Journal of Scientific & Industrial Research
Vol. 85, April 2026, pp. 312-324
DOI: 10.56042/jsir.v85i4.22732

Seeing Beyond Text: A Visual-Linguistic Dataset and Multimodal Framework for English-Hindi Video-Guided Translation

Binnu Paul[1,2]*, Dwijen Rudrapal[1], Kunal Chakma[1] & Anupam Jamatia[1]
[1]Department of CSE, NIT Agartala 799 046, India
[2]Department of Computer Science and Information Technology, SHUATS, Prayagraj 211 008, Uttar Pradesh, India

Received 16 August 2025; revised 26 December 2025; accepted 25 March 2026

Despite the progress in Neural Machine Translation (NMT), translating ambiguous and context rich content remains a major challenge, especially in low-resource language pairs like English-Hindi. Traditional NMT systems often fail to solve these challenges due to their reliance on textual data alone. Multimodal approaches, particularly those incorporating visual context, offer promising solutions to the task by resolving linguistic ambiguities. To address this, a novel solution is introduced through a Visual Scene-Aware Hindi Subtitles Dataset (VISA-HIN), designed specifically for English-Hindi Video-Guided Multimodal Machine Translation (VMMT). This dataset aligns English subtitles with corresponding video frames and provides Hindi translations. Alongside the dataset, this study propose a video-guided MMT framework that leverages visual cues to enhance translation quality. The results of the experiments show the potential of scene aware information to improve contextual understanding and fluency in English-to-Hindi translation, paving the way for more robust and accurate multimodal translation systems in low-resource settings.

Keywords: Artificial intelligence, Cross-lingual learning, Indian languages, Machine learning, Natural language processing

Introduction

Multimodal Machine Translation (MMT)integrates additional modalities like images, audio, and video with text to enrich the translation process.[1–4] MMT leverages cross-modal fusion to produce contextually grounded outputs following earlier research.[5] This addresses unimodal MT's core challenges like, semantic ambiguity, data scarcity in low-resource languages, and false perception or interpretation, and morphological complexities. Multiple modalities of information help to overcome these challenges by resolving ambiguities with idioms or implied actions and further enhance through unsupervised alignment, mitigating OOV issues and biases, with contrastive learning frameworks.[6] Video holds dominant importance among the modalities in MT tasks for its temporal richness, capturing gesture, action, sequences, and interactions that static images cannot. Inclusion of videos improve accuracy in narrative content by allowing dynamic disambiguation via frame progression.[7,8] This is crucial for applications span multimedia subtitling for e-learning tutorials, news dissemination, e-commerce descriptions, and promoting inclusive digital access. For example, consider the English sentence: "He is backing up while giving instructions." When presented as text alone, this sentence is semantically ambiguous. The phrase "backing up" may denote physical backward movement ("*वह पीछे की ओर हट रहा है*"), reversing a vehicle ("*वह वहां को रिवर्स में लेजा रहा है*"), or metaphorical support ("*वह समर्थन कर रहा है*"). The accompanying action "giving instructions" does not, by itself, resolve which interpretation is intended. Text or a static image cannot capture the temporal dynamics required to disambiguate these meanings. In contrast, video provides continuous cues such as motion direction, object interaction, and the coordination of movement with speech. These temporal signals allow the system to correctly infer the intended action and produce an accurate Hindi translation, demonstrating the essential role of video-based context in multimodal translation.

In India, a linguistically diverse nation with plentiful coexisting languages, high quality translation is critical for bridging communication gaps and delivering services across multilingual populations. In spite of Hindi's prominence in India, large-scale video-text aligned multimodal datasets for Hindi are limited.

*Author for Correspondence
E-mail: binnupaul16@gmail.com

Existing resources focus on image-text pairings, leaving a gap in video-grounded translation research for many Indian languages like Tamil, Bengali, and Marathi. This scarcity poses a challenge, as many expressions in these languages carry multiple meanings, requiring dynamic visual cues for accurate disambiguation. To address this, the Visual Scene-Aware Hindi Subtitles (VISA-HIN) dataset is introduced for English-Hindi translation. By integrating dynamic visual cues with text, VISA-HIN enhances translation models to produce linguistically correct, contextually rich, and culturally relevant translations, advancing multimodal translation research.

The contributions of this current research work are as follows:

- VISA-HIN Dataset: This study unveil VISA-HIN, a fresh dataset tying English-Hindi subtitles at the paragraph level to their video counterparts, tuned for scene-aware insights.
- Video-Guided MMT Approach: Proposal of a new Video-Guided Multimodal Machine Translation (VMMT) method for English-to-Hindi translation that leverages temporal video context for enhanced semantic understanding.
- Contextual Disambiguation: Demonstration of how video sequences improve the resolution of lexical ambiguities that are otherwise difficult to handle through text alone.
- Support for Low-Resource Languages: Exploration of the method's broader applicability to other low-resource Indian languages, reinforcing the value of multimodal data in handling morphological complexity.

Related Works

This section provides a thorough examination of works on machine translation. The study of Multimodal Machine Translation (MMT) has advanced to use additional modalities, like video and images, in addition to text, to improve translation accuracy, especially for languages with limited resources. While traditional text-based machine translation models have shown significant improvements for languages with abundant resources, they still struggle to perform well in languages like Hindi due to the absence of extensive parallel datasets.[9] To address these issues, recent advances in multimodal machine translation have introduced encoder–decoder calibration for enhanced image–text fusion and modal contrastive learning within an end-to-end framework to improve text–image alignment.[10,11] Some of the initial research investigated the creation of synthetic data and datasets tailored for low-resource environments.[12,13] Subsequent efforts highlighted improved model architectures and the incorporation of visual grounding.[14,15] Recent research proposed have continued to enhance translation accuracy and expand domain reach, while proposed models in focused on multilingual contexts and challenges related to low-resource scenarios.[16–20] One notable dataset, the Hindi Visual Genome (HVG) dataset[21] facilitates text and image-based MMT for the English-Hindi machine translation. The utilization of visual and video data in Multimodal Machine Translation (MMT) has been demonstrated to enhance translation quality.[7,22] A large-scale multilingual video description dataset is introduced for advancing video-guided MMT for English-Chinese language pair.[8] The samples in this dataset are collected from YouTube and annotated with multilingual descriptions. Since inception of Video-Guided Machine Translation Challenge 2020, various video-guided MMT system proposed to generate robust translated text. The proposed system utilizes key frame-based video features with positional encoding, whereas a dual-model framework combining a hierarchically attentive recurrent neural network and a shallow model is used for video-guided translation.[23,24] In a recent work, a more challenging, ambiguous subtitle dataset for visual scene-aware machine translation between Japanese and English is introduced which have more than one translation with different meanings.[25] Alongside this dataset, the work also proposed a model incorporating selective attention, frame attention loss, and ambiguity augmentation to enhance video-based disambiguation.

