances with bilingual annotations in the image. The work of [Li et al., 2022] introduced a visual hallucination framework called 'VALHALLA', where discrete visual tokens are predicted based on input text using a hallucination model. This approach offers greater flexibility in visual hallucination by using a transformer that autoregressively models text and image tokens within a unified data stream. The recent MMT approach in [Fei et al., 2023] explored unsupervised multimodal machine translation in an image-free inference time setting, introducing a visual scene hallucination mechanism to generate pseudo-visual features during inference. Their study also introduced SG-pivoting learning objectives for unsupervised translation training in the context of Retrieval-based MMT.

2.2 Generative Models:

Generative MMT models are typically based on training and inference data sets that include images annotated in both the source and target languages. In these types of MMT system, there is usually a precise alignment between the source and target sentence pairs and their corresponding images. This category can be further classified into two subtypes: Artificial Neural Network Model (ANN Model) and Joint Learning Model.

1. ***Artificial Neural Network Models:*** The MMT approaches in this category primarily focus on encoder-decoder architecture utilizing deep neural networks. It addresses limitations in sentence composition seen in retrieval-based. The core concept involves encoding a source modality into a vector representation and then employing a decoder module to generate the target modality in a single-pass pipeline. The first such approach by [Elliott et al., 2015] used a Recurrent Neural Network (RNN) encoder-decoder without attention, initializing both the encoder and decoder with an image vector. Although this multimodal setup outperforms text-only models, it achieved considerably better translation quality by utilising the attention mechanism [Bahdanau et al., 2015]. The early approaches often employed pre-trained models, such as ResNet [He et al., 2016] for encoding images, initialized hidden vectors [Calixto et al., 2016, Libovický and Helcl, 2017], word embeddings with additional input tokens [Huang et al., 2016, Calixto and Liu, 2017]. Recent approaches [Libovický et al., 2018][Zhou et al., 2018][Ive et al., 2019][Lin et al., 2020a] incorporate attention mechanisms to create decoder representations that take into account visual information. The proposed MMT approaches in [Calixto et al., 2017, Helcl et al., 2018, Arslan et al., 2018] add an additional visual attention layer between the source-target and feedforward layers of the decoder. The focus on improving decoders in these researches has been significant, but encoder-based approaches have received comparatively little attention. The approach of [Yao and Wan, 2020, Yin et al., 2020] in this direction developed a multimodal encoder to replace the conventional Transformer encoder. One significant challenge when incorporating multimodal information into translation is that an abundance of such information can introduce unwanted noise [2][Gain et al., 2021a]. Therefore, it is crucial to regulate the quantity of information integrated from secondary modalities. As a contribution in this direction, the work in [Wu et al., 2021] utilizes a gating matrix Λ to govern the degree of incorporation of visual information into textual representations. The higher values of Λ_{ij} within the range [0, 1] indicate greater utilization of visual context in translation, while lower values signify greater reliance on textual information. To further improve MMT Systems, the recent proposed approach by [Zhao et al., 2022] incorporates semantic image regions using two modality-specific attention mechanisms, the graph-based multimodal fusion encoder [Yin et al., 2023] for robust MMT.

2. ***Joint Learning Models:*** Joint learning is a technique that involves training a model on multiple subtasks simultaneously. In this approach, the model shares some parameters among these tasks and executes them concurrently. The primary goal is to learn representations of one modality through a task and leverage these representations to enhance the performance of other tasks involving different modalities, ultimately leading to improved translation quality. The approach proposed in [Elliott and Kádár, 2017, Zhou et al., 2018] adopts joint learning from various modalities in closely related subtasks to train the translation model holistically. Additionally, the MMT architecture introduced by [Helcl et al., 2018] divides multimodal translation into two subtasks: a translation task and an auxiliary visual reconstruction task. This method allows the model to capture a representation of the source sentence enriched with visual information. However, it is important to note that joint training approaches

[2] http://multicomp.cs.cmu.edu/research/multimodal-representation/

may not always be ideal due to potential misalignments between different modalities. These modalities can exhibit varying types and levels of noise at different points in time, which can pose challenges to seamless integration.

2.3 Other Models:

All the models discussed so far predominantly employ neural network-based methodologies for generating visual descriptions and establishing correlations between modalities to improve translation quality. In the case of the MMT framework introduced in [Chen et al., 2018], it treats the task as a multi-agent communication game in which a translator and a captioner collaborate to achieve translation. Other MMT approaches, as proposed in [Calixto et al., 2019, Elliott, 2018, Long et al., 2020], outline the MMT problem and tackle it using latent variable models and Generative Adversarial Networks (GANs). Furthermore, the Dynamic Context-guided Capsule Network for MMT, presented in [Lin et al., 2020b], employs a context-guided dynamic routing mechanism to extract visual and textual features for fusion and incorporation into the translation process. In the study conducted by [Sato et al., 2023], it is shown that masking approaches have a significant positive impact on the MMT task. The proposed work incorporates three masking strategies to learn different linguistic patterns.

Numerous MMT approaches have been proposed for a wide range of languages around the world. However, the development of MMT approaches for Indian languages has been relatively scarce, primarily due to several challenges elaborated in Section 6. Among the challenges discussed, it should be noted that most of the proposed MMT approaches for Indian languages are designed to support image-based translation, with only a limited number dedicated to video- or audio-based translation. In the subsequent section, we delve into the various MMT datasets and approaches introduced for Indian languages, encompassing languages such as Hindi (Hi), Malayalam (Ml), Tamil (Ta), Telugu (Te), Bengali (Bn), Assamese (As), and Mizo (Mz).

3 Dataset

A dataset represents a collection of different forms of digital information. The development of a sophisticated MMT model is critically dependent on a properly pre-processed and annotated dataset. As mentioned in the last section, some standard datasets have been developed for some of the widely spoken languages in the world, such as English, German, Spanish, Chinese, Japanese, etc. In this section, various multimodal datasets developed for Indian languages are discussed.

3.1 Hindi Visual Genome (HVG)

The dataset containing a Hindi variant subset of Visual Genome[3], as described in the work [Parida et al., 2019], is comprised of several elements. Within the dataset, each image includes three components: the original English caption from the Visual Genome, a rectangular area within the image, and finally a Hindi translation. To interpret captions from English to Hindi, a multimodal machine translation (MMT) model is employed, as detailed in [Vaswani et al., 2018]. In addition, a special challenge test set is also included,

[3] http://visualgenome.org

SN	Dataset	Translation Language Pair	Domain	No. of Sentences	Avg. Length of the Sentences	No. of Tokens	No. of Types	No. of Characters	Singletons	Tasks associated with the dataset
1	Synthetic En-Hi Dataset (Synthetic) [Meetei et al., 2023]	En-Hi	News	49,500	En:11 Hi:13	N.A	N.A	N.A	N.A	MT,MMT
2	Assamese Multimodal Dataset [Das and Singh, 2021]	En-As		13,000	N.A	N.A	N.A	N.A	N.A	MT,MMT
3	HVG [Parida and Bojar, 2019]	En-Hi	General	32,925	En:4.99 Hi:5.07	En:164137 Hi:157617	En:8669 Hi:8719	En:815998 Hi:765543	En:3745 Hi:655	MT,MMT
4	BnVG [Sen et al., 2022]	En-Bn		32,923	En:4.98 Bn:3.98	En:164076 Bn:130979	En:8609 Bn:9732	En:815768 Bn:750892	En:3717 Bn:736	MT, MMT
5	MlVG [Parida and Bojar, 2021]	En-Ml		32,923	En:4.98 Ml:3.72	En:164073 Ml:122478	En:8598 Ml:13741	En:815746 Ml:949285	En:3720 Ml:701	MT, MMT
6	AsVG [Laskar et al., 2021b]	En-As		31,325	N.A	En:156272 As:159065	N.A	N.A	N.A	MT,MMT
7	MzVG [Khenglawt et al., 2022]	En-Mz		31,325	N.A	En:161436 Mz:188316	N.A	N.A	N.A	MT,MMT

