Video-guided Machine Translation for Indic Languages

Pinaki Das[1,2,3]*, **Srikar Ayyagari**[1]*, **Raj Dabre**[4]†, **Thuy-Trang Vu**[3], **Virendra Singh**[1], **Pushpak Bhattacharyya**,[1] **Gholamreza Haffari**[3]

[1]Indian Institute of Technology Bombay, India [2]IITB-Monash Research Academy, India
[3]Monash University, Australia [4]AI4Bharat, IIT Madras

Correspondence: {pinaki, srikar, pb, viren}@cse.iitb.ac.in, prajdabre@gmail.com, {pinaki.das, trang.vu1, gholamreza.haffari}@monash.edu

Abstract

Video-guided machine translation (VMT) translates text by leveraging the context provided by the accompanying video. This additional contextual information helps improve translation quality by disambiguating ambiguous words or explicitly indicating attributes such as the colour or gender of objects and people. Prior works have extensively explored VMT, however, most approaches train models from scratch, which requires large amounts of data. Additionally, the majority of VMT research has focused on the English–Chinese language pair. In this paper, we explore VMT training strategies that use only a limited amount of data and extend the task to Indic languages, specifically Hindi, Marathi and Odia. We also propose a synthetic data framework that is effective for training VMT models in low-resource language settings. Our results demonstrate an improvement of **+1.75** BLEU compared to baseline models. Our code[1] and dataset[2] are publicly available.



Figure 1: Task of VMT: Odia translation of an English source sentence with and without the video (ambiguous span in red). Without video, blue one becomes ନୀଳ ରଙ୍ଗ (*nīḷa ranga*, blue colour). With video, it becomes ନୀଳ ରଙ୍ଗର ମଣିଟି (*nīḷa rangara maṇiṭi*, blue gem). The bottom box is a back-translation of the with-video output.

1 Introduction

Multimodal machine translation (MMT) improves the quality of the translated output as it is conditioned on the information derived from both textual and visual inputs. Initial MMT models focused mainly on image-guided machine translation, where images provided the necessary context to produce accurate translations. VMT, on the other hand, aims to translate sentences by conditioning them with the context derived from the accompanying videos.

VMT takes advantage of the temporal and multimodal nature of the videos, which combine textual, auditory and visual data into a single cohesive source of information. Unlike static images, videos provide better information about the sequence of events, actions and interactions. This makes VMT particularly effective for tasks such as subtitle translation of videos (see Figure 1), where the subtitles occur in the particular span or context of the corresponding video.

Prior works (Kang et al., 2023; Guan et al., 2025; Li et al., 2023b) have primarily trained VMT models from scratch using triplets of parallel sentences paired with corresponding videos. However, existing publicly available datasets for VMT are largely restricted to the English–Chinese (En–Zh) direction, which constrains the extension of VMT to other language families. Training from scratch further demands a substantial amount of high-quality triplet data. Yet, curating such resources is both labour-intensive and time-consuming. The challenge is exacerbated for low-resource languages, where assembling sufficient triplet data in itself is

* Equal contribution.
† Work done during employment at NICT, Japan.
[1] daspinaki/Video-Guided-MT-Indic
[2] Pinaki007/Ambiguous-VMT

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 1214–1235
November 6-10, 2026 ©2026 Association for Computational Linguistics

very difficult.

This work addresses these challenges by focusing on extending the VMT task to Indic languages by introducing an evaluation set for English-Hindi, English-Marathi and English-Odia directions. We also explore effective training strategies for VMT using a limited number of triplet data by leveraging strong text-only pre-trained translation models. Furthermore, we also propose a synthetic data pipeline to create visually grounded translations using multimodal large language models (MLLMs) for training VMT models in low-resource languages.

Our contributions are summarised as follows:

- **Dataset Contribution:** We construct an evaluation benchmark dataset for Hindi, Marathi, and Odia for the Video-guided Machine Translation task, each containing 158 sentences. The data is released under the CC BY 4.0 license (Section 4.1).
- **Synthetic Data Generation Framework for VMT:** We introduce a framework for generating visually grounded translations for low-resource languages using open-source MLLMs. (Section 3.3).
- **Training Strategies:** We explore video-guided machine translation frameworks that integrate video representations into a pre-trained text translation model through LoRA-based fine-tuning, and a symmetric contrastive loss designed to align video and text embeddings for improved semantic consistency. (Section 3.1).

2 Related Work

2.1 Image-guided Machine Translation

Initial MMT research largely focused on image-guided machine translation (IMT). The Multi30K dataset (Elliott et al., 2016) was instrumental in accelerating the development of MMT models, providing approximately 30,000 parallel sentence pairs in English and German, with corresponding images. Retrieval-based IMT (Fang and Feng, 2022; Tang et al., 2022; Zhang et al., 2020) retrieves relevant images according to the source sentence. Text-to-image guided machine translation (Calixto et al., 2019; Li et al., 2022a; Yuasa et al., 2023; Guo et al., 2023) focuses on generating or imagining synthetic images conditioned on the input and translating using them. IMT has also been explored using multitask training strategies (Elliott and Kádár, 2017; Nishihara et al., 2020; Zhu et al., 2023) where subtasks like image captioning and translation are involved. It has also been investigated through contrastive loss objectives (Cheng et al., 2023; Chen et al., 2023; Ji et al., 2022), and in unsupervised settings (Fu et al., 2023; Fei et al., 2023; Surís et al., 2022). To further take advantage of the large availability of data and compute, many pretraining approaches and pre-trained models (Futeral et al., 2024; Sato et al., 2023; Yawei and Fan, 2021) have been used for the IMT task. Specifically for IMT tasks, various efforts in Indic languages have been undertaken, such as the Hindi Visual Genome dataset (Sanayai Meetei et al., 2019) and Odia Visual Genome dataset (Parida et al., 2025), along with WAT workshops focused on English-Indic Multimodal translation (Parida and Bojar, 2025).

2.2 Video-guided Machine Translation

The use of video as an additional modality for machine translation was explored by Wang et al. (2019), who introduced a multilingual video description dataset for benchmarking video-guided translation. They used dual attention and dual encoders for text and video modalities. Hirasawa et al. (2020) introduced order-aware frames using positional embeddings to capture better the temporal order of the frames of the video. Gu et al. (2021), inspired by HAN (Miculicich et al., 2018), divided inputs into motion representation and spatial representation, which in turn is divided into a multilevel hierarchy representation of object level, frame level and video-level. Li et al. (2023b) introduced frame-level attention and ambiguity-aware attention. Li et al. (2023a) use video as a pivot between languages. Kang et al. (2023) introduced additional contrastive loss-based training. Guan et al. (2025) used LLMs to extract fine-grained audio features from the videos in natural language and further used them to train an Enc-Dec to enhance the audio awareness in VMT. Lv et al. (2025) integrated a selective attention module and the bidirectional attention module by taking inspiration from Li et al. (2021) and Tang et al. (2022). Their architecture utilises two encoders, each for video and source text and fuses the obtained representations using a cross-modal bidirectional attention mechanism.

Datasets: VaTex (Wang et al., 2019), EVA (Li

et al., 2023b), BIGVIDEO (Kang et al., 2023), MAD-VMT (Shurtz et al., 2024), Trifine (Guan et al., 2025) and TopicVD (Lv et al., 2025) are datasets for En-Zh direction whereas How2 (Sanabria et al., 2018) and VISA (Li et al., 2022b) are in En-Pt and En-Ja direction respectively. EVA also has an additional direction of En-Ja.

The datasets here, therefore, are limited to only a few language pairs, making VMT inaccessible to other language families. Also, since previous work in VMT mostly focuses on training from scratch with a large amount of data, it is important that we look at methods and techniques for low-resource languages with efficient training. Therefore, in our work, we introduce a new evaluation set of Indic languages and try to explore efficient ways to train VMT models with limited or no initial available data.

3 Methodology

To perform the video-guided machine translation, we assume access to a multimodal dataset $\mathcal{D} = \{V_i, x_i, y_i\}_{i=1}^{N}$ of N triplets, each triplet consists of a raw video clip V_i, the sentence x_i in the source language corresponding to the video clip, and the sentence y_i in the target language.

Our goal is to learn the function $f : (V_i, x_i) \mapsto y_i$ that generates fluent and accurate target-language sentences by leveraging both the textual and visual modalities. We extend strong pre-trained text-only translation models to the video-guided machine translation task, while relying on only limited multimodal data. We also refer to this multimodal data as *triplet data* in the upcoming sections.

3.1 Leveraging Pre-Trained Text-Only Models

State-of-the-art text-only translation models already capture robust linguistic knowledge. Therefore, instead of training a multimodal model from scratch for VMT, we build upon these pre-trained models and introduce video features in a parameter-efficient manner.