Recent research in Video-Guided Multimodal Machine Translation (VMMT) underscored the importance of pre-training with lexically diverse video datasets, demonstrating significant improvements in common benchmarks, showed that visual data and broader contextual signals significantly enhance translation effectiveness, especially in specialized content.[26,27] Conversely, the method proposed challenged the efficacy of multimodal input, arguing that visual components offer limited advantages unless the source text is flawed or incomplete.[2] Together, these studies indicate that the effectiveness of visual modalities is conditional, particularly when factors like domain adaptation, contextual alignment, and dataset quality are sufficiently managed. Some other research works as proposed in emphasized the role of motion-

aware and action-centric representations in aligning language with dynamic visual content, Motion-focused datasets and architectures, showed that explicitly modelling temporal dynamics improves vision–language alignment.[28–30]

Video-guided MMT model for the low-resource English-Hindi language pair is first introduced in the work of Meetei *et al.,* introducing a synthetic multimodal dataset comprising text and video modalities specifically for this purpose.[31] In contrast, the proposed dataset, VISA-HIN, is specifically tailored to address ambiguities in video-guided translation. It focuses on linguistic ambiguities—especially polysemy and omission—by providing parallel subtitles that are clearly grounded in visual context. The VISA-HIN dataset correlates each 10-second segment with a single sentence. This approach guarantees a precise one-to-one relationship between visual elements and textual descriptions, minimizing potential noise and redundancy. Unlike earlier methods that predominantly utilized static image features (such as ResNet), the current study leverage dynamic scene context through I3D and R-CNN features, combined using an adaptive cross-modal attention mechanism. This approach is especially vital for Indian languages, where ambiguities arising from polysemy and context sensitivity are common, making text-only models less effective at accurately resolving uncertainties.

Dataset preparation

Data Collections

The VISA-HIN dataset was formed by first translating the original Visual Scene-Aware Subtitles Dataset (VISA) into Hindi and then cleansing the translations for accuracy and relevance.[25] The original VISA dataset contains subtitles from various movies and TV series, which are designed for machine translation tasks. VISA is divided into two sections: Polysemy and Omission. The Polysemy category deals with subtitle ambiguity from words with multiple meanings and the Omission category addresses ambiguities caused by missing subjects or objects, leading to incomplete sentence understanding. To support multimodal learning, each subtitle line is paired with a corresponding video clip. Each video clip is standardized to a 10-second duration, recorded at frames per second (fps), capturing 5 seconds before and after the midpoint of the subtitle's time span. All audio contents were removed from the video clips to focus only on visual cues. This pre-determined length guarantees uniform visual context throughout samples and makes it easier to synchronize with translations at the sentence level. It also identifies important actions within a defined timeframe, allowing for improved alignment between visual elements and language components, especially important for Hindi, where the structure and meaning of sentences are intensely influenced by context.

Rationale and Importance of the Dataset

The VISA dataset has significant potential and suitability for multimodal machine translation from English to Hindi. The dataset covers conversational and narrative styles found in movies or shows, useful for training models on realistic, context-rich data. The core strength of VISA is its explicit pairing of ambiguous subtitles with visual scenes. The visual features learned from the English-Japanese pairs, particularly how visual cues resolve ambiguity. The same is equally important for Indian languages, including Hindi which exhibits ambiguities with gender, honorifics, and context-dependent word senses that text-only MT struggles with. The dataset VISA-HIN offers a valuable benchmark for advancing Video-Guided MMT and promotes research in low-resource, multilingual NLP settings.

Annotation

In order to prepare the VISA-HIN dataset, a semi-automatic multi-stage annotation approach is employed to ensure accurate annotations and high-quality translations. The process is described below:

1. Initial translation: The annotation process began with the automatic translation of 39,880 English subtitles from the original VISA dataset. This step utilized three state-of-the-art machine translation models: translation model proposed in The Tatoeba Translation Challenge,[32] Google Translate[a], and SeamlessM4T[b]. These models were chosen for their capacity to handle both general and contextually ambiguous sentences. For each subtitle, translations were generated by all three models. The best version was selected based on linguistic accuracy and contextual alignment.

[a]https://translate.google.co.in/?sl=en&tl=hi&op=translate
[b]https://huggingface.co/facebook/seamless-m4t-v2-large

2. Post-editing: The translation output from the above steps is further corrected by human annotators for generating most corrected translation. Three bilingual annotators, fluent in both English and Hindi, carried out comprehensive manual post-editing of the machine-generated Hindi subtitles. To facilitate human post-editing, a locally hosted web annotation tool is developed. The tool displays the original English subtitle, its Hindi translation, and the corresponding video clip. An intuitive user interface that enabled annotators to efficiently review and refine translations.

For unambiguous sentences, machine translations required only minor edits. However, ambiguous or visually dependent subtitles often needed substantial rewriting to capture the intended message. The video context played a pivotal role in disambiguation and helped ensure semantic fidelity.

Quality Verification

To assess annotation consistency, this study used the ROUGE-1 score, a metric that evaluates unigram overlap between annotations.[33] ROUGE-1 is well-suited for translation evaluation, as it reflects lexical similarity between reference and predicted translation. The analysis under this study revealed an inter-annotator agreement of 91.86%, indicating high consistency in how different annotators interpreted and translated the source text. This high score affirms the quality, reliability, and appropriateness of the dataset for research in machine translation. This degree of agreement demonstrates a strong correlation in word choice and fundamental sentence meaning. It also confirms the quality and dependability of the annotations. To further ensure annotation quality, a secondary validation step was conducted where a random sample of 3,980 subtitles (approximately 10% of the dataset) was manually reviewed by linguistic experts. This verification step helped to eliminate remaining errors and enhanced the overall reliability of the VISA-HIN dataset.

Statistical Insights

A comparative statistics of English and Hindi subtitles in the dataset is reported in Table 1. This comparison offers valuable insights into the distinct language patterns present in the corpus. However, the number of unique word types is higher in English compared to Hindi, indicating a broader vocabulary usage in the English text despite Hindi having higher values in other categories. Hindi often requires more words to express the same concepts due to differences in grammar, morphology, and expression styles.