Table 3: Major corpora statistics for MMT task in Indian languages

which includes 1,400 segments of text containing inconclusive words. In these cases, the visual feature serves to clarify the ambiguity. The dataset was utilized for the WAT 2019 shared task on multi-modal translation (MMT), as outlined in [Nakazawa et al., 2021a]. It comprises 32,925 sentences divided into 28,932 training sentences, 998 validation sentences, 1,595 test sentences, and 1,400 challenge sentences. The details of the Hindi Visual Genome corpus are provided in Table 3. HVG is also termed as Hindi Visual Genome 1.0. The updated version of Hindi Visual Genome i.e. Hindi Visual Genome 1.1 (HVG 1.1) is released on 2020, resolving translation issues reported during WAT 2019 multimodal task.

3.2 Bengali Visual Genome (BnVG)

The Bengali Visual Genome (BnVG) is a multimodal dataset developed by [Sen et al., 2022] includes text and images appropriate for multimodal machine translation (MMT) tasks for English to Bengali translation. The BnVG consists of short English segments along with their corresponding images, which were manually interpreted into Bengali by native Bengali speakers. A total of 28,930 segments are included in the training set. For test and development sets, another 998 and 1595 segments were provided, respectively, following the same sample of the original Hindi visual genome 1.1. In addition, a challenge set of 1400 segments has been prepared for the multimodal sharing task in WAT 2022. The challenge set was created by selecting inconclusive English words based on embedded similarity along with associated images which resolved the ambiguity.

3.3 Malayalam Visual Genome (MlVG)

This dataset [4] is developed by the work in [Parida and Bojar, 2021]. It is a multimodal dataset with text and images suitable for multimodal English-to-Malayalam translation research. As in Hindi Visual Genome 1.1, it follows a selection of short segments which are in English and accompanying images for translation. MlVG captions were automatically translated from English to Malayalam. However, the captions were later corrected manually by considering the corresponding images. A challenge test set was

[4] https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3533

also created for the multimodal shared task [5] in WAT 2019 by collecting ambiguous English words based on embedded similarity whose images help to resolve the ambiguity.

3.4 Assamese Visual Genome (AsVG)

The dataset is developed by [Laskar et al., 2021b] for the English-Assamese multimodal machine translation system. Using the Baseline English-Assamese Neural Machine Translation (NMT) model, the English segments in Hindi Visual Genome 1.1 are interpreted into Assamese. Furthermore, the translated text is manually verified and corrected using the Assamese typing tool [6]. The post-translation editing task was carried out by native Assamese speakers. Another English-Assamese multimodal machine translation dataset is developed by [Das and Singh, 2021] which includes 10,000 images and captions. Further, this dataset is enhanced in the work by [Das and Singh, 2022] with 13,000 numbers of images along with their captions. All images and captions were collected from different local Assamese e-Newpapers.

3.5 Mizo Visual Genome (MzVG)

This dataset was developed by [Khenglawt et al., 2022] for the English-Mizo multimodal machine translation system. For the creation of Mizo Visual Genome, the English segments of Hindi Visual Genome 1.1 were utilized. The English segments were translated into Mizo using Google translation tools; however, the translation was not accurate and approximately 97% of the sentences required corrections. Consequently, a post-editing task was undertaken by native Mizo speakers.

4 Multimodal Machine Translation Approaches for Indian Languages

India has a wide range of languages, each with its own standards and scripts. The language barrier makes it difficult to spread information, communicate in languages, and provide or access various services. Therefore, extensive research in MMT tasks for major Indian languages is necessary. However, other than Hindi, very limited MMT approaches have been proposed for other Indian languages. The research progress of MMT for Indian languages is discussed in this subsection.

4.1 English to Hindi MMT Systems

Hindi is the first language spoken widely in India. As an initial contribution towards English-Hindi (En-Hi) multimodal research, the work in [Chowdhury et al., 2018] proposed an attention-based multimodal machine translation. The work integrates visual features along with textual features during encoding and decoding in translation. The encoding and decoding procedure of the translation system followed the structure proposed by [Calixto and Liu, 2017]. The work also introduced a synthetic dataset for the En-Hi multimodal machine translation task as one of the first datasets in this domain. The conclusion of the work shows that a multimodal neural machine translation system

[5] https://ufal.mff.cuni.cz/malayalam-visual-genome/wat2022-english-malayalam-multi

[6] https://www.lipikaar.com/

Sl No	Work	Approaches	Modalities	Datasets	Language Pair	Dir.	Evaluation Metrics
1	[Chowdhury et al., 2018]	AT-MMT	Text Image	Flickr30k	Hi-En	U	BLEU, METEOR
2	[Laskar et al., 2019]	DAD-MMT		HVG	En-Hi	U	BLEU, RIBES AMFM
3	[Meetei et al., 2019]	AT-MMT		HVG	En-Hi	U	BLEU, RIBES AMFM
4	[Laskar et al., 2020]	DAD-MMT		HVG 1.1	En-Hi	U	BLEU, RIBES AMFM
5	[Laskar et al., 2021a]	DAD-MMT		HVG 1.1	En-Hi	U	BLEU, RIBES AMFM
6	[Gain et al., 2021b]	DAD-MMT		HVG 1.1	En-Hi	U	BLEU, RIBES AMFM
7	[Gupta et al., 2021]	ViTA		HVG	En-Hi	U	BLEU, RIBES AMFM
8	[Gain et al., 2021a]	TR-MMT DAD-MMT		HVG 1.1	En-Hi	U	BLEU, RIBES
9	[Meetei et al., 2021]	DAD-MMT		Synthetic	En-Hi	U	BLEU
10	[Laskar et al., 2022b]	DAD-MMT		HVG 1.1	En-Hi	U	BLEU, RIBES
11	[Shi and Yu, 2022]	TR-MMT		HVG 1.1	En-Hi	U	BLEU
12	[Singh et al., 2021b]	DAD-MMT		Multi-30K	En-Hi En-Te	B	BLEU
13	[Parida et al., 2022]	TR-MMT		HVG 1.1 MlVG BnVG	En-Hi En-Ml En-Bn	U	BLEU
14	[Parida et al., 2021b]	TR-MMT		HVG MlVG	En-Hi En-Ml	U	BLEU
15	[Chakravarthi et al., 2019]	DAD-MMT		Multi30K	En-Ta En-Kn En-Ml	B	BLEU
16	[Parida et al., 2021a]	ViTA		BnVG	En-Bn	U	BLEU
17	[Laskar et al., 2022a]	DAD-MMT		BnVG	En-Bn	U	BLEU, RIBES
18	[Laskar et al., 2021b]	DAD-MMT		AsVG	En-As	B	BLEU, TER RIBES, METEOR F-measure
19	[Kumar and Lalithamani, 2022]	AT-MMT		COCO	En-Ta	U	BLEU
20	[Lekshmy and Jayaraman, 2022]	AT-MMT		MlVG	En-Ml	U	BLEU
21	[Khenglawt et al., 2022]	AT-MMT		MzVG	En-Mz	B	BLEU, TER RIBES, METEOR F-measure

Table 4: Summary of MMT systems for Indian languages

can be trained for under-described language pairs with the help of synthetic data and also found that the translation system can benefit from visual features.