Multimodal fusion can be performed in two general ways:
- *Early fusion*, where both modalities are jointly encoded, and
- *Late fusion*, where video features are injected into the model at later stages (e.g., in the decoder).

We adopt the late fusion strategy where the encoder processes only the source text, while the decoder attends to both text and video representations. This preserves the strengths of the text encoder while allowing video context to influence translation.

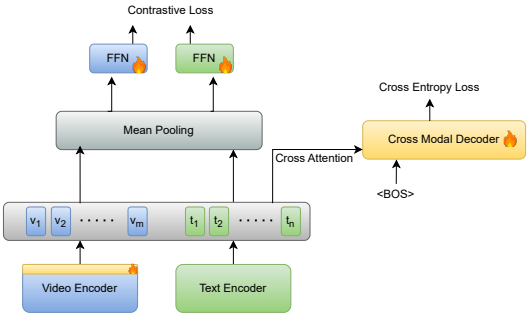


Figure 2: Standard VMT architecture. We make use of a video encoder and text encoder to process the source video and text. We mean pool them and put a contrastive loss along with standard cross-entropy loss. We only fire the decoder and FFNs required for contrastive loss.

Video Feature Extraction Given a video clip V_i of duration T_{total}, we uniformly sample frames at a fixed rate $\text{fps}_{\text{sample}}$, resulting in $T = \lfloor T_{\text{total}} \times \text{fps}_{\text{sample}} \rfloor$ frames $\{I_1, I_2, \ldots, I_T\}$, where $I_t \in \mathbb{R}^{H \times W \times C}$ and H, W, and C denote the height, width, and number of channels of each frame, respectively. A pre-trained vision encoder (VE) produces frame-level features, which are projected into the hidden space dimension d of the translation model via a feed-forward network, giving the final sequence of video features $X_v = \{z_1, z_2, \ldots, z_T\} \in \mathbb{R}^{T \times d}$, where $z_t = \text{FFN}(\text{VE}(I_t))$, and FFN is a feed-forward network.

Incorporating Video features We use adaptation techniques to train our model to attend to video features. Specifically, we use bottleneck adapters (Houlsby et al., 2019) in our decoder layers to learn the new modality. We also use LoRA (Low-rank Adaptation) (Hu et al., 2022) to fine-tune our model.

We use an additional symmetric Contrastive Loss for aligning the video and text embeddings. As seen in Figure 2, mean-pooled embeddings from the text and video embeddings are projected into a common space via linear heads, then normalised to facilitate cosine similarity comparisons. For a batch of size B, the per-pair similarity is computed, and the cross-entropy loss is applied

such that correct text-image pairs (i.e., those sharing the same batch index) are maximised. Let $\bar{h}_x$ and $\bar{h}_v$ denote the mean-pooled text and video embeddings, respectively. Then:

$$\ell_{ij} = \frac{\bar{h}_{x,i}^{\intercal} \bar{h}_{v,j}}{\tau}$$

$$\mathcal{L}_{\mathcal{C}} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp(\ell_{ii})}{\sum_{j=1}^{B} \exp(\ell_{ij})}$$

where τ is a temperature hyperparameter.

The final loss is a combination of the above contrastive loss ($\mathcal{L}_{\mathcal{C}}$) and the cross-entropy loss ($\mathcal{L}_{CE}$):

$$\mathcal{L}_{CE} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j} \log(p(y_{i,j} | y_{i,<j}, x_i, V_i))$$

$$\mathcal{L}_{total} = \mathcal{L}_{CE} + \lambda * \mathcal{L}_{C}$$

where λ is a hyperparameter to control the weight given to the contrastive loss term.

3.2 Fusion Techniques

Since the dimensionality of the video and text encoders may differ, we first project video embeddings into the hidden space d of the translation model. We then study two late-fusion strategies:

Fused Cross-Modal Attention In the first case, we have the cross-attention of the decoder attending to both text and video encoder outputs. Let the outputs of the encoders be X_t and X_v, of length L_t and L_v respectively. Then we pass

$$X_{concat} = [X_v; X_t] \in \mathbb{R}^{(L_t+L_v)\times d}$$

to the cross-attention matrices of the decoder. The key and value pairs will then be modified to:

$$K = X_{concat} \cdot W^k \in \mathbb{R}^{(L_t+L_v)\times d_k}$$

$$V = X_{concat} \cdot W^v \in \mathbb{R}^{(L_t+L_v)\times d_v}$$

The query matrix, Q, remains the same.

Video-Conditioned Self-Attention In the second case, we pass the video embeddings as a prefix to the decoder input. In this scenario, the self-attention mechanism attends to the video embeddings along with the target sequence tokens. We have the decoder sequence to be:

$$X_{\text{decoder}} = [X_v; X_{\text{target}}] \in \mathbb{R}^{(L_v+L_{\text{target}})\times d}$$

where X_{target} represents the target sequence embeddings (e.g., shifted target tokens during training or generated tokens during inference).

In this prefix-based approach, the self-attention mechanism operates on the concatenated sequence:

$$Q = X_{\text{decoder}} \cdot W^q \in \mathbb{R}^{(L_v+L_{\text{target}})\times d_k}$$

$$K = X_{\text{decoder}} \cdot W^k \in \mathbb{R}^{(L_v+L_{\text{target}})\times d_k}$$

$$V = X_{\text{decoder}} \cdot W^v \in \mathbb{R}^{(L_v+L_{\text{target}})\times d_v}$$

We use video encoders, which are mentioned in detail in Appendix A.4.

3.3 Synthetic Data Generation

Training of a VMT model requires the availability of triplet data, i.e., videos along with parallel subtitles. Therefore, extracting such triplets becomes crucial. However, for low-resource languages, the availability of such triplets is a challenge. Recent studies (Long et al., 2024) have shown that the instruction-following ability of LLMs and VLMs can help create synthetic data which can be reliably used to train better models across tasks. Therefore, we introduce a framework, as shown in Figure 3, which employs an MLLM to extract video features and a baseline text-only translation model to produce translations. These outputs are combined to generate visually grounded translations, which refine and improve upon the text-only baseline.

Formulation Formally, a VMT dataset can be represented as $\mathcal{D} = \{(V_i, x_i, y_i)\}_{i=1}^{N}$, where V_i denotes the raw video clip, x_i denotes the source language subtitle, and y_i denotes the target language translation. However, for low-resource languages, such triplets are often unavailable. Instead, we typically have access to partial data $\mathcal{D}_p = \{(V_i, x_i)\}_{i=1}^{N}$. To construct synthetic triplets, we first generate a baseline translation $y_i^{\text{base}} = f_\theta(x_i)$ using a text-only machine translation model f_θ. Next, we obtain the visual context in natural language by prompting a Multimodal Large Language Model (MLLM) g_ϕ with the video V_i and a fixed instruction prompt p_{desc}, giving $c_i = g_\phi(V_i, p_{\text{desc}})$, where c_i is a textual description of the scene, entities, objects, actions, and other visual cues relevant for disambiguation. Representing the visual context as text rather than as latent features lets it be passed directly to a text-only LLM, so no alignment between the representation spaces

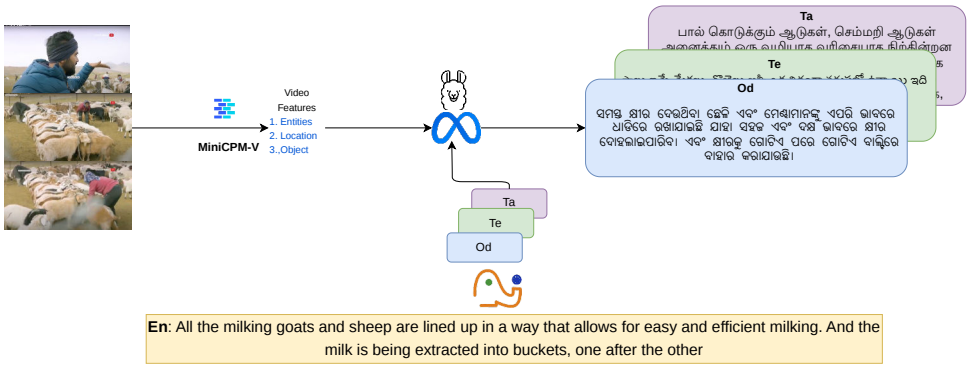


Figure 3: Synthetic data generation pipeline. An MLLM *(MiniCPM-V-2.6)* describes the video in natural language, and a text-only MT model *(IndicTrans2)* translates the English source. A text-only LLM *(Llama-3.1-8B)* then revises the baseline translation using the video description to produce a visually grounded translation.

of the two models is required. We then generate a visually grounded translation y_i^{syn} by prompting an LLM h_ψ with the source sentence, the baseline translation, and the description c_i, where $y_i^{\text{syn}} = h_\psi(x_i, y_i^{\text{base}}, c_i, p_{\text{ground}})$, where the grounding prompt p_{ground} instructs the LLM to revise y_i^{base} only where c_i resolves an ambiguity in x_i, and to otherwise leave the translation unchanged. The resulting synthetic triplet dataset is therefore $\mathcal{D}_{syn} = \{(V_i, x_i, y_i^{\text{syn}})\}_{i=1}^{N}$, which can be used to train a VMT model. Finally, the VMT model M_ω is trained to predict the target translation conditioned on both video and source text, $\hat{y}_i = M_\omega(V_i, x_i)$, and is optimised using the standard cross-entropy objective.