Table 1 — Statistical summary of word counts for English and Hindi subtitles

Statistic	English	Hindi
Total number of sentences	39,880	39,880
Average sentence length (tokens)	8.869	9.846
Total number of tokens	353,727	392,691
Median sentence length (tokens)	8.0	9.0
Total characters	1,458,095	1,589,532
Singleton token types	8,201	6,951
Vocabulary size (unique token types)	16,459	14,641
Average token length (characters)	3.43	3.26
Standard deviation of sentence length	3.41	4.35
Minimum / maximum sentence length (tokens)	1 / 34	1 /44
Median (50th percentile)	7.0	8.0

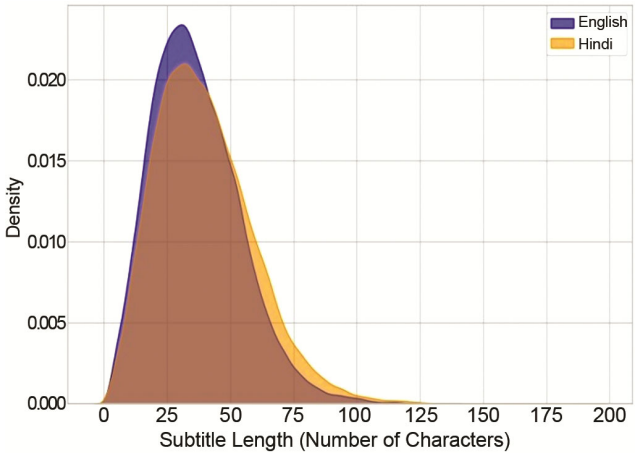

Fig. 1 — Kernel density estimation (KDE) plot depicting the distribution of subtitle lengths

Kernel Density Estimation (KDE) plots of subtitle lengths measured in character count and presents in Fig. 1. The plot reveals that both English and Hindi follow a similar distribution curve, with peaks around 20–40 characters. Hindi subtitles show a longer tail, with the orange curve extending further right, indicative of longer average sentence lengths. This also reveals a common pattern in machine translation that low-resource or morphologically rich language like Hindi often require more characters or tokens to convey equivalent meanings.

Finally, Table 2 offers the volume of data in training, test and validation set for both English and Hindi along with word statistics for the Omission and Polysemy sets. The table emphasizes important metrics such as maximum, minimum, and average word counts, in addition to total tokens and vocabulary size.

Table 2 — Word statistics and dataset splits for omission and polysemy sets

Set	Data	Lang.	Wordstats			Tokens	Vocab	Split Size
			Max	Min	Avg			
Omission	Test	En	28	2	8.767	8767	1808	1,000
		Hi	25	2	9.672	9672	1923	1,000
	Training	En	38	2	8.895	153119	10396	17,214
		Hi	49	1	9.903	170474	9517	17,214
	Validation	En	22	2	9.012	9012	1869	1,000
		Hi	28	1	10.084	10084	1909	1,000
Polysemy	Test	En	25	2	8.832	8832	1872	1,000
		Hi	33	1	9.794	9794	1899	1,000
	Training	En	37	1	8.842	165047	11206	18,666
		Hi	40	1	9.788	182710	10204	18,666
	Validation	En	23	2	8.950	8950	1882	1,000
		Hi	29	1	9.957	9957	1951	1,000

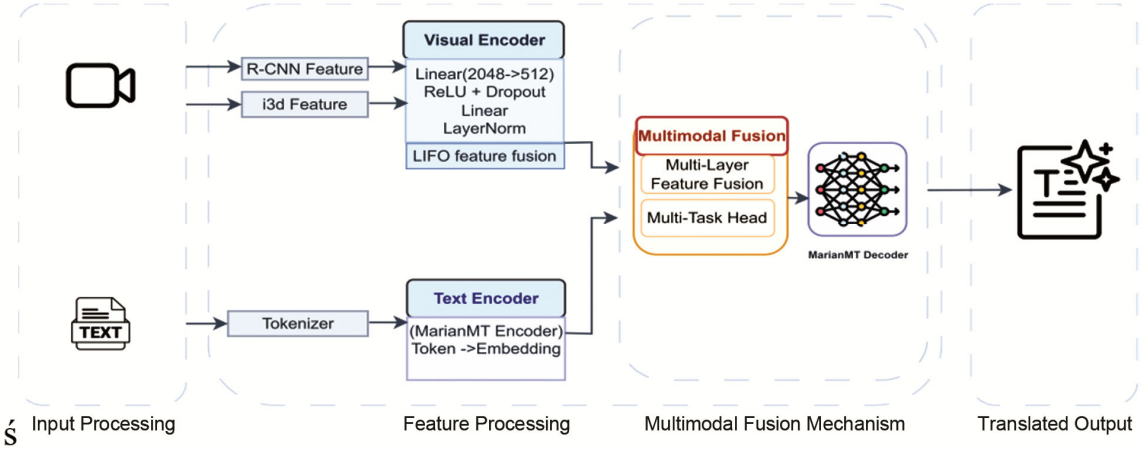

Fig. 2 — Proposed multimodal machine translation framework

Methodology

Given a source text sentence $T = \{t_1, t_2, t_3, \ldots\ldots, t_n\}$ and its corresponding video context V, the objective of this study is to generate a contextually-aware translation $\hat{Y} = \{y_1, y_2, y_3, \ldots\ldots, y_m\}$ in the target language. The key challenge lies in effectively fusing textual and visual information to produce translations that are not only linguistically accurate but also contextually appropriate given the visual scene.

This was formulated as a multimodal sequence-to-sequence learning problem, where the translation model f_θ is conditioned on both textual and visual inputs. The multimodal translation model predicts the target sentence $\hat{Y}$ from textual input (T) and visual representation (V):

$$\hat{Y} = f_\theta(T, V) = \arg max_Y P(Y \mid T, V;\ \theta) \quad \ldots (1)$$

where, θ represents the learnable parameters of the multimodal translation system.

The approach proposed in this study consists of five main components: *Input Processing*, *Feature Processing*, *Feature Fusion Module*, *Multimodal Fusion Module*, and *Translator Module*. To the best of our knowledge, this is the first study to incorporate not only LIFO-based feature fusion, but also multi-layer feature fusion and multi-task learning for multimodal machine translation involving text and video. The overall architecture of the system is illustrated in Fig. 2.[14]

Input Processing Module

The input processing module is an essential part of the Multimodal Machine Translation (MMT) system, designed to manage and process both textual and visual inputs. The video undergoes feature extraction to capture relevant visual data, while the text is prepared for translation. These features are then aligned and merged for further processing in the translation workflow.