Another attention-based MMT approach proposed by [Meetei et al., 2019] for translating English captions into Hindi using text and image modalities. The proposed work experimented with the Hindi Visual genome dataset [Nakazawa et al., 2019] introduced as a shared multimodal translation task at the Asian Translation Workshop (WAT 2019). In this study, visual features were used in conjunction with textual features to improve the accuracy of En-Hi language translations. The proposed approach reports satisfactory performance for the evaluation and challenge dataset published in the shared task by outperforming their Text-Only Translation (TOT) and Hindi-only Image Captioning (HOC) modules in the same work. The approach in [Laskar et al., 2019] proposed an MMT system for the English-Hindi MMT task on the Hindi Visual Genome dataset provided in each track of the WAT 2019 MMT task as in the work of [Nakazawa et al., 2019]. The proposed model is trained by a doubly attentive Decoder where two different attention mechanisms are used, one for the source language words and the other for image features in a single RNN decoder. The proposed multimodal model not only showed a better performance than pure text translation, but also clearly outperformed other approaches presented in the shared task for En-Hi multimodal translation.

Furthermore, in WAT 2020, the translation task proposed by [Laskar et al., 2020] discovered that the quality of MMT for En-Hi can be further enhanced using monolingual data. This idea evolved from previous research done for the NMT system introduced in the work by [Sennrich et al., 2016] and [Zhang and Zong, 2016]. A monolingual corpus is used in the pre-training stage of the translation process to enhance the performance of target contents. Through experiments, it is found that the proposed approach is very effective, especially in the case of scarce-resource language translation. The proposed approach was further improved in the work by [Laskar et al., 2021a] using phrase pairs from the source and target languages through data augmentation.

The MMT approach in [Meetei et al., 2021] proposed a doubly attentive Decoder model for MMT for a domain-specific dataset. The proposed approach obtained two monolingual news dataset published in English and Hindi that were linked to images to create a synthetic aligned corpus in English and Hindi. The work reported a systematic analysis of the translation task for domain-specific data other than open data used in all the previous research works. The work by [Gupta et al., 2021] focuses on the quality of translation by fine-tuning the mBART model for En-Hi multimodal machine translation. By adding the object tags detected from the image, the visual information was added to the text. The faster R-CNN with ResNet-101-C4 backbone [7] was used to detect the list of objects present in the image. To explore the effectiveness of different state-of-art MMT techniques for En-Hi or vice versa MMT task, the work by [Gain et al., 2021a] used Transformer based multimodal machine translation approach and Double attentive decoder approach in their experiments for comparative study of different MMT Systems developed for En-Hi Language. The work concludes that among the systems they implemented and experimented with, object tag extraction from images enhanced the performance and outperformed other MMT systems, as well as text-only NMT systems.

Multiple captions in a multilingual setup are used in the approach proposed by [Singh et al., 2021b] and generated improved translated content compared to previous research results. The work concluded that the system can be made more robust to manage the sparse word translation using multiple captions, and also added that the translation can be alleviated with the additional image features which lay out the context information. Considering the quality of translated content, another work by [Laskar et al., 2022b] proposed an approach based on transliteration for augmenting phrase pairs to improve multimodal translation performance from English to Hindi.

In a recent research work, [Shi and Yu, 2022] proposed an approach based on the transformer that merged visual features with text information in the process of text encoding for translation. In the study, various feature extraction models were analyzed with respect to the impact of visual feature information on translation results, and the ResNet model was shown to be more effective in generating more accurate translations than the other models.

[7] https://github.com/facebookresearch/detectron2

4.2 English to Bengali MMT Systems

In the past few years, some promising research approaches have been proposed for English-Bengali (En-Bn) Multimodal Machine Translation tasks. The approach proposed by [Parida et al., 2021a] is the first approach to develop an MMT system for an En-Bn language translation pair. The Bengali Visual Genome (BnVG) 1.0 [Sen et al., 2022] dataset is used in the work for experimentation and model generation. The proposed approach encodes English sentences using the mBART model [Liu et al., 2020] along with object tags extracted from images following the ViTA approach introduced by [Gupta et al., 2021] and then decodes the text to produce Bengali translations.

As a further contribution in the same direction, the work proposed by [Laskar et al., 2022a] performed 4 operations, namely phrase pair expansion based on transliteration, data pre-processing, model training, and testing to translate text from English to Bengali. The tool 'OpenNMT-py' [Klein et al., 2017] is used to create multimodal and plain text models separately. The aim of this approach, which is based on transliteration, is to enable the sharing of vocabulary at the sub-word level between the initial and the final sentences shared during the process of training. During the Multimodal Machine training phase, a bi-directional RNN in the encoder and a double-attentive RNN in the decoder are used as explained in the work by [Calixto and Liu, 2017, Calixto et al., 2017].

4.3 English to Tamil MMT Systems

The research work in [Kumar and Lalithamani, 2022] studied the COCO dataset and parallel English and Tamil datasets from [Jain et al., 2020] for English-Tamil language translation. The work also proposed four different MMT models to translate English captions into Tamil captions for the given image. All four different models have been experimented with in the work for evaluation of the proposed approach. First, two different models were trained at the same time to translate English into Tamil captions. The first model generates a caption in English for a given image using Inception V3 (for CNN Image Encoder) introduced in the work by [Boonyuen et al., 2019]. The second model is used to transform the produced English captions to the desired output of Tamil captions. The generated Tamil captions are based on a model transformer and positional encoding using the encoder and decoder architectures [Jain et al., 2020].

Second, in addition to the previous model, before training, they had input the sentence with <start> and the <end> tags to train the model. The ResNet architecture is used to extract the features present in the image. During training, the model used the cross-entropy loss to mask steps where the input tokens are paddings denoted by the <pad> symbol. The trained model generated captions in English. Then the generated English captions were translated into Tamil using the same Tamil translator which was used in the previous model. Third, in this approach, instead of 2 models as in previous approaches, a single model is trained on the Tamil captions dataset and using Inception V3 with decoder architecture English caption of image translated into Tamil. This proposed approach eliminates the need to re-translate the generated captions to get the output. Fourth, in this approach, the previous MMT model is enhanced by increasing the epochs count with various batch sizes from 8 to 64.

4.4 English to Malayalam MMT Systems

A multimodal machine translation model for English-Malayalam (En-Ml) languages was proposed by [Lekshmy and Jayaraman, 2022]. It was developed using an encoder-decoder architecture. On the encoder side, this model uses a three-layer network. The encoder receives Text Embedding and Image Embedding as input at each time step. The model updates the hidden vectors at each time step and contains the full context details of the sentence, which is fed as input and image at the last time step. Later, this context vector is inserted into the decoder. The decoder predicts the target words one by one and later concatenates them to finally produce the final translation output. It uses an additional weight matrix to create probabilities for every word in the output vocabulary before advancing to the next time step. This way for each time steps the most probable word is predicted. Depending on the morphological features of the target language decoder outputs can be variable. During experiments and evaluation, 990 images and corresponding English captions are used for the validation segment while 1400 images and corresponding English captions are included for test data.