Since extending the VMT task to Indic languages requires triplet data in Indic languages, we make use of the synthetic data creation pipeline to create grounded translations for a low-resource Indic language. An example of this can be seen in Figure 4. To verify that the grounded translations genuinely incorporate visual information, we conduct a human verification study on 1,000 randomly sampled synthetic triplets: 82.1% are judged fully grounded, 13.4% partially grounded, and only 4.5% not grounded (Appendix A.8). We further show that the synthetic translations closely match the distribution of real subtitles at the semantic and syntactic level (Appendix A.10).

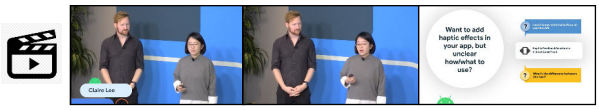
4 Experimental Setup

4.1 Data

We curate a high-quality English-Indic language parallel video-text dataset through a systematic extraction and filtering pipeline applied to the existing multimodal datasets. We filtered video IDs containing Indic language subtitles, and for each extracted video ID, we downloaded English and Indic subtitle tracks. We restricted ourselves to only content creators who uploaded subtitles to ensure high-quality subtitles. The subtitles are aligned using a timestamp-based alignment process that matches English and Indic language subtitle segments occurring within the same temporal windows. This alignment process ensured that paired text segments corresponded to identical video intervals. Using the obtained triplet data and previously available test sets (Kang et al., 2023), we randomly sample 1000 sentences and ask expert annotators in Hindi, Marathi and Odia to first identify sentences for which videos are necessary for accurate translations, i.e., cases where semantic ambiguities are present. After identifying such sentences, we ask the annotators to translate the filtered sentences into Hindi, Marathi and Odia. Such sentences form the Ambiguous set, whereas those which did not make it into the Ambiguous set form the Unambiguous set (as can be seen in Table 1). 42 samples were also dropped because of quality issues upon manual inspection. Detailed annotation instructions and the procedure of this process are provided in Appendix A.7.

We chose Odia (previously called Oriya) as the low-resource language for investigating the synthetic data pipeline to train for the VMT task in En-Od direction. This is because Odia (Od) was found to have no parallel subtitle data available in English, which qualifies it as a low-resource language for the VMT task.

For training data, we used the rest of the manually extracted triplet data, after removing those 1000 triplets mentioned above. To ensure data quality and semantic coherence, we apply two



Source
So you want to add haptic effects for your app, and you found these two ***constants***

Translation w/o grounding
ତେଣୁ ଆପଣ ନିଜ ଆପ୍ ପାଇଁ ହ୍ୟାପ୍ଟିକ୍ ପ୍ରଭାବ ଯୋଡ଼ିବାକୁ ଚାହୁଁଛନ୍ତି, ଏବଂ ଆପଣ ଏହି ଦୁଇଟି ***ସ୍ଥିରତାକୁ*** ପାଇଛନ୍ତି

Grounded Translation
ତେଣୁ ଆପଣ ଆପଣଙ୍କର ଆପ୍ଲିକେସନ୍ ପାଇଁ ହାପ୍ଟିକ୍ ପ୍ରଭାବ ଯୋଗ କରିବାକୁ ଚାହୁଁଛନ୍ତି, ଏବଂ ଆପଣ ଏହି ଦୁଇଟି ***ସ୍ଥାୟୀ ମୂଲ୍ୟ*** ପାଇଛନ୍ତି

Description
The video features an indoor scene with two individuals, a man and a woman, standing against a blue background. The setting appears to be a presentation or lecture environment, as suggested by the presence of a plant in the corner and the formal attire of the participants. 1. Scene context: This scene is set indoors, likely in a conference room or auditorium, indicated by the professional backdrop and lighting. The time of day cannot be determined .....might submit queries about using specific APIs related to haptic feedback in app development. 7. Ambiguity resolution: The text bubbles provide context for the discussion topic, clarifying that it revolves around adding haptic effects in ***apps*** and understanding how to use certain ***APIs*** effectively.

Grounded Improvement
Changed from "ସ୍ଥିରତା" (sthirata - stability/steadiness) to ***"ସ୍ଥାୟୀ ମୂଲ୍ୟ"*** (sthayi mulya - fixed/constant value).

Figure 4: Evidence of grounding in the synthetic data pipeline. The English source sentence contains the ambiguous word *constants*, which can mean either stability or the concept of fixed values in programming languages. Without grounding, the text-only model renders it as ସ୍ଥିରତା (*sthiratā*, stability), giving the back-translation *So you want to add haptic effects for your own app, and you have found these two stabilities*. MiniCPM-V-2.6 then generates a description of the video, which shows a presentation on mobile development and explicitly mentions APIs and app development. Using this description, the LLM revises the translation to ସ୍ଥାୟୀ ମୂଲ୍ୟ (*sthāyī mūlya*, constant value), giving the back-translation *So you want to add haptic effects for your application, and you have found these two constant values*. The visual context thus allows the model to treat constants as a programming term rather than a general noun for stability.

Set Type	En–Hi	En–Mr	En–Od
Ambiguous	158	158	158
Non-Ambiguous	800	800	800
Total	958	958	958

Table 1: Test Data Statistics for Ambiguous and Non-Ambiguous Sets

quality-control steps. First, we use Language-agnostic BERT Sentence Embeddings (LaBSE) (Feng et al., 2022) to compute cosine similarity between the embeddings of aligned subtitle pairs and retain only those exceeding a threshold (0.9 for Hindi and 0.8 for Marathi), removing mistranslations and misalignments. Second, we apply a denoising step that includes removing pairs with length-ratio outliers, near-duplicate segments, and non-target script characters. This results in 46K sentence pairs for Hindi and 35K for Marathi, with corresponding video segments for each pair. For Odia, we extract triplets using the synthetic data generation pipeline (as seen in Section 3.3), filter and denoise it down to 25K sentence pairs for training. We use a threshold of 0.8 for this case. For the synthetic En–Od data, we also show in Table 14 that both steps substantially improve downstream VMT performance.

4.2 Baselines

Our multimodal translation model is built on top of two strong pre-trained text-only systems: **IndicTrans2** (Gala et al., 2023) for English–Indic language directions, and the **Helsinki-NLP Opus** (Tiedemann and Thottingal, 2020) models for the English–Chinese pair. These models provide robust linguistic representations and strong baseline translation quality. We extend them by introducing video modality through LoRA-based adaptation combined with contrastive training. In our results, we refer to the use of Fused Cross-Modal Attention (Section 3.2) simply as LoRA.

For VMT of low-resource language in the En-Od direction, we use IndicTrans2 (Gala et al., 2023) to first obtain the machine-translated output. We use MiniCPM-V (Yao et al., 2024) to generate video descriptions, which aid in identifying features that enhance the machine-translated outputs (Appendix A.3). Finally, we use Llama-3.1-8B (Grattafiori et al., 2024) to combine video descriptions and machine-translated outputs to get the visually grounded translations for Odia. We use the grounding prompt described in Appendix A.9.

4.3 Evaluation

For evaluation, we compare against the original text-only models in both zero-shot and fine-tuned settings to measure the gains achieved by incorporating video features.

Evaluation Metrics We use the BLEU (Bilingual Evaluation Understudy) score (Papineni et al., 2002) and the COMET Score (Rei et al., 2020) for

evaluating our MT outputs. We also use an additional chrF (Popović, 2015) score for Chinese MT outputs. We also perform human evaluation as given in A.8.