***Text Input*:** The text input processing adheres to the default neural machine translation pipeline with tokenization and extraction of embeddings. For a given input text sequence, the system initially uses the Stanford Tokenizer[c] to transform the text into token IDs and create attention masks.

$$T = \text{tokenizer}(x) \rightarrow \{input\_ids, attention\_mask\} \quad \ldots (2)$$

***Video Input*:** The video inputs used in this research are directly taken from the VISA dataset, where all video clips have undergone pre-processing and standardization. Each clip is a 10 second MP4 file at 25 frames per second, resized to 224 × 224 pixels, extracted from the midpoint of subtitle segments, covering 5 seconds prior and following. These clips are provided without audio to emphasize the visual context. The selection of a 10-second duration in the VISA dataset was made to guarantee sufficient temporal context for understanding ambiguous subtitles. These pre-processed clips are used along with their corresponding subtitle pairs, for the Multimodal Translation (MMT) tasks.

Feature Processing

To enable effective VMMT, this study carefully selected and analysed three following feature classes critical for integrating textual and visual information to resolve linguistic ambiguities, such as polysemy (words with multiple meanings) and omission (missing subjects or objects), thereby enhancing translation accuracy for English to Hindi tasks. Each feature class, its extraction process, and its importance to the VMMT task are described below.

- ***Contextual Features*:** Rich contextual features were extracted from the source text using the MarianMT model. In the VISA HIN setting, contextualization is derived purely from auxiliary linguistic cues within the source sentence, rather than from neighbouring video frames or temporal visual windows. This model generates contextualized embeddings that capture both semantic (meaning) and syntactic (structural) information of text. These embeddings represent each word or token in the sentence based on its surrounding context, allowing the model to distinguish semantic variations of words in sentences, such as differentiating the word "bank" in "river bank" versus "money bank". contextualized embeddings serve as high-quality input features for the subsequent translation task, facilitating better disambiguation, coherence, and fluency in the target language. The textual input is processed using the standard MarianMT encoder pipeline, which includes positional encoding, multi-head self-attention, and a feed-forward network to obtain contextualized token embeddings.
- ***Spatial Semantics (Object Presence and Detection):*** Object-based spatial semantics provide critical visual context for resolving ambiguities in text. For instance, in the sentence "She sat on the bench", the presence of a bench object in the video can clarify whether "bench" refers to a seat or a workbench. By identifying objects and their positional relationships, RCNN features help the model ground textual references in the visual scene, enhancing translation accuracy for nouns and phrases prone to mis-interpretation. This is particularly valuable for Hindi, where word order and context play significant roles in meaning. Spatial semantics are captured through object presence and detection using a Region-based Convolutional Neural Network (RCNN). RCNN generates feature vectors representing object categories and their relative positions, which are then normalized and projected into a lower dimensional space for integration with textual embeddings.
- ***Temporal Dynamics (Action Recognition Features):*** Action recognition features capture the dynamic aspects of videos, such as movements or activities, which are important for interpreting verbs and action related ambiguities in text. For example, in the sentence "He is cutting", "cutting" refers to slicing food or editing a film by identifying relevant actions (e.g., chopping motions or computer use). Temporal dynamics are extracted using Inflated 3D ConvNet (I3D), pretrained on the Kinetics-400 dataset, which contains 400 action classes. The model outputs feature vectors representing action categories (e.g., running, cooking), which are normalized and projected for compatibility with textual and spatial features.

LIFO Feature Fusion

The Last-In-First-Out (LIFO) feature fusion strategy was employed, to combine the heterogeneous video features into a unified representation. Each feature type f_i is first projected to a common embedding space of dimension $d = 512$:

[c]https://nlp.stanford.edu/software/tokenizer.shtml

$$h_i = W_i f_i + b_i \quad \ldots (3)$$

where, f_i denotes the $(\mathrm{i}^{\{\mathrm{th}\}})$input feature vector, h_i represents the projected feature embedding, $W_i \in \mathbb{R}^{d \times d_i}$ and $b_i \in \mathbb{R}^d$are learnable projection parameters for feature type i. The projected features are then processed through a Transformer encoder layer to capture inter-feature dependencies:

$$v = TransformerEncoder(h_1, h_2, h_3, h_4) \quad \ldots (4)$$

where, $v \in \mathbb{R}^d$ denotes the unified multimodal video representation, which is obtained by mean pooling across the feature dimensions. Compared to simple concatenation, LIFO fusion enables ordered and progressive integration of multimodal features, allowing visual cues to refine textual representations more effectively and reducing cross modal interference.

Multimodal Fusion Module

This module comprises with following two sub-modules,

1. **Multi-Layer Feature Fusion Module**: The Multi-Layer Feature Fusion module is the core of the proposed model, designed to hierarchically integrate textual and visual information. It employs a three-layer structure, where each layer models increasingly complex interactions between the modalities. To begin, text and video features are projected into a common hidden space. At each layer, cross modal attention is used were text attends to video and vice versa, followed by self-attention within each modality. These steps allow the model to align low level correspondences in early layers and capture high level semantics in deeper layers. A learnable gating mechanism was used to control the flow of information between the modalities, and the final output is a weighted combination of layer outputs. This hierarchical approach results in richer, more context-aware multimodal representations, which are passed on to downstream tasks.

2. **Multi-Task Head Module**: The Multi-Task Head module improves the quality of the learned multimodal representations by introducing auxiliary learning objectives. Specifically, tasks like Scene Classification, Action Recognition and Translation Quality Estimation are incorporated. Each task receives the fused multimodal representation and processes it through a task specific neural branch. An auxiliary feature quality estimation module is then applied over the fused representation (v):

$$q(v) = \sigma(W_q^T . ReLU(W_{q1} v + b_{q1}) + b_q) \quad \ldots (5)$$

W_{q1} and W_q^T are learnable weight matrices, b_{q1} and $b_q$represent bias parameters, ReLU is the activation function, and σdenotes the sigmoid activation function. This design ensures that the model learns robust representations that incorporate semantic, temporal, emotional, and quality-related cues from the multimodal input.

Translator Module

The Multi-Layer Multi-Task Translator integrates all components into a cohesive system. It uses a pretrained MarianMT encoder to process the input text and enriches these representations using the fused multimodal features (v).

The enhanced text representation $h_{enhanced}$is computed via attention-based integration:

$$h_{enhanced} = h_{text} + \gamma \cdot Attention(h_{text}, v, v) \quad \ldots (6)$$

where, h_{text} denotes the textual encoder representation, *Attention* denotes the cross-modal attention mechanism, and γ is a learnable scaling parameter. In the attention formulation $\mathrm{Attention}(\mathrm{Q}, \mathrm{K}, \mathrm{V})$, the first (v) is used as the Key to compute attention scores, while the second (v) serves as the Value that provides the information to be aggregated. This enhanced representation is then passed to the decoder for translation generation.