4.5 English to Assamese MMT Systems

As a contribution towards the MMT system for English-Assamese (En-As) languages, [Laskar et al., 2021b] developed a dataset i.e. Assamese Visual Genome (AsVG) and based on this dataset, an En-As paired MNMT (Multimodal Neural Machine Translation) system proposed to fill the research gap in English-Assamese MMT. This proposed MMT approach followed the multimodal setup described in [Calixto et al., 2017]. Visual Global and local features were extracted using the publicly available pre-trained CNN model VGG-19 [Nakazawa et al., 2021b]. During encoding, the hidden states are initialized using a single-layer feed-forward neural network for each RNN, and for decoding, an RNN is used with double attention. There are three computations followed in the proposed MMT system, computation of the hidden states, computation of the attention to the source language hidden state, and computation of the final hidden state from the attention state.

4.6 English to Mizo MMT Systems

To contribute to the development of the Multimodal Machine Translation system for English-Mizo, [Khenglawt et al., 2022] introduced the utilization of Mizo Visual Genome (MzVG). This proposed MMT model expands on the attention-based NMT model by incorporating spatial visual elements through the integration of a visual component [Calixto et al., 2017] [Calixto and Liu, 2017]. Visual features, both global and local, were extracted with the assistance of a pre-trained CNN model, VGG-19 [Nakazawa et al., 2021b]. The researchers employed two models in their experiment: the Bidirectional Recurrent Network (BRNN) and the Recurrent Neural Network (RNN). The results from the experiment demonstrated enhanced translation quality when compared to Text-Only NMT.

Performance evaluation of MMT systems is a crucial task both for determining the effectiveness of proposed MMT systems and optimizing their performances. The development of MMT approaches relies on assessing the quality of a system through systematic evaluation and performance comparison. However, the task of quality evaluations or translation comparisons is challenging due to the possibility of more than one correct translation. Moreover, it is difficult to differentiate the quality of translation when a sentence can be represented differently by preserving its meaning. Table 4 presents a precise summary of the MMT approaches proposed for Indian languages along with the techniques used, the information modalities, the domain of the data set, the pair of translation languages, the direction and the evaluation metrics. The column labeled 'Direction' (Dir.) specifies whether the task entails unidirectional (U) or bi-directional (B) translation.

5 Results and Discussion

BLEU and RIBES scores are predominantly utilized as evaluation metrics for MMT tasks. The challenging nature of performance analysis in MMT systems is attributed to the complexity of multiple modalities, their varying degrees of importance, and the diversity of data. It should be noted that there is no flawless single evaluation metric, and, typically, a combination of various metrics and human evaluation is employed to obtain a comprehensive and reliable performance analysis of MMT systems. Selection of evaluation metrics is dependent on the specific needs, objectives, information modalities, and domain used in the MMT system translation process. During the analysis of performance evaluation reported by promising MMT approaches, different approaches are experimented on different datasets in various settings, rendering the analysis process quite intricate. For instance, models like [Laskar et al., 2019] and [Parida et al., 2021a] are trained on similar size multi-domain datasets (HVG 1.0 and BnVG datasets) as shown in Table 3. However, due to the different language pairs for translation, drawing comparative conclusions about their performance is challenging.

It has also been observed that several proposed approaches used the same dataset to demonstrate the effectiveness of their model architectures. For example, the HVG 1.0 data set was used for the performance evaluation and their performance in terms of the BLEU and RIBES scores was reported by the models proposed in [Laskar et al., 2019, Meetei et al., 2019, Gupta et al., 2021, Parida et al., 2021a]. Given their experimental setup, it was observed that the approach presented in [Gupta et al., 2021] has outperformed the other approaches, and BLEU scores of 51.60 and 44.64 were achieved on challenge and evaluation datasets, respectively. Although the RIBES score of their experiment was not reported by the work of [Parida et al., 2021a], it was observed that the work of [Gupta et al., 2021] achieved RIBES scores of 0.86 and 0.82 in the challenge and evaluation datasets, respectively, which were 0.22 and 0.06 higher for the challenge and evaluation datasets, respectively, compared to the score of all approaches for the same dataset. Relatively poor performance was observed on the same data set for the approach proposed in [Meetei et al., 2019], with the lowest BLEU and RIBES scores among the approaches. Progress in the performance of the proposed approach in the HVG 1.0 dataset is shown in Figure 2. Furthermore, several MMT approaches [Laskar et al., 2020, Laskar et al., 2021a, Gain et al., 2021a, Shi and Yu, 2022, Laskar et al., 2022a, Parida et al., 2022] experimented with the HVG 1.1 dataset, which is an extension of the previous dataset, HVG 1.0, for performance evaluation. The performance of all the approaches surveyed in this category showed very close BLEU scores ranging from 33.57 as reported in [Laskar et al., 2020] to 42.70 as reported by [Shi and Yu, 2022] for both the challenge and the evaluation datasets. However, [Shi and Yu, 2022] demonstrated that improved transformer architectures with VGG11, VGG19, and ResNet50 models significantly improve the MMT translation system. Figure 3 provides a detailed analysis of the system performances experimented on the HVG 1.1 dataset.

In comparison to the approaches in the former dataset HVG 1.0, approaches in the latter dataset perform a bit lower considering the best-performed system. However, the latter approaches show more consistent performance than the former approaches, considering the overall performance of the compared approaches. To determine the best model for the considered data set (HVG and HVG 1.1), both the BLEU score and the RIBES score must be considered. From Table 5, it is evident that the work by [Gupta et al., 2021] has the highest BLEU score of 51.60, followed by [Shi and Yu, 2022] and [Parida et al., 2021a] with a score of 42.70 and 43.29 respectively. Upon examining the models that have both BLEU and RIBES scores listed, it is evident that the approach [Gupta

Sl. No.	Paper	Dataset	Domain	BLEU SCORE		RIBES	
				Challenge	Evaluation	Challenge	Evaluation
1	[Chowdhury et al., 2018]	Flickr30k	Multi-domain	N.A	24.20	N.A	N.A
2	[Meetei et al., 2021]	Synthetic	News	N.A	33.23	N.A	N.A
3	[Laskar et al., 2019]	HVG 1.0	Multi-domain	20.37	40.55	0.64	0.76
4	[Meetei et al., 2019]			12.58	28.45	0.48	0.63
5	[Gupta et al., 2021]			51.60	44.64	0.86	0.82
6	[Parida et al., 2021b]			43.29	42.11	N.A	N.A
7	[Laskar et al., 2020]	HVG 1.1		33.57	40.51	0.75	0.80
8	[Laskar et al., 2021a]			39.28	39.46	0.79	0.80
9	[Gain et al., 2021b]			37.50	42.47	0.79	0.81
10	[Shi and Yu, 2022]			42.29	42.70	N.A	N.A
11	[Laskar et al., 2022b]			39.30	39.40	0.79	0.80
12	[Parida et al., 2022]			39.10	42.00	N.A	N.A
13	[Singh et al., 2021b]	Multi-30K		48.60	N.A	N.A	N.A
14	[Parida et al., 2021a]	BnVG		26.80	43.50	N.A	N.A
15	[Laskar et al., 2022a]			28.70	43.90	0.69	0.78
16	[Parida et al., 2022]	MlVG		N.A	41.00	N.A	N.A
17	[Parida et al., 2021b]			N.A	31.30	N.A	N.A
18	[Laskar et al., 2021b]	AsVG		N.A	23.84	N.A	0.59
19	[Khenglawt et al., 2022]	MzVG		N.A	10.03	N.A	0.17

Table 5: Summary of the performance evaluation of MMT systems for Indian languages

et al., 2021] (with the HVG 1.0 dataset) has the highest RIBES score of 0.86, which is slightly higher than the approaches in [Laskar et al., 2020, Laskar et al., 2021a, Laskar et al., 2022a, Gain et al., 2021a].