Evaluation Sets For the English-Hindi translation task, we created two evaluation sets. A test set that is a random split from our full dataset $\mathcal{D}$, the **In-domain**, and a heldout set, with the set forming a heldout cluster when k-means is applied to the video features for each data point. This set is denoted by **Out-of-domain**. For English-Marathi, English-Odia and English-Chinese directions, the test set used is a random split from our filtered dataset. We also test on our created ambiguous test set for En–Hi, En-Mr and En-Od. The statistics of the ambiguous set are mentioned in Appendix A.11.

5 Results and Analysis

We evaluate our proposed VMT systems across English–Hindi, English–Marathi, and English-Odia language pairs. The results in Tables 2 and 3 highlight the impact of introducing video modality, different vision encoders, and contrastive alignment losses.

Effect of Fine-tuning Text-only Models IndicTrans2, when used in a zero-shot setting, provides a reasonable baseline (BLEU 41.94 on the En–Hi In-domain test set). Fine-tuning with LoRA on parallel text alone leads to large improvements (BLEU +15.38), underscoring the importance of domain adaptation even without additional modalities.

Contribution of Video Modality Adding video features further improves translation quality. For English–Hindi, our VMT model with LoRA and CLIP embeddings achieves the best performance (BLEU 59.07 and COMET 90.26 on the In-domain test set), outperforming the text-only fine-tuned baseline by +1.75 BLEU. The gains are more pronounced on the Out-of-domain test set, where video-conditioned models outperform the text-only fine-tuned model by up to +1.51 BLEU. This suggests that the video modality provides stronger disambiguation when the test set differs more from training data distributions. For English–Marathi and English-Odia, the addition of video features similarly improves translation with increased BLEU and COMET scores. Paired bootstrap resampling confirms that the improvements are significant (Appendix A.12 and Table 20).

Impact of Contrastive Loss Incorporating a contrastive loss (CTR) to align video and text representations yields mixed but generally positive effects. For English–Hindi, contrastive loss with CLIP embeddings achieves the highest scores on the Out-of-domain test set (BLEU 55.01, COMET 85.33), establishing that alignment objectives help models exploit visual context more effectively in challenging test scenarios. However, on some subsets (e.g., English–Marathi with ViT and English-Odia), the contrastive loss slightly reduces BLEU, suggesting sensitivity to the choice of video encoder and dataset scale.

English-Chinese Translation We also evaluate our approach on the English–Chinese language pair. As can be seen from Table 7, incorporating the video modality consistently improves performance across all evaluation metrics, demonstrating that our method generalises beyond Indic languages and is effective across diverse language pairs.

Ambiguity Resolution We further evaluate on a curated ambiguous subset, where translations require contextual grounding. As shown in Table 4, the addition of video features combined with contrastive loss achieves the best results (BLEU 15.14, COMET 83.61), matching or surpassing the zero-shot baseline and outperforming the fine-tuned text-only model. For Odia, we see a marginal increase in performance using videos. However, for Marathi, performance drops. Since the training set for Marathi is general domain, we can consider ambiguous Marathi examples to be both out-of-domain and ambiguous, which makes it tougher for the model to discern. These findings indicate that video guidance is beneficial in scenarios where textual information alone is insufficient, but it is possible that it may not show up in automatic metrics as the edits are very small. Table 6 and Appendix A.1 show representative examples where video context helps the model select the correct native term where the text-only model falls back on transliteration and adds visual detail absent from the source, such as translating *spoons* as बड़े चम्मच (tablespoons) when the video shows a large spoon. In this case, an improvement in BLEU is penalised since the reference lacks the grounded version. Error cases show that the model can also over-correct, replacing an appropriate loanword with a native

Model	Video Features	In-domain BLEU	In-domain COMET	Out-of-domain BLEU	Out-of-domain COMET
IndicTrans2 (zero-shot)	-	41.94	88.86	44.54	84.51
IndicTrans2 FT (LoRA)	-	57.32	90.12	53.43	85.33
Ours (Video-Conditioned Self-Attention) + LoRA	ViT	48.34	87.36	50.70	83.71
	CLIP	48.13	87.02	51.16	83.52
Ours (LoRA)	ViT	59.02	89.77	53.93	85.07
	CLIP	**59.07**	90.26	54.94	85.18
Ours (LoRA) + CTR	ViT	58.31	90.29	54.15	**85.45**
	CLIP	58.96	**90.31**	**55.01**	85.33

Table 2: Model performance comparison for En-Hi with different video features

Model	Video Features	En-Mr BLEU	En-Mr COMET	En-Od BLEU	En-Od COMET
IndicTrans2 FT (LoRA)	-	22.32	62.29	19.41	64.96
Ours (LoRA)	ViT	22.85	62.16	**20.77**	66.89
	CLIP	**23.18**	62.4	19.61	**67.81**
Ours (LoRA) + CTR	ViT	22.85	62.51	19.34	67.55
	CLIP	23.08	**62.54**	19.69	67.28

Table 3: Model performance on unambiguous dataset for En–Mr and En–Od

Model	BLEU	COMET
IndicTrans2 (zero-shot)	15.13	83.61
IndicTrans2 FT (LoRA)	14.00	82.20
Ours (LoRA)	14.27	82.41
Ours (LoRA) + CTR	**15.14**	**83.61**

Table 4: Model performance on ambiguous dataset for En-Hi with ViT features

Model	En–Mr BLEU	En–Mr COMET	En–Od BLEU	En–Od COMET
IndicTrans2 FT	**10.00**	**61.36**	7.44	71.24
Ours	9.98	61.17	**7.67**	**71.25**
Ours + CTR	9.69	61.08	**7.67**	**71.25**

Table 5: Model performance on the ambiguous test sets for En–Mr and En–Od with ViT features. All models are fine-tuned with LoRA.

Source	Text-only	VMT	Outcome
bead	बीड (translit.)	मनका (bead)	Resolved
5 spoons	5 चम्मच (5 spoons)	5 बड़े चम्मच (5 tablespoons)	Grounded (video shows a large spoon)
round	राउंड (round)	दौर (phase, bout)	Wrong sense
Which one is that?	वह कौन सा है? (which one is that?)	जो वह है? (which that is?)	Meaning lost

Table 6: Qualitative En–Hi examples (glosses in parentheses). VMT can recover native terms and add visually grounded detail, but can also select an inappropriate sense or distort meaning (last two rows). Full sentences, video frames, glosses, and back-translations are in Table 11 (Appendix A.1).

term of the wrong sense, or distort the source meaning.

Human Evaluation Since video typically induces small, targeted edits that automatic metrics only partly capture, we also conduct human evaluation (Appendix A.8). VMT outputs are rated highly fluent (4.53–4.79) and adequate (3.60–4.76), with video-consistency scores of 3.80–3.97 across languages (Table 15). In pairwise comparison on the ambiguous sets, annotators prefer VMT over the text-only model in 139 cases versus 89 (pooled across languages) in Table 17.

6 Additional Experiments

We conduct ablation studies to understand how different training strategies affect the extent to which the model attends to video features. Results on the unambiguous set are reported in Table 9 and 10.

Two-stage Training In this setting, we first train the model with the masked source text such that the decoder is forced to rely on video features to gen-

Model	BLEU	COMET	chrF
Opus (zero-shot)	32.01	71.01	30.25
Opus FT (LoRA)	43.82	80.5	38.98
Ours (LoRA)	**44.98**	82.11	**40.69**
Ours (LoRA) + CTR	44.31	**82.21**	40.33

Table 7: Model performance for En-Zh with CLIP features

Model	En–Hi	En–Mr	En–Od
GPT-5.2	12.02	7.10	6.68
GPT-4o-mini	11.03	9.49	6.03
MiniCPM-V-2.6	5.02	0.40	0.70

Table 8: BLEU scores for English to Hindi, Marathi, and Odia translation.

erate translations. In the second stage, we introduce both full source text and video features to the model. This is designed to encourage the model to initially ground its translations in video, before allowing it to balance both modalities (Gupta et al., 2023). While this approach improves the model's reliance on video, it achieves slightly lower scores (BLEU 50.18, COMET 84.30) than our standard LoRA fine-tuning. This suggests that although the curriculum can bias the model toward visual grounding, it may limit the overall fluency that comes from strong textual alignment.

Masked Source Text Here, we randomly mask a small percentage of tokens in the source sentence during training, forcing the decoder to recover missing information from video features. This strategy is less aggressive than the two-stage curriculum and allows for smoother integration of textual and visual information. Compared to two-stage training, masked source text yields stronger results (BLEU 50.99, COMET 84.77), indicating that partial masking encourages the model to exploit video context without overly constraining its ability to use the source text.

Comparison with Standard LoRA Our default LoRA-based multimodal fine-tuning still achieves the best overall performance (BLEU 53.93, COMET 85.07). Nonetheless, both ablations confirm that encouraging the model to depend more on video features can enhance multimodal grounding, even if at the cost of slight reductions in overall translation quality.