The enhanced multimodal representation is then used by the decoder to compute the primary translation loss:

$$L_{\{trans\}} = -\sum_{\{t=1\}}^{\{N\}} \log\left(P\left(Y_t \mid y < t, h_{enhanced}\right)\right) \quad \ldots (7)$$

where, Y_t represents the target token at decoding step (t), (y<t) denotes all previously generated target tokens, and is the enhanced multimodal representation used for prediction.

The overall loss optimized during training combines the main translation objective and the auxiliary tasks:

$$\mathscr{L}_{system} = \mathscr{L}_{trans} + \lambda \sum_{i=1}^{4} \beta_i L_{aux}^{(i)} \quad \ldots (8)$$

where, $\mathscr{L}_{trans}$ denotes the primary translation loss, $L_{aux}^{(i)}$represents the auxiliary loss corresponding to task (i), β_i are task-specific weighting coefficients, and λcontrols the contribution of auxiliary objectives during optimization

This architecture enhances translation quality by aligning cross-modal signals, learning auxiliary semantics, and effectively integrating visual context into the translation process.

Algorithm 1: LIFO-Based Multimodal Translation

Input: Source sentence (X), video features (F=$\{f_1, f_2, f_3, f_4\}$)

Output: Translated sentence $\hat{Y}$

Step 1: Tokenization (T);

Tokenize the input sentence X to obtain input IDs and attention mask;

Step 2: Text Encoding;

Encode the tokenized text using the MarianMT encoder to obtain the text representation h_{text};

Step 3: Video Feature Projection;

Project all video features f_i into a shared embedding space to obtain projected embeddings h_i;

Step 4: LIFO Video Fusion;

Stack projected video features h_1, h_2, h_3, h_4 in LIFO order and apply a Transformer encoder followed by mean pooling to obtain the video representation (v);

Step 5: Multimodal Fusion;

Fuse text and video representations (h_{text}, v) using multi-layer cross-modal attention and gated integration;

Step 6: Context Enhancement;

Enhance the text representation $h_{enhanced}$ using attention over the fused multimodal representation (v);

Step 7: Translation Generation;

Generate the translated sentence $\hat{Y}$ using the MarianMT decoder;

Step 8: Model Optimization;

Train the model using a weighted combination of translation and auxiliary task losses $L_{\{trans\}}$;

return $\hat{Y}$

Experiments and Result Analysis

To assess the proposed VMMT framework for translating from English to Hindi, two datasets were used. The first is VISA-HIN, a new dataset created specifically for this study, consisting of 39,880 video clips matched with parallel English Hindi subtitles. The evaluation focused on the validation split of VISAHIN, which contains 2,000 video clips 1,000 featuring polysemous expressions and 1,000 aimed at evaluating the omission handling capabilities, as illustrated in Table 2. Additionally, this study used the publicly accessible VaTeX v1.1 dataset, which provides 3,000 video clips with corresponding English captions in its validation set.[8] From this dataset, 2,000 clips were chosen for evaluation to assess the model's generalization performance on out-of-domain data. The same semi-automated translation method as described in Annotation Section was applied to this subset of VaTeX to maintain consistency in performance evaluation and comparison. The quality of automatic translation was assessed using two well-known metrics: BLEU and chrF.

Experimental Setup

The proposed VMMT system was implemented using PyTorch and trained on an NVIDIA A100 GPU (40GB). The translation model utilizes the MarianMT transformer encoder-decoder structure from the[35] checkpoint, which was selected for its robust baseline performance in translating between English and Hindi. To enhance the translation process with visual context, a hierarchical visual encoding approach was incorporated. Specifically, I3D was used to capture temporal motion features, while RCNN was applied to key frames to extract object level spatial features. Both sets of features were ℓ_2 normalized and linearly projected before being input into multi-head attention and gated fusion layers.

Each video sample was normalized to a duration of 10 seconds at 25 frames per second, resized to 224 × 224 pixels, and had audio removed to concentrate on visual semantics. Translation outputs were produced using beam search (with a beam width of 3) while capping the maximum length at 50 tokens. To minimize repetition, a 3-gram penalty was employed during the decoding process. Training was carried out on the VISA-HIN dataset with a batch size of 32, a learning rate of 5×10^{-5}, and a dropout rate of 0.3. Early stopping was activated based on improvements in validation BLEU scores.

Results

The current $MMT_{Proposed}$ approach was experimented and compared with the following models:

- NMT Baseline: A text-only transformer model.[32]
- MMTmp model (MMT_{mp}): A video-guided model using mean-pooled visual features and Byte Pair Encoding (BPE) for tokenization.
- MMTs model (MMT_s): Similar to MMTmp model but uses segment-based visual feature extraction, offering more localized visual grounding.
- Multimodal LLM (MMT_{LLM}): For comparative analysis, GPT-4o[36], a multimodal large language

Table 3 — Evaluation of NMT and VMMT systems in terms of BLEU and chrF scores

Models	T_{visa}		T_{vtx}	
	BLEU	chrF	BLEU	chrF
NMT	45.7	0.62	25.6	0.46
MMT_{mp}	46.2	0.63	21.0	0.44
MMT_s	47.9	0.66	30.5	0.53
MMT_{LLM}	48.5	0.68	30.3	0.53
$MMT_{Proposed}$	48.8	0.69	30.5	0.55

model that can process both text and images, was utilized.

A comparative performance analysis of the proposed approach and other experimented models on validation split of VISA-HIN (T_{visa}) and translated subset of VaTeX(T_{vtx}) dataset is presented in Table 3.

The evaluation of the NMT and VMMT systems, as shown in Table 3, emphasizes the superior effectiveness of the proposed multimodal translation framework, $MMT_{Proposed}$, in terms of automatic and manual evaluation metrics. On the T_{visa} dataset, $MMT_{Proposed}$ obtains the highest BLEU score of 48.8 and a chrF score of 0.69, surpassing the baseline NMT system, which achieves a BLEU score of 45.7 and a chrF score of 0.62. Among other VMMT systems, MMT_{LLM} and MMT_s performs admirably with a BLEU score of 48.5 and 47.9 and a chrF score of 0.68 and 0.66, yet it still does not reach the performance of $MMT_{Proposed}$.