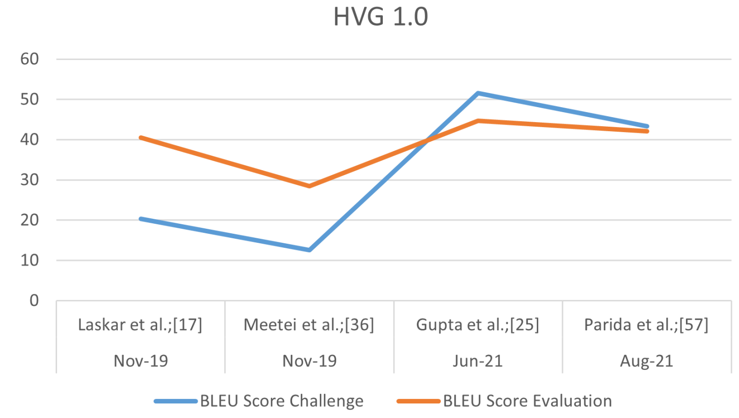


Figure 2: Performance of approaches on HVG 1.0

Therefore, based on the available data, it appears that the model proposed by [Laskar et al., 2020] is the best model among the listed ones, as it has the highest RIBES score and a competitive BLEU score. However, it is important to note that this conclusion is

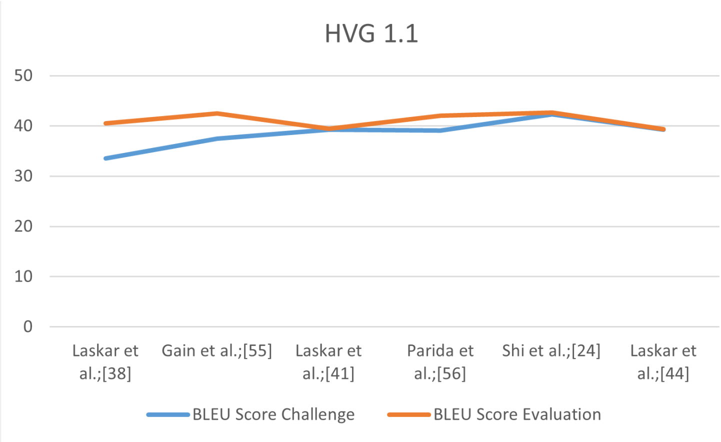


Figure 3: Performance of approaches on HVG 1.1

based on the limited information provided in the table, and other factors such as dataset size, model complexity, and training methods could also impact the overall performance of the models. Comparing the results of the HVG and HVG 1.1 datasets, we can see that the HVG 1.1 dataset has generally performed better than the HVG dataset across all the approaches, as evidenced by the higher BLEU and RIBES scores.

6 Research Challenges and Future Direction of Research

6.1 Research Challenges

Over the past few years, a promising development has been witnessed in multimodal machine translation for Indian languages. New research opportunities have been provided to work in this area due to the emergence of the MMT approach and the substantial demand for reliable information about Indian languages containing multiple modalities. Currently, researchers are focused on the application of machine learning algorithms to extract features from images and to use them to enhance text translation. However, challenges in multimodal machine translation in Indian languages are encountered, resulting in a decrease in the pace of the exploitation of research tasks. The research challenges in this domain are influenced by the following factors.

Representation of multimodal data: The possibility of capturing and gaining an in-depth understanding of events is offered by multimodal sources of information. Complementary and redundant information is contained within multiple modalities of information for representation. Therefore, the research challenge lies in the representation of multimodal data by filtering out noise and redundancy.

Embedding information from various modalities: The capturing of corresponding information and more robust predictions are enabled by the use of information from multiple modalities. However, this approach is susceptible to overfitting. Generalization rates may not be uniform across different modalities. Consequently, the training strategy may not be considered optimal most of the time. The combination of multiple modalities is sometimes challenged by the nature of heterogeneous information in multiple modalities.

Conveying knowledge between modalities: There are 22 regional languages recognized as scheduled languages by the Constitution of India. Morphologically rich, cohesive, and difficult-to-analyze structured sentences are found in the regional languages of India. A distinction between an aspirated and an unaspirated stop is not shown by the Dravidian dialect, such as Tamil, but many Indo-Aryan loanwords and a large number of compound characters are supported by other Dravidian languages, such as Kannada and Malayalam, resulting from the association of two consonant symbols [Kumar et al., 2015]. Although datasets are available in a few languages, resources for different modalities for many languages are very limited, including a lack of modality descriptions, noise, and ambiguous labels. In such scenarios, depending on the knowledge gained from resource-rich modalities is deemed quite beneficial. However, the transfer of knowledge between modalities presents a significant research challenge.

6.2 Future Direction of Research

Appropriate description of multimodal information: The identification of objects and the generation of relevant text [Summaira et al., 2021] are key factors in MMT. However, approaches for the generation of image, video, and audio descriptions in Indian languages have been proposed very sparingly or extremely rarely. Furthermore, the performance of existing approaches is not satisfactory due to the insufficient quality and quantity of training data. Therefore, the implementation of image, video, and audio description models on Indian language datasets may be considered a significant research perspective in the future.

Bridging modality gaps: The importance of information from each modality in the development of a robust translation system is emphasized in [Singh et al., 2021a]. Therefore, the methodology must be explored for the interpretation and integration of the most relevant correlation between the information modalities to achieve more accurate translations.

Standard dataset: India is a land of diversity, characterized by 22 registered languages and more than 1,500 other spoken languages. Regrettably, digital information is severely lacking, or in many cases nearly nonexistent, for the majority of Indian languages. This deficiency can be primarily attributed to the absence of script and document standardization, as noted in [Singh et al., 2021a]. Furthermore, even within the same language, variations exist in both written and spoken forms, featuring different grammatical standards or, in some cases, no standards at all. Consequently, a lack of annotated corpora is prevalent in most Indian languages, which, in turn, poses challenges for MMT tool development. Future research direction needs to invest in developing annotated corpora for various Indian languages for MMT tasks. As we look to the future, it is imperative for research to focus on the development of annotated corpora for diverse Indian languages to facilitate MMT tasks.

Handling translation variations: A sentence can be rewritten in various ways without altering the semantic structure [Rudrapal and Das, 2017]. Therefore, during translation, a sentence in the source language may have more than one possible translated sentence in the target language. Future research should explore paraphrase-generation techniques [Zhou and Bhat, 2021], as well as the handling of noisy paraphrases in the Indian language scenario.

Exploring new dimension in multimodalities for Indian languages: Other prominent research works on AI, such as sign language recognition [Albahri et al., 2023] and generation hold substantial potential for advancing multimodal machine translation for Indian languages. This research domain can facilitate more inclusive and efficient multimodal machine translation services by adapting sign language recognition systems [Qahtan et al., 2023, Şenol et al., 2024] for Indian sign language users, accurately identifying and translating content in various Indian languages, improving sign language recognition for Indian contexts, developing Indian script handwriting recognition [Kuncan et al., 2020] for data pre-processing, and optimizing data processing for improved translation services [Noyan et al., 2022], particularly in the linguistically diverse Indian landscape.