Adapters vs LoRA Finally, comparing adaptation strategies (Table 10), we find that LoRA adaptation significantly outperforms bottleneck adapters (BLEU 53.93 vs. 39.28), suggesting that low-rank adaptation is better suited for integrating video modality in our setting. We experimented with adapters of different sizes and reported the results for the best-performing one.

Model	BLEU	COMET
Ours (LoRA)	**53.93**	**85.07**
Two-stage Training	50.18	84.30
Masked source text	50.99	84.77

Table 9: Comparison of different training strategies

Model	BLEU	COMET
Ours (Adapters)	39.28	80.12
Ours (LoRA)	**53.93**	**85.07**

Table 10: Comparison of different Adaptation techniques

MLLMs Performance We carry out zero-shot experiments on MLLMs using all video frames as input (Table 19 and Table 8) and zero/few-shot experiments using text descriptions of the video (Table 21). We observe that the zero-shot performance of GPT-5.2 is better than both GPT-4o-mini and MiniCPM-V-2.6. We also observe that increasing the number of examples helps in better translation. Lastly, we experiment by giving only select frames (Table 19) to the MLLM. We observe that using one frame yields better results than using 3 frames while prompting MLLMs. This suggests that the MLLM gets redundant information from irrelevant frames and thus requires the most precise and relevant frame to give the best translation.

7 Conclusion

In this paper, we introduce a VMT benchmark with annotated evaluation sets for three Indic languages: English–Hindi, English–Marathi, and English–Odia. We explore efficient training strategies using pre-trained models, avoiding training from scratch. Our results show that these methods generalise to other language pairs, such as English–Chinese. We also present a synthetic data generation framework that enables visually grounded translations for low-resource languages like Odia. We see that the VMT task finds it difficult to upstage the baselines by a large margin but performs small edits to improve the translation quality.

Limitations

Our experiments are constrained by the size of the available dataset. While larger datasets could potentially improve model performance, collecting and curating additional high-quality parallel data is challenging. Furthermore, generating synthetic data to augment the dataset requires substantial computational resources, which may limit the scalability of such approaches. Also, many datasets and their sources of data for VMT may require paid licenses, which is an added challenge.

Acknowledgements

We thank the anonymous reviewers for their constructive feedback, which helped improve this work. We also thank Prof. Preethi Jyothi at IIT Bombay for her valuable guidance and suggestions throughout this work. Finally, we thank the CFILT Lab at IIT Bombay for funding the annotation and human evaluation efforts.

Ethics Statement

Our dataset consists of video IDs and parallel subtitle data created from openly available sources. We do not redistribute the videos or the creator-uploaded subtitles, which remain subject to the terms of their original sources. All training data was collected from publicly available data, and the annotated evaluation sets will be released under a CC BY 4.0 license. We will put in place proper guidelines so that the data is only used for research purposes. All annotators were compensated at market rates. We used AI assistants only for grammar correction and correcting stylistic issues during writing.

References

Iacer Calixto, Miguel Rios, and Wilker Aziz. 2019. Latent variable model for multi-modal translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6392–6405, Florence, Italy. Association for Computational Linguistics.

Zhuo Chen, Fei Yin, Qing Yang, and Cheng-Lin Liu. 2023. Cross-lingual text image recognition via multi-hierarchy cross-modal mimic. *IEEE Transactions on Multimedia*, 25:4830–4841.

Xuxin Cheng, Zhihong Zhu, Yaowei Li, Hongxiang Li, and Yuexian Zou. 2023. Das-cl: Towards multimodal machine translation via dual-level asymmetric contrastive learning. pages 337–347.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

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

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

Qingkai Fang and Yang Feng. 2022. Neural machine translation with phrase-level universal visual representations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5687–5698, Dublin, Ireland. Association for Computational Linguistics.

Hao Fei, Qian Liu, Meishan Zhang, Min Zhang, and Tat-Seng Chua. 2023. Scene graph as pivoting: Inference-time image-free unsupervised multimodal machine translation with visual scene hallucination. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5980–5994, Toronto, Canada. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Chengpeng Fu, Xiaocheng Feng, Yichong Huang, Wenshuai Huo, Hui Wang, Bing Qin, and Ting Liu. 2023. Enabling unsupervised neural machine translation with word-level visual representations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12608–12618, Singapore. Association for Computational Linguistics.

Matthieu Futeral, Cordelia Schmid, Benoît Sagot, and Rachel Bawden. 2024. Towards zero-shot multimodal machine translation. *arXiv preprint arXiv:2407.13579*.

Jay Gala, Pranjal A Chitale, A K Raghavan, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar M, Janki Atul Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. Indictrans2: Towards high-quality and accessible machine translation models

for all 22 scheduled Indian languages. *Transactions on Machine Learning Research*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Weiqi Gu, Haiyue Song, Chenhui Chu, and Sadao Kurohashi. 2021. Video-guided machine translation with spatial hierarchical attention network. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 87–92, Online. Association for Computational Linguistics.

Boyu Guan, Yining Zhang, Yang Zhao, and Chengqing Zong. 2025. TriFine: A large-scale dataset of vision-audio-subtitle for tri-modal machine translation and benchmark with fine-grained annotated tags. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8215–8231, Abu Dhabi, UAE. Association for Computational Linguistics.

Wenyu Guo, Qingkai Fang, Dong Yu, and Yang Feng. 2023. Bridging the gap between synthetic and authentic images for multimodal machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2863–2874, Singapore. Association for Computational Linguistics.

Devaansh Gupta, Siddhant Kharbanda, Jiawei Zhou, Wanhua Li, Hanspeter Pfister, and Donglai Wei. 2023. CLIPTrans: Transferring visual knowledge with pre-trained models for multimodal machine translation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2863–2874.

Tosho Hirasawa, Zhishen Yang, Mamoru Komachi, and Naoaki Okazaki. 2020. Keyframe segmentation and positional encoding for video-guided machine translation challenge 2020. *ArXiv*, abs/2006.12799.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning*.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Baijun Ji, Tong Zhang, Yicheng Zou, Bojie Hu, and Si Shen. 2022. Increasing visual awareness in multimodal neural machine translation from an information theoretic perspective. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6755–6764, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Liyan Kang, Luyang Huang, Ningxin Peng, Peihao Zhu, Zewei Sun, Shanbo Cheng, Mingxuan Wang, Degen Huang, and Jinsong Su. 2023. BigVideo: A large-scale video subtitle translation dataset for multimodal machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8456–8473, Toronto, Canada. Association for Computational Linguistics.

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

Mingjie Li, Po-Yao Huang, Xiaojun Chang, Junjie Hu, Yi Yang, and Alex Hauptmann. 2023a. Video pivoting unsupervised multi-modal machine translation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3918–3932.

Yi Li, Rameswar Panda, Yoon Kim, Chun-Fu Chen, Rogério Schmidt Feris, David D. Cox, and Nuno Vasconcelos. 2022a. Valhalla: Visual hallucination for machine translation. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5206–5216.

Yihang Li, Shuichiro Shimizu, Chenhui Chu, Sadao Kurohashi, and Wei Li. 2023b. Video-helpful multimodal machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4281–4299, Singapore. Association for Computational Linguistics.

Yihang Li, Shuichiro Shimizu, Weiqi Gu, Chenhui Chu, and Sadao Kurohashi. 2022b. VISA: An ambiguous subtitles dataset for visual scene-aware machine translation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6735–6743, Marseille, France. European Language Resources Association.

Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. On LLMs-driven synthetic data generation, curation, and evaluation: A survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11065–11082, Bangkok, Thailand. Association for Computational Linguistics.

Jinze Lv, Jian Chen, Zi Long, Xianghua Fu, and Yin Chen. 2025. TopicVD: A topic-based dataset of video-guided multimodal machine translation for documentaries. *Preprint*, arXiv:2505.05714.

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-level neural

machine translation with hierarchical attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2947–2954, Brussels, Belgium. Association for Computational Linguistics.

Tetsuro Nishihara, Akihiro Tamura, Takashi Ninomiya, Yutaro Omote, and Hideki Nakayama. 2020. Supervised visual attention for multimodal neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4304–4314, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Shantipriya Parida and Ondřej Bojar. 2025. Findings of WAT2025 English-to-Indic multimodal translation task. In *Proceedings of the Twelfth Workshop on Asian Translation (WAT 2025)*, pages 103–108, Mumbai, India. Association for Computational Linguistics.

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

Maja Popović. 2015. chrF: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

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

Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. 2018. How2: A large-scale dataset for multimodal language understanding. *ArXiv*, abs/1811.00347.