For T_{vtx} dataset, both $MMT_{Proposed}$ and MMT_s reach the highest BLEU score of 30.5, but $MMT_{Proposed}$ edges out MMT_s with a higher chrF score of 0.55 compared to 0.53. The baseline NMT system, on the other hand, shows relatively poorer results on this dataset, attaining BLEU and chrF scores of 25.6 and 0.46, respectively. The MMT_{mp} system displays notably limited performance on T_{vtx} with BLEU and chrF scores of 21.0 and 0.44, illustrating its difficulties with out-of-domain data.

Manual evaluation of the translation outputs on the VISA test set showed that the text-only translation system produced 1,292 correct translations, 490 partially correct translations, and 218 incorrect translations. In comparison, the multimodal translation system achieved 1,403 correct translations, 448 partially correct translations, and 149 incorrect translations. Among the multimodal translations, 163 cases specifically resolved ambiguities present in the text-only outputs, and 1,851 translations were judged to be reasonable overall.

Table 4 — Example of Highlighting VMMT's Contextual Enhancement

Text – Gota gunshot wound to the upper left bicep.

Reference - ऊपरी बाए बाइसेप्स में एक गोली का घाव है |

NMT - ऊपर बाई ओर की ओर एक गोली मार दी गई |

MMT_{prop} - ऊपरी बाए बाइसेप्स में एक गोली का घाव है |

$VMMT_{imp}$ - ऊपरी बाए हाथ के हिस्से में एक गोली लगी है |

$VMMT_S$ - ऊपरी बाए हिस्से में गोली का निशान है |

Table 5 — Translation Improvement by Removing Ambiguity

Text - Park on the street.

Reference - सड़क पर पार्क करे।

NMT - सड़क पर पार्क है।

$MMT_{Proposed}$ - सड़क पर पार्क करे।

$VMMT_{imp}$ - सड़क पर पार्क करे।

$VMMT_S$ - सड़क पर पार्क करे।

Readability was evaluated on a subset of the challenge set using Average Sentence Length (ASL), Average Word Length (AWL), and language-model–based perplexity (PPL). The proposed MMT system yields a lower ASL (9.90) than the NMT baseline (10.08), indicating simpler sentence structures, while its PPL (8.86) is comparable to NMT (8.79) and closely aligns with the Golden reference (8.87), reflecting improved fluency. Although the AWL of MMT (3.43) is marginally higher than that of NMT (3.39), it remains closer to the Golden reference (3.44), suggesting more appropriate lexical selection. Overall, the consistent shift of MMT outputs toward the Golden reference across all metrics demonstrates improved readability over the NMT baseline. An in-depth analysis of the translation output demonstrates how the proposed Video-guided Multimodal Machine Translation ($MMT_{Proposed}$) model significantly improves translation quality by resolving contextual ambiguities that text-only NMT systems struggle with. In Table 4, the sentence "Got a gunshot wound to the upper left bicep" is misinterpreted by the NMT system as an act of shooting rather than an injury. $MMT_{Proposed}$ accurately captures the passive wound description, matching the reference translation. The phrase "Park on the street", in Table 5 is incorrectly

interpreted by the NMT system as a location statement. In contrast, $MMT_{Proposed}$ correctly translates it as an imperative, showing superior contextual understanding enabled by multimodal inputs.

Ablation Study

To assess the contribution of individual components within the proposed Multimodal Machine Translation (MMT) framework T_{visa}, a series of ablation experiments were conducted on a validation set. The baseline NMT model was first evaluated, which relies solely on textual input and does not incorporate visual information, and compared it against the full MMT system that integrates contextual linguistic features, spatial semantics (object presence and detection), and temporal dynamics (action recognition features). The results indicate that each component contributes complementary benefits: temporal features are particularly effective for short, action-centric sentences (e.g., "She is running"), whereas spatial semantics play a more prominent role in resolving object-related references (e.g., "She is holding the glass"). This analysis highlights how different visual cues address distinct types of ambiguities in multimodal translation.

***Effectiveness of Contextual Features*:** This element captures cues from surrounding sentences within a video transcript and leverages a wider linguistic context from nearby frames to improve translation uniformity. Eliminating contextual features coerce the model to depend solely on the individual sentence without any accompanying cues or language influences from the video. The lack of contextual features led to a decrease in BLEU from 48.7 to 47.2 and chrF from 0.69 to 0.66, indicating that even translations of single sentences gain from a minimal understanding of context across a video segment, thereby enhancing referent clarity and word selection.

***Effectiveness of Spatial Semantics (Object Presence and Detection)*:** Spatial semantics offer insights into the objects detected in the video and their interrelations, linking terms like "man", "table" or "ball" to visual proof. Disabling this component removes object presence vectors, hindering the model's ability to utilize visual grounding for clarifying ambiguous nouns and references within the scene. The absence of spatial semantics resulted in a decrease in BLEU from 48.7 to 46.9 and chrF from 0.69 to 0.65, demonstrating that object detection plays a crucial role in improving translation quality by supplying the model with specific visual references for more precise word choice.

***Effectiveness of Temporal Dynamics (Action Recognition Features*)** Temporal dynamics facilitate the understanding of actions and motion patterns in visual frames, enabling the model to functionally translate verbs and phrases related to actions. This ablation caused a decrease in BLEU from 48.7 to 47.1 and chrF from 0.69 to 0.66, underscoring the importance of temporal signals in grasping the dynamic elements of visual scenes, which results in translations that are more accurate and contextually appropriate. Overall, the results indicate that each component meaningfully contributes to performance improvements, with the full multimodal model outperforming the text-only baseline (BLEU: 45.7 → 48.7, chrF: 0.62 → 0.69) by effectively leveraging context, positional grounding, and temporal understanding from video input.

Error Analysis

Despite the notable improvements achieved by the proposed multimodal machine translation (MMT) system, several limitations persist. These issues primarily stem from the inherent complexity of language, which can challenge even visually grounded translation models. The key error types observed are as follows:

***Inadequate Handling of Idiomatic Expressions*:** The MMT model tends to translate idiomatic phrases too literally, lacking semantic depth. For instance, in the below given example, the idiom

"kicked the bucket" is translated literally as "*बाल्टी मारी*" (hit the bucket) which holds no significance in Hindi. This shortcoming arises from the model's dependency on surface-level token matching rather than deeper semantic understanding. Since idiomatic expressions often lack visual representation, adding video doesn't assist the translation. The example of translation is,

English Input: "He kicked the bucket right after his big promotion, and everyone was taken aback."