7 Conclusion

With the popularity of smart digital devices and various online platforms, people express views in multiple modalities. The possibility of apprehending an in-depth understanding and interpretation of text is offered by multimodal sources of information. As a result, an emerging field in research communities is represented by research in multimodal machine translation. The scope and necessity become manifold when the translation task pertains to Indian languages. Promising MMT approaches for major Indian languages have been proposed by different research communities in recent years. A comprehensive survey of these approaches is presented, along with techniques and datasets for the MMT task. The MMT task is defined, and existing works are categorized based on various techniques. The performance of the proposed approaches has also been briefly analyzed based on the claimed results and comparative evaluations. Finally, the challenges associated with the MMT task for Indian languages and some possible directions for future research in this domain are highlighted. An in-depth overview of research progress is provided by this survey work, and further research in MMT for more Indian or other languages is encouraged.

Acknowledgements

We would like to express our deep gratitude to anonymous reviewers whose meticulous assessments, constructive comments, and expertise significantly improved the quality and rigour of this manuscript. Their invaluable contributions have greatly contributed to the refinement and excellence of this work.

References

[Albahri et al., 2023] Albahri, O., AlSattar, H., Garfan, S., Qahtan, S., Zaidan, A., Ahmaro, I. Y., Alamoodi, A., Zaidan, B., Albahri, A., Al-Samarraay, M. S., et al. (2023). Combination of fuzzy-weighted zero-inconsistency and fuzzy decision by opinion score methods in pythagorean m-polar fuzzy environment: A case study of sign language recognition systems. *International Journal of Information Technology & Decision Making*, 22(04):1341–1369.

[Arslan et al., 2018] Arslan, H. S., Fishel, M., and Anbarjafari, G. (2018). Doubly attentive transformer machine translation. *arXiv preprint arXiv:1807.11605*.

[Bahdanau et al., 2015] Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015*.

[Barrault et al., 2018] Barrault, L., Bougares, F., Specia, L., Lala, C., Elliott, D., and Frank, S. (2018). Findings of the third shared task on multimodal machine translation. In *THIRD CONFERENCE ON MACHINE TRANSLATION (WMT18)*, volume 2, pages 308–327.

[Boonyuen et al., 2019] Boonyuen, K., Kaewprapha, P., Weesakul, U., and Srivihok, P. (2019). Convolutional neural network inception-v3: a machine learning approach for leveling short-range rainfall forecast model from satellite image. In *International Conference on Swarm Intelligence*, pages 105–115. Springer.

[Calixto et al., 2016] Calixto, I., Elliott, D., and Frank, S. (2016). Dcu-uva multimodal mt system report. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 634–638.

[Calixto and Liu, 2017] Calixto, I. and Liu, Q. (2017). Incorporating global visual features into attention-based neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 992–1003, Copenhagen, Denmark. Association for Computational Linguistics.

[Calixto et al., 2017] Calixto, I., Liu, Q., and Campbell, N. (2017). Doubly-attentive decoder for multi-modal neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1913–1924, Vancouver, Canada. ACL.

[Calixto et al., 2019] Calixto, I., Rios, M., and Aziz, W. (2019). Latent variable model for multimodal translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6392–6405.

[Chakravarthi et al., 2019] Chakravarthi, B. R., Priyadharshini, R., Stearns, B., Jayapal, A. K., Sridevy, S., Arcan, M., Zarrouk, M., and McCrae, J. P. (2019). Multilingual multimodal machine translation for dravidian languages utilizing phonetic transcription. In *Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages*, pages 56–63.

[Chen et al., 2018] Chen, Y., Liu, Y., and Li, V. (2018). Zero-resource neural machine translation with multi-agent communication game. In *Proceedings of the aaai conference on artificial intelligence*, volume 32.

[Chowdhury et al., 2018] Chowdhury, K. D., Hasanuzzaman, M., and Liu, Q. (2018). Multimodal neural machine translation for low-resource language pairs using synthetic data. In *Proceedings of the Workshop on Deep Learning Approaches for Low-Resource NLP*, pages 33–42.

[D'Andrea et al., 2021] D'Andrea, A., Caschera, M. C., Ferri, F., and Grifoni, P. (2021). Mubefe: Multimodal behavioural features extraction method. *JUCS - Journal of Universal Computer Science*, 27(3):254–284.

[Das and Singh, 2021] Das, R. and Singh, T. D. (2021). Image caption generation framework for assamese news using attention mechanism. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pages 231–239.

[Das and Singh, 2022] Das, R. and Singh, T. D. (2022). Assamese news image caption generation using attention mechanism. *Multimedia Tools and Applications*, 81(7):10051–10069.

[Elliott, 2018] Elliott, D. (2018). Adversarial evaluation of multimodal machine translation. In Riloff, E., Chiang, D., Hockenmaier, J., and Tsujii, J., editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2974–2978, Brussels, Belgium. Association for Computational Linguistics.

[Elliott et al., 2017] Elliott, D., Frank, S., Barrault, L., Bougares, F., and Specia, L. (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. ACL.

[Elliott et al., 2015] Elliott, D., Frank, S., and Hasler, E. (2015). Multilingual image description with neural sequence models. *arXiv preprint arXiv:1510.04709*.

[Elliott et al., 2016] Elliott, D., Frank, S., Sima'an, K., and Specia, L. (2016). Multi30k: Multilingual english-german image descriptions. *arXiv preprint arXiv:1605.00459*.

[Elliott and Kádár, 2017] Elliott, D. and Kádár, Á. (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.

[Fei et al., 2023] Fei, H., Liu, Q., Zhang, M., Zhang, M., and Chua, T.-S. (2023). Scene graph as pivoting: Inference-time image-free unsupervised multimodal machine translation with visual scene hallucination. *arXiv preprint arXiv:2305.12256*.

[Gain et al., 2021a] Gain, B., Bandyopadhyay, D., and Ekbal, A. (2021a). 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.

[Gain et al., 2021b] Gain, B., Bandyopadhyay, D., and Ekbal, A. (2021b). IITP at WAT 2021: System description for English-Hindi multimodal translation task. In *Proceedings of the 8th Workshop on Asian Translation*, pages 161–165, Online. ACL.

[Grubinger et al., 2006] Grubinger, M., Clough, P., Müller, H., and Deselaers, T. (2006). The iapr tc12 benchmark: A new evaluation resource for visual information systems. *Workshop Ontoimage*.

[Gupta et al., 2021] Gupta, K., Gautam, D., and Mamidi, R. (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.

[He et al., 2016] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

[Helcl et al., 2018] Helcl, J., Libovickỳ, J., and Variš, D. (2018). Cuni system for the wmt18 multimodal translation task. *arXiv preprint arXiv:1811.04697*.

[Honegger et al., 2021] Honegger, F., Feng, Y., and Rauterberg, M. (2021). Multimodality for passive experience: Effects of visual, auditory, vibration and draught stimuli on sense of presence. *JUCS - Journal of Universal Computer Science*, 27(6):582–608.

[Huang et al., 2016] Huang, P.-Y., Liu, F., Shiang, S.-R., Oh, J., and Dyer, C. (2016). Attention-based multimodal neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 639–645.

[Ive et al., 2019] Ive, J., Madhyastha, P., and Specia, L. (2019). Distilling translations with visual awareness. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6525–6538, Florence, Italy. Association for Computational Linguistics.