Loitongbam Sanayai Meetei, Thoudam Doren Singh, and Sivaji Bandyopadhyay. 2019. WAT2019: English-Hindi translation on Hindi visual genome dataset. In *Proceedings of the 6th Workshop on Asian Translation*, pages 181–188, Hong Kong, China. Association for Computational Linguistics.

Julia Sato, Helena Caseli, and Lucia Specia. 2023. Choosing what to mask: More informed masking for multimodal machine translation. In *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.

Ammon Shurtz, Lawry Sorenson, and Stephen D. Richardson. 2024. The effects of pretraining in video-guided machine translation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15888–15898, Torino, Italia. ELRA and ICCL.

Dídac Surís, Dave Epstein, and Carl Vondrick. 2022. Globetrotter: Connecting languages by connecting images. *Preprint*, arXiv:2012.04631.

ZhenHao Tang, XiaoBing Zhang, Zi Long, and XiangHua Fu. 2022. Multimodal neural machine translation with search engine based image retrieval. In *Proceedings of the 9th Workshop on Asian Translation*, pages 89–98, Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Jörg Tiedemann and Santhosh Thottingal. 2020. Opus-mt–building open translation services for the world. In *Annual Conference of the European Association for Machine Translation*, pages 479–480. European Association for Machine Translation.

Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. 2019. VaTeX: A large-scale, high-quality multilingual dataset for video-and-language research. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4580–4590.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, and 4 others. 2024. MiniCPM-V: A GPT-4V level MLLM on your phone. *Preprint*, arXiv:2408.01800.

Kong Yawei and Kai Fan. 2021. Probing multi-modal machine translation with pre-trained language model. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3689–3699, Online. Association for Computational Linguistics.

Ryoya Yuasa, Akihiro Tamura, Tomoyuki Kajiwara, Takashi Ninomiya, and Tsuneo Kato. 2023. Multimodal neural machine translation using synthetic images transformed by latent diffusion model. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 76–82, Toronto, Canada. Association for Computational Linguistics.

Zhuosheng Zhang, Kehai Chen, Rui Wang, Masao Utiyama, Eiichiro Sumita, Z. Li, and Hai Zhao. 2020.

Neural machine translation with universal visual representation. In *International Conference on Learning Representations*.

Yaoming Zhu, Zewei Sun, Shanbo Cheng, Luyang Huang, Liwei Wu, and Mingxuan Wang. 2023. Beyond triplet: Leveraging the most data for multimodal machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2679–2697, Toronto, Canada.

A Appendix

A.1 Qualitative Examples

Table 11 presents qualitative En–Hi examples comparing text-only and VMT outputs, along with three frames from the corresponding video segment. Each example includes the English source, the reference, a word-by-word Hindi gloss (HG), and a back-translation (BT). In (a), the text-only model transliterates *bead* as *bīḍ*, while VMT, seeing the beadwork in the video, uses the native Hindi term *mankā*. In (b), VMT renders *spoons* as *baṛe chammach* (tablespoons), matching the large spoon shown in the video. This grounded detail is penalised by BLEU, since the reference does not contain it. The two error cases illustrate the limitations of VMT. In (c), the model over-corrects when it replaces the conventional loanword *rāuṇḍ* with the native term *daur* ("phase, bout"), which is inappropriate in a beadwork context. In (d), the model turns the question *vah kaun sā hai* (which one is that?) into a relative clause, losing the meaning of the source.

A.2 chrF++ Scores

We report the chrF++ scores in Tables 12 and 13 to complement the results in Section 5. We also compare our results against baselines such as VaTex in Table 22 and Table 23.

A.3 Description of Prompts

This section presents the prompts used to obtain video descriptions from *MiniCPM-V*.

MiniCPM-V-2.6

```
You are given a video. Provide a
detailed, structured description that
highlights visual cues useful for
Video-Guided Machine Translation.
Focus on the following aspects:
1. Scene context: Specify indoor/outdoor,
time of day, lighting, and background
environment.
2. Objects: List all visible objects
with size, shape, colour, material,
and relative positions. 3. Entities:
Describe people, animals, or other beings
- include appearance, clothing, facial
expressions, posture, and interactions.
4. Actions and dynamics: Explain
what each entity is doing, gestures,
movements, and how they interact with
objects or others.
5. Temporal flow: Summarise the sequence
of events across frames rather than one
static moment.
6. Communication cues: If someone is
speaking, describe who appears to be
speaking, their body language, and
the reactions of others. 7. Ambiguity
resolution: Add details that can
disambiguate meaning (e.g., homonyms,
cultural context, visible numbers, signs,
or symbols).

Important instructions: - Be factual and
precise; avoid guessing or inventing
details. - Use clear and concise
sentences. - Write in a way that helps
disambiguate translations by grounding
abstract terms with visible details.
```

A.4 Video Encoders

We consider two different vision encoders for our experiments.

1. ViT (Dosovitskiy et al., 2021) is a Vision Transformer model that structures images into patches and applies transformer-based self-attention to capture complex spatial relationships in visual data. ViT models are typically trained using classification losses such as cross-entropy loss and masked image modelling losses. [3]

2. CLIP (Contrastive Language-Image Pre-training) (Radford et al., 2021) trains a dual-encoder (image + text) model by maximising the cosine similarity between embeddings of paired images and texts, while minimising similarity for incorrect pairs in the same batch. The CLIP loss is a symmetric contrastive loss that ensures that image and text representations are close to each other. [4]

A.5 Hyperparameters

We set the LoRA rank to 16 after experimenting with ranks of 8, 16, and 32. While higher ranks increased the number of trainable parameters, they did not lead to consistent improvements in translation quality, making rank 16 a good balance between efficiency and performance. For

[3]The ViT model used:timm/vit_base_patch16_224.mae
[4]The CLIP model used: openai/clip-vit-base-patch32

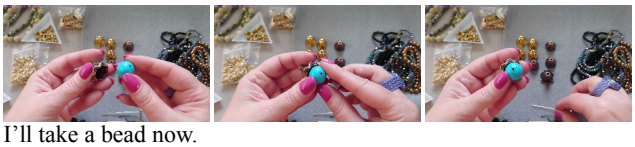
	Sentence	Comment
(a) Transliteration vs. translation		
Frames		Video shows beadwork; the text-only output transliterates **bead** (*bīḍ*), while VMT uses the Hindi word *mankā* (bead).
Source	I'll take a bead now.	
Reference	मैं अब एक मनका लूँगा। HG: I now one bead take-will	
Text-only	मैं अब एक **बीड** लूँगा। HG: I now one **"bīḍ"** take-will	
VMT	मैं अब एक **मनका** लूँगा। BT: I will take a bead now.	
(b) Grounded specification		
Frames		The video shows a large spoon; VMT adds *baṛe* (big), rendering spoons as **tablespoons**. This grounded detail is penalised by BLEU, since the reference lacks it.
Source	Add 5 spoons of hot water.	
Reference	गर्म पानी के 5 चम्मच जोड़ें। HG: hot water of 5 spoons add	
Text-only	गर्म पानी के 5 चम्मच जोड़ें। (identical to reference)	
VMT	5 **बड़े** चम्मच गर्म पानी डालें। HG: 5 **big** spoons hot water pour BT: Pour in 5 tablespoons of hot water.	
(c) Error: lexical ambiguity		
Frames		Video shows an indoor beadwork activity, where *rāuṇḍ* is the conventional term. VMT instead selects *daur* (phase, bout), an inappropriate sense.
Source	Now, I'm doing a round again.	
Reference	अब, मैं फिर से एक राउंड कर रहा हूँ। HG: now, I again one round do-PROG am	
Text-only	अब, मैं फिर से एक **राउंड** कर रहा हूँ। (identical to reference)	
VMT	अब, मैं फिर से एक **दौर** कर रहा हूँ। HG: now, I again one **phase** do-PROG am BT: I am doing another phase now.	
(d) Error: meaning distortion		
Frames		VMT replaces the question *vah kaun sā hai* (*which one is that?*) with the relative clause *jo vah hai*, so the question is lost.
Source	Well, there is nothing wrong here; this is good. Which one is that?	
Reference	वैसे यहाँ कुछ भी गलत नहीं है, यह अच्छा है। वह कौन सा है? HG: well here anything wrong not is, this good is. that which one is?	
Text-only	खैर यहाँ कुछ भी गलत नहीं है, यह अच्छा है। वह कौन सा है? BT: Well, nothing is wrong here, this is good. Which one is that?	
VMT	खैर यहाँ कुछ भी गलत नहीं है, यह अच्छा है **जो वह है**? HG: well here anything wrong not is, this good is **which that is**? BT: Well, nothing is wrong here, this is good, which it is?	