MMT Output: "उसने बाल्टी मारी उसके बड़े प्रमोशनके तुरंत बाद, और हर कोई पीछे लिया गया।"

***Lexical Choice and Tone Misalignment*:** The model also struggles with appropriate lexical selection and tone preservation, especially in emotionally charged contexts. The expression "everyone was taken aback" in the above example indicates surprise and emotional reaction, but the MMT output "*हर कोई पीछे लिया गया।*" (everyone was taken aback) comes across as

Table 6 — Impact of individual multimodal features on translation performance

Feature configuration	Included	BLEU	chrF
Text-only baseline	No	45.7	0.62
Full multimodal model	Yes (All)	48.7	0.69
Contextual features removed	No	47.2 (−1.5)	0.66 (−0.03)
Spatial semantics removed	No	46.9 (−1.8)	0.65 (−0.04)
Temporal dynamics removed	No	47.1 (−1.6)	0.66 (−0.03)

incompetent and lacking in emotion. This shortcoming is noted where the model overly emphasizes tangible visual aspects and fails to adequately capture emotional meanings.

***Grammatical and Syntactic Errors Due to Modality Conflict*:** The integration of visual and textual modalities sometimes results in syntactic errors, such as verb-object mismatches or incorrect clause structures. This indicates a necessity for improved cross-modal alignment techniques and temporal attention strategies to maintain syntactic accuracy.

Quantitative and Qualitative Analysis

The quantitative results further substantiate the qualitative observations regarding the strengths and limitations of the proposed Multimodal Machine Translation (MMT) framework. As shown in Table 3, the full multimodal model consistently outperforms the text-only NMT baseline on the VISA-HIN dataset, achieving an improvement of +3.1 BLEU points (45.7 → 48.8) and +0.07 chrF (0.62 → 0.69). These gains indicate that incorporating visual information provides measurable benefits beyond purely textual cues, particularly for visually grounded and context-dependent translations. The ablation results presented in Table 6 provide a more fine-grained understanding of the contribution of individual multimodal components. Removing spatial semantics leads to the largest performance degradation (−1.8 BLEU, −0.04 chrF), highlighting the importance of object presence and positional grounding for resolving visually ambiguous noun phrases. In comparison, removing temporal dynamics results in a drop of −1.6 BLEU, demonstrating the role of action recognition features in translating short, action-centric expressions. The removal of contextual linguistic features also causes a noticeable decline (−1.5 BLEU), confirming that strong textual contextualization remains essential even in multimodal settings. This trend aligns with the qualitative analysis, where spatial features improve object-related translations (e.g., "get in"→"*अंदर जाना*", while temporal cues help disambiguate action-oriented expressions (e.g., *"aim"* in weapon-related contexts).

Spatial semantics are critical for resolving noun ambiguity. For example, "Being a head" is rendered as "*सिर बनना*" by NMT, while multimodal grounding enables the correct leadership interpretation "*प्रधान बनना*". On the other way temporal features are essential for accurate action and aspect modeling. Like, the translation of "She is aiming," by the ablated model is "*निशाना लगाने*", missing the ongoing nature of the action, whereas the multimodal system generates "*निशाना बनाने*".

Conclusions

The current research advances the field of VMMT by introducing VISA-HIN, a visual scene-aware subtitle dataset tailored for English to Hindi translation. Alongside this, the proposed model, effectively integrates visual cues through adaptive cross-modal attention to enhance translation quality in context-sensitive scenarios. Performance evaluation of developed dataset demonstrates that the proposed VMMT model consistently outperforms the baseline NMT and existing VMMT models and improves domain adaptation, contextual coherence, and ambiguity resolution. Manual evaluations further validate its strength to capture nuanced meanings, particularly where text-only systems fail. This work has direct applicability to real-world scenarios such as multilingual video subtitling, content localization, assistive technologies, and instructional video translation, where visual grounding plays a critical role in resolving linguistic ambiguity. The future direction of current research work may focus on addressing challenges in handling emotional tone, syntactic precision, further, the VISA-HIN dataset may be expanded to more modalities such as audio or speech to enrich context modeling. In addition, strategies to improve performance in out-of-domain scenarios and idiom translation through pre-training in diverse multimodal corpora remain vital areas for exploration.

Acknowledgement

We used Grammarly and ChatGPT (version 5.2) solely for grammar checking and language editing. These tools did not replace our intellectual contributions, critical analysis, or scholarly judgment. All content was carefully reviewed and revised by us, and we take full responsibility for the originality, accuracy, and integrity of the submitted manuscript.

Funding Information

No funds or grants were received for conducting this study.

Conflict of Interest

The authors declare that they have no conflict of interest.

Data Availability Statement

The dataset used in this research work is publicly available in this link https://github.com/ku-nlp/VISA. Collected data are further annotated to fit in our current research work which may be available on request to the corresponding author.

References

1 O'Sullivan C, Introduction: multimodality as challenge and resource for translation, *J Spec Transl*, **20** (2013) 2–14.

2 Cui S, Duan K, Ma W & Shinnou H, Does multimodal machine translation can improve translation performance?, *Neural Comput Appl*, **36**(22) (2024) 13853–13864, doi:10.1007/s00521-024-09705-y.

3 Parida S, Bojar O & Dash S R, Hindi visual genome: a dataset for multimodal English to Hindi machine translation, *Comput Sist*, **23(4)** (2019) 1499–1505.

4 Chowdhury K D, Hasanuzzaman M & Liu Q, Multimodal neural machine translation for low-resource language pairs using synthetic data, *Proc Workshop Deep Learn Approaches Low-Resour NLP*, (2018) 33–42, doi: 10.18653/v1/W18-3405.

5 Elliott D, Frank S, Sima'an K & Specia L, Multi30K: multilingual English–German image descriptions, *Proc 5th Workshop Vision Lang*, (2016) 70–74, doi: 10.18653/v1/w16-3210.

6 Kang L, Huang L, Peng N, Zhu P, Sun Z, Cheng S, Wang M, Huang D & Su J, Bigvideo: a large-scale video subtitle translation dataset for multimodal machine translation, *Findings of the Association for Computational Linguistics: ACL,(2023)* 8456-8473, doi: 10.18653/v1/2023.findings-acl.535.

7 Sanabria R, Caglayan O, Palaskar S, Elliott D, Barrault L, Specia L & Metze F, How2: a large-scale dataset for multimodal language understanding, *arXiv*, (2018) 1811.00347.

8 Wang X, Wu J, Chen J, Li L, Wang Y-F & Wang W Y, Vatex: A large-scale, high-quality multilingual dataset for video-and-language research, *Proc IEEE Int Conf Comput Vis*, (2019) 4581–4591, doi: 10.1109/ICCV.2019.00468.