[Jain et al., 2020] Jain, M., Punia, R., and Hooda, I. (2020). Neural machine translation for tamil to english. *Journal of Statistics and Management Systems*, 23(7):1251–1264.

[Khenglawt et al., 2022] Khenglawt, V., Laskar, S. R., Manna, R., Pakray, P., and Khan, A. K. (2022). Mizo visual genome 1.0 : A dataset for english-mizo multimodal neural machine translation. In *2022 IEEE Silchar Subsection Conference (SILCON)*, pages 1–6.

[Klein et al., 2017] Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. (2017). OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

[Kumar et al., 2015] Kumar, A., Padró, L., and Oliver, A. (2015). Learning agglutinative morphology of indian languages with linguistically motivated adaptor grammars. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 307–312.

[Kumar and Lalithamani, 2022] Kumar, V. V. and Lalithamani, N. (2022). English to tamil multimodal image captioning translation. In *2022 IEEE World Conference on Applied Intelligence and Computing (AIC)*, pages 332–338. IEEE.

[Kuncan et al., 2020] Kuncan, M., Vardar, E., Kaplan, K., and Ertunç, H. M. (2020). Turkish handwriting recognition system using multi-layer perceptron. *Journal of Mechatronics and Artificial Intelligence in Engineering*, 1(2):41–52.

[Laskar et al., 2022a] Laskar, S. R., Dadure, P., Manna, R., Pakray, P., and Bandyopadhyay, S. (2022a). 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.

[Laskar et al., 2021a] Laskar, S. R., Khilji, A. F. U. R., Kaushik, D., Pakray, P., and Bandyopadhyay, S. (2021a). Improved english to hindi multimodal neural machine translation. In *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*, pages 155–160.

[Laskar et al., 2020] Laskar, S. R., Khilji, A. F. U. R., Pakray, P., and Bandyopadhyay, S. (2020). Multimodal neural machine translation for english to hindi. In *Proceedings of the 7th Workshop on Asian Translation*, pages 109–113.

[Laskar et al., 2021b] Laskar, S. R., Paul, B., Paudwal, S., Gautam, P., Biswas, N., and Pakray, P. (2021b). Multimodal neural machine translation for english–assamese pair. In *2021 International Conference on Computational Performance Evaluation (ComPE)*, pages 387–392. IEEE.

[Laskar et al., 2022b] Laskar, S. R., Singh, R., Karim, M. F., Manna, R., Pakray, P., and Bandyopadhyay, S. (2022b). Investigation of english to hindi multimodal neural machine translation using transliteration-based phrase pairs augmentation. In *Proceedings of the 9th Workshop on Asian Translation*, pages 117–122.

[Laskar et al., 2019] Laskar, S. R., Singh, R. P., Pakray, P., and Bandyopadhyay, S. (2019). English to hindi multi-modal neural machine translation and hindi image captioning. In *Proceedings of the 6th Workshop on Asian Translation*, pages 62–67.

[Lekshmy and Jayaraman, 2022] Lekshmy, H. and Jayaraman, S. (2022). English-malayalam vision aid with multi modal machine learning technologies. In *2022 ICICCS*, pages 1469–1476. IEEE.

[Li et al., 2022] Li, Y., Panda, R., Kim, Y., Chen, C.-F. R., Feris, R. S., Cox, D., and Vasconcelos, N. (2022). Valhalla: Visual hallucination for machine translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5216–5226.

[Libovický and Helcl, 2017] Libovický, J. and Helcl, J. (2017). Attention strategies for multi-source sequence-to-sequence learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 196–202, Vancouver, Canada. Association for Computational Linguistics.

[Libovický et al., 2018] Libovický, J., Helcl, J., and Mareček, D. (2018). Input combination strategies for multi-source transformer decoder. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 253–260, Brussels, Belgium. ACL.

[Lin et al., 2020a] Lin, H., Meng, F., Su, J., Yin, Y., Yang, Z., Ge, Y., Zhou, J., and Luo, J. (2020a). Dynamic context-guided capsule network for multimodal machine translation. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 1320–1329, New York, NY, USA. Association for Computing Machinery.

[Lin et al., 2020b] Lin, H., Meng, F., Su, J., Yin, Y., Yang, Z., Ge, Y., Zhou, J., and Luo, J. (2020b). Dynamic context-guided capsule network for multimodal machine translation. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 1320–1329, New York, NY, USA. Association for Computing Machinery.

[Liu et al., 2020] Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., and Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

[Long et al., 2020] Long, Q., Wang, M., and Li, L. (2020). Generative imagination elevates machine translation. *arXiv preprint arXiv:2009.09654*.

[Meetei et al., 2023] Meetei, L. S., Singh, S. M., Singh, A., Das, R., Singh, T. D., and Bandyopadhyay, S. (2023). Hindi to english multimodal machine translation on news dataset in low resource setting. *Procedia Computer Science*, 218:2102–2109.

[Meetei et al., 2019] Meetei, L. S., Singh, T. D., and Bandyopadhyay, S. (2019). Wat2019: English–hindi translation on hindi visual genome dataset. In *Proceedings of the 6th Workshop on Asian Translation*, pages 181–188.

[Meetei et al., 2021] Meetei, L. S., Singh, T. D., and Bandyopadhyay, S. (2021). Low resource multimodal neural machine translation of english-hindi in news domain. In *Proceedings of the First Workshop on Multimodal Machine Translation for Low Resource Languages (MMTLRL 2021)*, pages 20–29.

[Nakazawa et al., 2019] Nakazawa, T., Ding, C., Dabre, R., Kunchukuttan, A., Doi, N., Oda, Y., Bojar, O., Parida, S., Goto, I., and Mino, H. (2019). Proceedings of the 6th workshop on asian translation. In *Proceedings of the 6th Workshop on Asian Translation*, pages 1–225, Hong Kong, China. Association for Computational Linguistics.

[Nakazawa et al., 2021a] Nakazawa, T., Nakayama, H., Ding, C., Dabre, R., Higashiyama, S., Mino, H., Goto, I., Pa, W. P., Kunchukuttan, A., Parida, S., et al. (2021a). Overview of the 8th workshop on asian translation. In *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*, pages 1–45.

[Nakazawa et al., 2021b] Nakazawa, T., Nakayama, H., Goto, I., Mino, H., Ding, C., Dabre, R., Kunchukuttan, A., Higashiyama, S., Manabe, H., Pa, W. P., et al. (2021b). Proceedings of the 8th workshop on asian translation (wat2021). In *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*, pages 1–249, Online. Association for Computational Linguistics.

[Novotný et al., 2022] Novotný, V., Štefánik, M., Ayetiran, E. F., Sojka, P., and Řehůřek, R. (2022). When fasttext pays attention: Efficient estimation of word representations using constrained positional weighting. *JUCS - Journal of Universal Computer Science*, 28(2):181–201.

[Noyan et al., 2022] Noyan, T., Kuncan, F., Tekin, R., and Kaya, Y. (2022). A new content-free approach to identification of document language: Angle patterns. *Journal of the Faculty of Engineering and Architecture of Gazi University*, 37(3):1277–1292.

[Parida and Bojar, 2019] Parida, S. and Bojar, O. (2019). Hindi visual genome 1.0. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

[Parida and Bojar, 2021] Parida, S. and Bojar, O. (2021). Malayalam visual genome 1.0. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

[Parida et al., 2019] Parida, S., Bojar, O., and Dash, S. R. (2019). Hindi visual genome: A dataset for multi-modal english to hindi machine translation. *Computación y Sistemas*, 23(4):1499–1505.