Table 11: Qualitative En–Hi examples comparing text-only and VMT outputs: Two cases where video helps (a, b) and two error cases (c, d). Three frames from the corresponding video segment are shown for each example. HG = word-by-word gloss; BT = back-translation into English. Words that differ between systems are in bold.

Model	In-domain	Out-of-domain
IndicTrans2 (zero-shot)	73.99	72.68
IndicTrans2 FT (LoRA)	77.57	74.04
Ours (LoRA)	79.07	75.14
Ours (LoRA) + CTR	78.82	75.24

Table 12: chrF scores on in-domain and out-of-domain datasets for English–Hindi translation.

Model	chrF
IndicTrans2 FT (LoRA)	53.33
Ours (LoRA)	53.90
Ours (LoRA) + CTR	54.07

Table 13: chrF scores on English–Marathi translation

the contrastive loss, we found that assigning a weight of 1.0 for Hindi and 0.5 for Marathi gave the best trade-off between leveraging video features and maintaining fluency. The temperature for the contrastive objective was tuned among $0.001, 0.005, 0.01$, with 0.005 yielding the most stable training and highest scores. These hyperparameters reflect the best-performing configuration identified through empirical exploration. We also perform significance testing as shown in Table 20.

A.6 Ambiguity types in VMT

In VMT, ambiguity arises at multiple levels and can be effectively resolved through visual grounding. Beyond *lexical ambiguity* (e.g., bat, bank, shoot) and *coreference ambiguity* (resolving pronouns via visual cues), VMT must also handle *terminological ambiguity*, where words like cell, crane, or mouse take on domain-specific meanings that are disambiguated by the scene. Additionally, *action ambiguity* (e.g., *run* as physical motion vs. operating a system), *spatial ambiguity* (e.g., *left* relative to camera or subject), and *event-level ambiguity* (unclear interactions or intent in multi-entity settings) are more difficult cases of VMT.

A.7 Annotation Guidelines

For data annotation, two language experts were employed for each language: one for primary annotation and one for validation during translation. The annotators were evenly split by gender, with one male and one female per language. All annotators were compensated according to market rates of the region, and the average age of the annotators was 40 years. The annotators have consented to this data being publicly available.

Data Setting	Encoder	IT2 FT	Ours (LoRA)	+ CTR
Original	ViT	15.89	17.81	17.83
	CLIP		17.69	17.54
+ LaBSE filtering	ViT	16.79	18.52	18.63
	CLIP		18.41	18.57
+ LaBSE filtering	ViT	**19.41**	**20.77**	19.34
+ Denoising	CLIP		19.61	**19.69**

Table 14: Effect of LaBSE filtering and denoising on the synthetic En–Od training data (BLEU). IT2 FT or IndicTrans2 FT (LoRA) denotes the text-only IndicTrans2 model fine-tuned with LoRA, which uses no video features. Each preprocessing step improves all systems, and the full pipeline gives the best results.

Annotation Guideline for VMT

1. Task Overview

You are required to translate English sentences (subtitles of a given video) into Hindi, Marathi, or Odia. Unlike standard translation, this task is **video-guided**, meaning that the translation should not only convey the meaning of the English sentence, but also integrate additional information from visual or contextual cues in the video. Your work will populate the following columns in the annotation sheet:

- **VideoIds**: The link to the corresponding video.
- **English Source**: The English subtitle (already provided).
- **Visually Grounded Translation in Hindi/Marathi/Odia**: Your final translation in the target language, guided by both text and video.
- **Extra Information Added**: Any additional details incorporated from the video (not explicit in the English subtitle).
- **Notes (if any)**: Observations, doubts, or clarifications.

2. Principles of Translation

Accuracy

- Preserve the original meaning of the English source sentence.

- Ensure no factual distortions are introduced.

Video-Grounded Context Use the video to clarify ambiguities in the English source. If the video reveals who is speaking, what is happening, or other context, integrate that naturally in translation.

English Source: Put it there.
Video: The speaker points at the table.
Hindi Translation: इसे मेज़ पर रखो।
(Put it on the table.)

Naturalness & Fluency Translations must sound natural to a native speaker; avoid overly literal translations.

English: She broke down in tears.
Hindi (literal): वह आँसुओं में टूट गई।
(She broke down in tears.)
Hindi (better): वह रोने लगी। (She started crying.)

Register and Tone Match the tone of the dialogue (formal, informal, humorous).

Handling Cultural Differences Adapt idioms or humour when needed; use culturally appropriate paraphrases.

3. Using Video Cues Effectively
When watching the video, pay attention to the following cues:
Visual Objects & Actions

English: Give me that.
Video: Speaker pointing to a red book.
Hindi: मुझे वह लाल किताब दो। (Give me that red book.)

Speaker Identity & Gender

English: I am tired.
Video: A woman speaking.
Hindi: मैं थक **गई** हूँ। (I am tired. The feminine गई(gaī) marks a female speaker.)

Setting / Location

English: It's hot in here.
Video: People are in a kitchen.
Marathi: स्वयंपाकघरात खूप गरम आहे।
(It is very hot in the kitchen.)

Emotional Expressions & Tone

English: Oh, that's just great! (said sarcastically)
Odia: ବାହ! ବହୁତ ଭଲ! (sarcastic tone) (Wow! Very good!)

4. Extra Information Column
Use this column to record extra visual details added to your translation.

- Added **table** because the video shows object placement.
- Clarified the gender of the speaker.
- Added **red book** instead of generic **that**.

5. Notes Column

- Doubts or alternative translations.
- Unclear video or audio segments.
- Annotation issues (e.g. subtitle mismatch).

6. Examples Across Languages

Example 1
English Source: He's coming now.
Video: A young boy walking toward the speaker.
Hindi Translation: वह लड़का अब आ रहा है।
Extra Info Added: Identified subject as "boy."
(The boy is coming now.)
Notes: None.

Example 2
English Source: Take it.
Video: A woman handing over a pen.
Marathi Translation: हे पेन घे।
Extra Info Added: Object identified as **pen**.
(Take this pen.)
Notes: None.

Example 3
English Source: I can't do this anymore.
Video: Speaker crying while speaking.
Odia Translation: ମୁଁ ଆଉ ଏହା କରିପାରିବି ନାହିଁ, ମୋର ହିମ୍ମତ ଭଙ୍ଗିଗଲା।
(I can't do this anymore, my heart is broken) Extra Info Added: Emotional tone included.
Notes: Video context indicates emotional breakdown.

A.8 Human Evaluation

Because VMT annotations require high cognitive load, creating a test set is an inherently expensive task. We therefore created a relatively small test set as given in Table 1 and conducted thorough human evaluation using three criteria: *fluency*, *adequacy*, and *video consistency*, each rated from 1 to 5 on a Likert scale by three native annotators. As shown in Table 15, the system achieves high fluency and adequacy, indicating grammatically well-formed translations that largely preserve source meaning. The video consistency score suggests that the model incorporates visual context in many cases, though challenges remain in grounding fine-grained visual details. We also perform pairwise human evaluation to establish that humans prefer the VMT translations more than the text-only translations (Table 17).

Human Evaluation Guidelines

1. Purpose of the Evaluation This evaluation assesses the quality of English→Hindi video-guided machine translation systems, with a particular focus on visual grounding and ambiguity resolution. Annotators evaluate translations along four clearly separated dimensions:

- Fluency (language quality)
- Adequacy (meaning preservation)
- Video Consistency / Grounding (alignment with visual content)
- Ambiguity Resolution (only for ambiguous sentences)

2. What Annotators Will See For each example, annotators are shown:

- A short video clip
- An English source sentence
- A Hindi translation (system output)

3. Evaluation Dimensions

Fluency Fluency measures whether the Hindi sentence is grammatically correct, natural, and easy to understand for a native speaker.
Ignore the English sentence and the video while judging fluency.
Guiding Question:

> "Would a native Hindi speaker say it like this?"

Example

- English: The boy is running
- Fluent: लड़का दौड़ रहा है (The boy is running)
- Disfluent: लड़का है दौड़ता

Fluency Scale (1–5)

Score	Description
5	Perfectly natural Hindi
4	Minor awkwardness
3	Understandable but unnatural
2	Hard to read
1	Broken/incorrect Hindi

Adequacy Adequacy measures whether the meaning of the English sentence is preserved in Hindi.
Ignore stylistic issues unless they change meaning.
Guiding Question:

"Does the Hindi sentence convey the same meaning as the English sentence?"