9 Vaswani A, Attention is all you need, *Adv Neural Inf Process Syst*, **30** (2017) 5998–6008.

10 Tayir T, Li L, Li B, Liu J & Lee K A, Encoder–decoder calibration for multimodal machine translation, *IEEE Trans Artif Intell*, **5**(8) (2024) 3965–3973, doi: 10.1109/TAI2024.3354668.

11 Ma C, Han X, Wu L, Zhang Y, Zhao Y, Zhou Y & Zong C, Modal contrastive learning based end-to-end text image machine translation, *IEEE ACM Transaction in Audio Speech Language Process*, **32** (2023) 2153–2165, doi:10.1109/TASLP.2023.3324540.

12 Meetei L S, Singh T D & Bandyopadhyay S, Wat2019: English–Hindi translation on Hindi visual genome dataset, *Proc 6th Workshop Asian Transl*, (2019) 181–188, doi: 10.18653/v1/D19-5224.

13 Laskar S R, Singh R P, Pakray P & Bandyopadhyay S, English to Hindi multimodal neural machine translation and Hindi image captioning, *Proc 6th Workshop Asian Transl*, (2019) 62–67, doi: 10.18653/v1/D19-5205.

14 Laskar S R, Khilji A F U R, Pakray P & Bandyopadhyay S, Multimodal neural machine translation for English to Hindi, *Proc 7th Workshop Asian Translation*, (2020) 109–113, doi: 10.18653/v1/2020.wat-1.11.

15 Laskar S R, Khilji A F U R, Kaushik D, Pakray P & Bandyopadhyay S, Improved English to Hindi multimodal neural machine translation, *Proc 8th Workshop Asian Transl*, (2021) 155–160, doi: 10.18653/2021.wat-1.17.

16 Gupta K, Gautam D & Mamidi R, Vita: visual-linguistic translation by aligning object tags, *Proc 8th Workshop Asian Translation (2021)* 166-173, doi: 10.18653/v1/2021.wat-1.19.

17 Parida S, Panda S, Kotwal K, Dash A R, Dash S R, Sharma Y, Motlicek P & Bojar O, Nlphut's participation at WAT2021, *Proc 8th Workshop Asian Translation*, (2021) 146–154, doi: 10.18653/v1/2021.wat-1.16.

18 Gain B, Bandyopadhyay D & Ekbal A, Experiences of adapting multimodal machine translation techniques for Hindi, *Proc First Workshop MMT Low Resource Lang*, (2021) 40–44, doi: 10.26615/978/954/452-073-1_007.

19 Meetei L S, Singh T D & Bandyopadhyay S, Low resource multimodal neural machine translation of English–Hindi in news domain, *Proc First Workshop Multimodal Mach Transl Low Resour Lang*, (2021) 20–29, doi: 10.26615/978-954-452-073-1_004.

20 Meetei L S, Singh S M, Singh A, Das R, Singh T D & Bandyopadhyay S, Hindi to English multimodal machine translation on news dataset in low resource setting, *Procedia Comput Sci*, **218** (2023) 2102–2109, doi: 10.1016/j.procs.2023.01.186.

21 Sulubacak U, Caglayan O, Gronroos S-A, Rouhe A, Elliott D, Specia L & Tiedemann J, Multimodal machine translation through visuals and speech, *Mach Transl*, **34** (2020) 97–147, doi: 10.1007/s10590-020-09250-0.

22 Hirasawa T, Yang Z, Komachi M & Okazaki N, Keyframe segmentation and positional encoding for video-guided machine translation challenge 2020, *arXiv*, (2020) 2006.12799, doi: 10.48550/arXiv.2006.12799.

23 Yang Z, Hirasawa T, Komachi M & Okazaki N, Why videos do not guide translations in video-guided machine translation? An empirical evaluation of video-guided machine translation dataset, *J Inf Process*, **30** (2022) 388–396.

24 Li Y, Shimizu S, Gu W, Chu C & Kurohashi S, Visa: An ambiguous subtitles dataset for visual scene-aware machine translation, *Proc Thirteenth Language Resources Evaluat Conf* (2022) 6735–6743.

25 Shurtz A, Sorenson L & Richardson S D, The effects of pretraining in video-guided machine translation, *Proc LREC-COLING*, (2024) 15888–15898.

26 Lv J, Chen J, Long Z, Fu X & Chen Y, Topicvd: a topic-based dataset of video-guided multimodal machine translation for documentaries, *Int Conf Appl Natural Language Info Syst (2026)* 442–456, doi: 10.1007/978-3-031-97141-9_30.

27 Ding H, Liu C, He S, Ying K, Jiang X, Loy C C & Jiang Y-G, Mevis: a multi-modal dataset for referring motion expression video segmentation, *IEEE Trans Pattern Anal Mach Intell*, **47(12)** (2025) 11400–11416, doi:10.1109/ TPAMI.2025.3600507.

28 Zhao Z, Huo Y, Yue T, Guo L, Lu H, Wang B, Chen W & Liu J, Efficient motion-aware video MLLM, *Proc IEEE ConfComput Vis Pattern Recognit*, (2025) 24159–24168.

29 Lee J, Chung H & Kim B-H, Aligning video LLMs with video diffusion for fine-grained temporal understanding, *arXiv*, (2024).

30 Meetei L S, Singh A, Singh T D & Bandyopadhyay S, Do cues in a video help in handling rare words in a machine translation system under a low-resource setting? *Nat Lang Process J*, **3** (2023) 100016, doi: 10.1016/j.nlp.2023.100016..

31 Tiedemann J, The Tatoeba translation challenge–realistic datasets for low resource and multilingual machine translation, *Proc 5th Conf Mach Transl*, (2020) 1174–1182, doi: 10.18653/v1/2020.wmt-1.139.

32 Lin C-Y, Rouge: A package for automatic evaluation of summaries, *Proc ACL Workshop Text Summarization*, (2004) 74–81.

33 Cai Y, Li X, Zhang Y, Li J, Zhu F & Rao L, Multimodal sentiment analysis based on multi-layer feature fusion and multi-task learning, *Sci Rep*, **15**(1) (2025) 2126, doi: 10.1038/s41598-025-85859-6.

34 Junczys-Dowmunt M, Grundkiewicz R, Dwojak T, Hoang H, Heafield K, Neckermann T, Seide F, Germann U, Aji A F, Nikolay B, Martins A F T & Birch A, Marian: Fast neural machine translation in C++, *Proc ACL (2018)* 116-121, doi: 10.18653/v1/P18-4020.

35 Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman F L, McGrew B & others, Gpt-4 technical report, *arXiv*, (2023) 2303.08774, doi: 10.48550/ arXiv.2303.08774.