[Parida et al., 2021a] Parida, S., Panda, S., Biswal, S. P., Kotwal, K., Sen, A., Dash, S. R., and Motlicek, P. (2021a). Multimodal neural machine translation system for english to bengali. In *Proceedings of the First Workshop on Multimodal Machine Translation for Low Resource Languages (MMTLRL 2021)*, pages 31–39.

[Parida et al., 2022] Parida, S., Panda, S., Grönroos, S.-A., Granroth-Wilding, M., and Koistinen, M. (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.

[Parida et al., 2021b] Parida, S., Panda, S., Kotwal, K., Dash, A. R., Dash, S. R., Sharma, Y., Motlicek, P., and Bojar, O. (2021b). Nlphut's participation at wat2021. In *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*, pages 146–154.

[Qahtan et al., 2023] Qahtan, S., Alsattar, H. A., Zaidan, A. A., Deveci, M., Pamucar, D., and Martinez, L. (2023). A comparative study of evaluating and benchmarking sign language recognition system-based wearable sensory devices using a single fuzzy set. *Knowledge-Based Systems*, 269:110519.

[Rudrapal and Das, 2017] Rudrapal, D. and Das, A. (2017). Measuring the limit of semantic divergence for english tweets. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 618–624.

[Sato et al., 2023] Sato, J., Caseli, H., and Specia, L. (2023). Choosing what to mask: More informed masking for multimodal machine translation. In Padmakumar, V., Vallejo, G., and Fu, Y., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 244–253, Toronto, Canada. Association for Computational Linguistics.

[Sen et al., 2022] Sen, A., Parida, S., Kotwal, K., Panda, S., Bojar, O., and Dash, S. R. (2022). Bengali visual genome: A multimodal dataset for machine translation and image captioning. In *Intelligent Data Engineering and Analytics*, pages 63–70. Springer.

[Sennrich et al., 2016] Sennrich, R., Haddow, B., and Birch, A. (2016). Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

[Şenol et al., 2024] Şenol, A., Kaya, M., and Canbay, Y. (2024). A comparison of tree data structures in the streaming data clustering issue. *Journal of the Faculty of Engineering and Architecture of Gazi University*, 39(1):217–231.

[Shah et al., 2016] Shah, K., Wang, J., and Specia, L. (2016). Shef-multimodal: Grounding machine translation on images. In *Proceedings of the First Conference on Machine Translation*, volume 2, pages 660–665. ACL.

[Shi and Yu, 2022] Shi, X. and Yu, Z. (2022). Adding visual information to improve multimodal machine translation for low-resource language. *Mathematical Problems in Engineering*, 2022.

[Silberer and Lapata, 2012] Silberer, C. and Lapata, M. (2012). Grounded models of semantic representation. In *Proceedings of the 2012 Joint Conference on EMNLP and ACL*, pages 1423–1433, Jeju Island, Korea. Association for Computational Linguistics.

[Singh et al., 2021a] Singh, M., Kumar, R., and Chana, I. (2021a). Machine translation systems for indian languages: review of modelling techniques, challenges, open issues and future research directions. *Archives of Computational Methods in Engineering*, 28:2165–2193.

[Singh et al., 2021b] Singh, S. M., Meetei, L. S., Singh, T. D., and Bandyopadhyay, S. (2021b). Multiple captions embellished multilingual multi-modal neural machine translation. In *Proceedings of the First Workshop on Multimodal Machine Translation for Low Resource Languages (MMTLRL 2021)*, pages 2–11.

[Singh and Singh, 2020] Singh, S. M. and Singh, T. D. (2020). Unsupervised neural machine translation for english and manipuri. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, pages 69–78.

[Singh and Hujon, 2020] Singh, T. D. and Hujon, A. V. (2020). Low resource and domain-specific english to khasi smt and nmt systems. In *2020 International Conference on Computational Performance Evaluation (ComPE)*, pages 733–737. IEEE.

[Stein et al., 2009] Stein, B. E., Stanford, T. R., and Rowland, B. A. (2009). The neural basis of multisensory integration in the midbrain: its organization and maturation. *Hearing research*, 258(1-2):4–15.

[Sulubacak et al., 2020] Sulubacak, U., Caglayan, O., Grönroos, S.-A., Rouhe, A., Elliott, D., Specia, L., and Tiedemann, J. (2020). Multimodal machine translation through visuals and speech. *Machine Translation*, 34:97–147.

[Summaira et al., 2021] Summaira, J., Li, X., Shoib, A. M., Li, S., and Abdul, J. (2021). Recent advances and trends in multimodal deep learning: A review.

[Tamura et al., 2020] Tamura, H., Hirasawa, T., Kaneko, M., and Komachi, M. (2020). TMU Japanese-English multimodal machine translation system for WAT 2020. In *Proceedings of the 7th Workshop on Asian Translation*, pages 80–91, Suzhou, China. Association for Computational Linguistics.

[Vaswani et al., 2018] Vaswani, A., Bengio, S., Brevdo, E., Chollet, F., Gomez, A., Gouws, S., Jones, L., Kaiser, Ł., Kalchbrenner, N., Parmar, N., Sepassi, R., Shazeer, N., and Uszkoreit, J. (2018). Tensor2Tensor for neural machine translation. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 193–199, Boston, MA. Association for Machine Translation in the Americas.

[Wang et al., 2020] Wang, X., Wu, J., Chen, J., Li, L., Wang, Y.-F., and Wang, W. Y. (2020). Vatex: A large-scale, high-quality multilingual dataset for video-and-language research.

[Wu et al., 2021] Wu, Z., Kong, L., Bi, W., Li, X., and Kao, B. (2021). Good for misconceived reasons: An empirical revisiting on the need for visual context in multimodal machine translation. *arXiv preprint arXiv:2105.14462*.

[Yao and Wan, 2020] Yao, S. and Wan, X. (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.

[Yin et al., 2020] Yin, Y., Meng, F., Su, J., Zhou, C., Yang, Z., Zhou, J., and Luo, J. (2020). A novel graph-based multi-modal fusion encoder for neural machine translation. *arXiv preprint arXiv:2007.08742*.

[Yin et al., 2023] Yin, Y., Zeng, J., Su, J., Zhou, C., Meng, F., Zhou, J., Huang, D., and Luo, J. (2023). Multi-modal graph contrastive encoding for neural machine translation. *Artificial Intelligence*, 323:103986.

[Young et al., 2014] Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the ACL*, 2:67–78.

[Zhang and Zong, 2016] Zhang, J. and Zong, C. (2016). Exploiting source-side monolingual data in neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545.

[Zhang et al., 2019] Zhang, Z., Chen, K., Wang, R., Utiyama, M., Sumita, E., Li, Z., and Zhao, H. (2019). Neural machine translation with universal visual representation. In *International Conference on Learning Representations*.

[Zhao et al., 2022] Zhao, Y., Komachi, M., Kajiwara, T., and Chu, C. (2022). Region-attentive multimodal neural machine translation. *Neurocomputing*, 476:1–13.

[Zhou and Bhat, 2021] Zhou, J. and Bhat, S. (2021). Paraphrase generation: A survey of the state of the art. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5075–5086.

[Zhou et al., 2018] Zhou, M., Cheng, R., Lee, Y. J., and Yu, Z. (2018). A visual attention grounding neural model for multimodal machine translation. *arXiv preprint arXiv:1808.08266*.