Example

- English: The girl ate an apple
- Adequate: लड़की ने सेब खाया (The girl ate an apple.)
- Inadequate: लड़की ने संतरा खाया (The girl ate an orange.) (object changed)

Adequacy Scale (1–5)

Score	Description
5	Meaning fully preserved
4	Minor detail missing
3	Some meaning loss
2	Major meaning error
1	Completely incorrect

Video Consistency Video consistency measures whether the translation aligns with what is shown in the video.
Use the video and translation (and English if needed).
Guiding Question:

"Does the translation correctly reflect the visual content?"

Example: Gender Ambiguity

- English: The doctor is here
- Video: Female doctor treating a patient
- Video-consistent: डॉक्टर **आई** हैं (The doctor [female] has come)
- Video-inconsistent: डॉक्टर **आए** हैं (The doctor [male] has come)

Example: Object Ambiguity

- English: He picked up the bat
- Video: Cricket scene
- Correct grounding: उसने क्रिकेट का बैट उठाया (He picked up a cricket bat.)
- Wrong grounding: उसने डंडा उठाया (He picked up a stick)

Video Consistency Scale (1–5)

Score	Description
5	Fully consistent with video
4	Mostly consistent
3	Partially grounded
2	Mostly inconsistent
1	Contradicts video

Pair	Metric	Fluency	Adequacy	V.C.
En–Hi	Mean	4.57	3.60	3.80
	Std. Dev.	0.32	0.44	0.71
En–Mr	Mean	4.79	4.76	3.97
	Std. Dev.	0.49	0.51	0.85
En–Od	Mean	4.53	4.33	3.97
	Std. Dev.	0.90	1.24	0.89

Table 15: Combined human evaluation results for video-guided machine translation across language pairs (V.C. denotes Video-Consistency).

A.9 Grounding Prompt for Synthetic Translation

This section describes the prompt given to the text-only LLM (Llama-3.1-8B) in the final stage of the synthetic data generation pipeline (Section 3.3). For each example, the prompt is filled with the English source sentence and the natural-language video description produced by MiniCPM-V-2.6 using the prompt in Appendix A.3. The LLM returns the visually grounded Odia translation.

The prompt is designed to keep edits conservative. The LLM is asked to use the visual context only when it resolves a lexical, semantic, or referential ambiguity in the source sentence, and otherwise to produce a direct translation without altering the meaning. This prevents the LLM from inserting visual details that are not implied by the source, which would make the synthetic targets diverge from real subtitles. The LLM is also asked to output only the translation, so that no post-hoc extraction is needed. The resulting translations then undergo the same LaBSE filtering and denoising as the real training data (Section 4.1), and their grounding is verified by human annotators (Table A.8).

Criterion	Percentage (%)	Count
Grounded	82.1	821
Partially Grounded	13.4	134
Not Grounded	4.5	45
Total	100.0	1000

Table 16: Human verification of visual grounding in the synthetic En–Od training data. For 1,000 randomly sampled triplets, verifiers were shown the source sentence, the text-only baseline translation, the MLLM-generated video description, and the grounded translation. They judged whether the grounded translation correctly incorporates the relevant visual information from descriptions, labelling it as Grounded, Partially Grounded, or Not Grounded.

Pair	Tie (%)	VMT Win (%)	Text-Only Win (%)
En-Hi	48.10	32.91	18.99
En-Mr	51.90	28.48	19.62
En-Od	55.70	26.58	17.72

Table 17: Pairwise Human Evaluation Results on the Ambiguous Test Set (Percentages Only)

Grounding Prompt (Llama-3.1-8B)

You are an expert multilingual translator and visual reasoning assistant.

Your task is to translate an English sentence into Odia. You will also be provided with visual context from an associated video.

Instructions:
1. Carefully analyse the English source sentence.
2. Examine the provided visual context to identify whether it resolves any lexical, semantic, or referential ambiguity in the source sentence.
3. If the visual context provides useful disambiguating information, produce the most accurate and contextually appropriate Odia translation using that information.
4. If the visual context does NOT provide any additional disambiguating information, or if the source sentence is already unambiguous, translate the sentence directly without altering its meaning.
5. If there is no visual disambiguation benefit, the translation should be identical to what a text-only translator would produce.
6. Preserve the original meaning, tone, and fluency.
7. Output only the final Odia translation and nothing else.

Input:
English sentence: {source_sentence}
Visual context: {visual_description}

Output:
Odia translation:

A.10 Distribution analysis of the synthetic data generated

From 18, the lexical distributions of synthetic and real subtitles differ moderately, with a Jensen Shannon (JS) divergence of 0.36. Their semantic distributions, however, are closely aligned. To measure this, we compute two distances between the sentence embeddings of the two datasets. The first is a Fréchet distance (0.013). The second is the Maximum Mean Discrepancy (MMD) (0.00025). Both values are very low, which shows that the synthetic translations are semantically close to real subtitles. Finally, the divergence between their part-of-speech (POS) distributions is near zero. This suggests that the synthetic data also preserves natural syntactic structure. These results indicate that the synthetic translations introduce limited semantic drift despite moderate lexical variation. This analysis is conducted in addition to the manual grounding verification task described in Table 16.

A.11 Evaluation set statistics

Figure 5 shows in detail the distribution of sentence length, vocabulary size, and the length ratio between the source and target language of the ambiguous evaluation set.

A.12 Significance Tests

In Table 20, we perform paired bootstrap resampling to check the significance of the obtained results for En-Hi, En-Mr and En-Od pairs for Unambiguous test sets.

A.13 MLLM Experiments

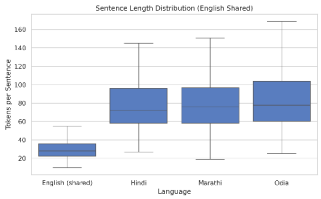
The MLLM experiments are mentioned in Table 19 and Table 21.

Category	Metric	Value	Interpretation
Lexical Distribution	KL Divergence	0.582073	Moderate lexical divergence between synthetic and real subtitles
	JS Divergence	0.356883	Moderate vocabulary shift with reasonable overlap
Semantic Distribution	Mean Embedding Distance	0.015871	Very small semantic shift; strong alignment
	Embedding FID	0.012903	Very close semantic manifold between datasets
	MMD	0.000252	Nearly identical embedding distributions
Syntactic Distribution	POS JS Divergence	0.000000	Almost identical syntactic structure

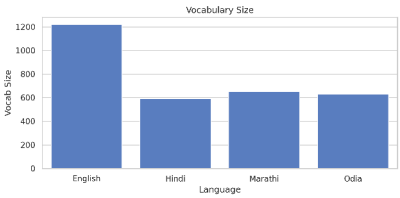
Table 18: Distribution matching statistics between synthetic and real subtitles.

Setting	En-Hi	En-Mr
0-shot	11.341	7.3882
0-shot with 1 frame	11.84	9.7201
0-shot with 3 frames	11.5989	4.7552

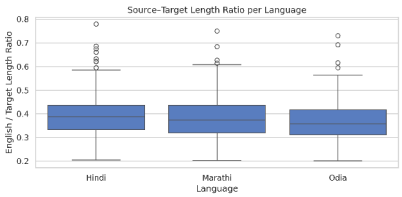
Table 19: GPT-4o-mini Performance with Multi-Image/Frame Inputs. We report BLEU scores here.



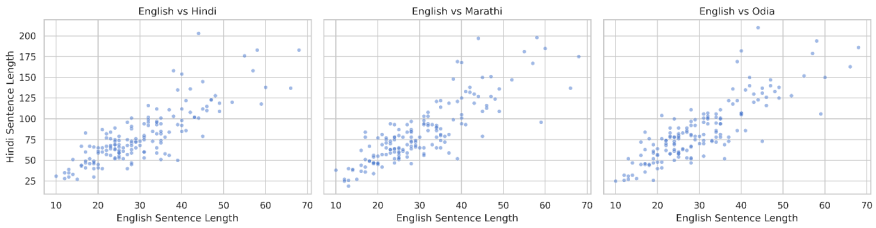
(a) Sentence Length



(b) Vocabulary Size



(c) Length Ratio



(d) Src vs Tgt Length

Figure 5: Overview of the VMT evaluation test set across all target languages. Top row: sentence length, vocabulary size. Middle row: Length ratio comparison. Bottom: source–target correlation.

Metric	Statistic	Hi	Mr	Od
BLEU	Δ	-0.0314	-0.0221	-0.0851
	CI	[-0.0449, -0.0183]	[-0.0286, -0.0164]	[-0.0920, -0.0788]
	Significant	Yes	Yes	Yes
COMET	Δ	-0.0047	-0.0046	-0.0214
	CI	[-0.0084, -0.0010]	[-0.0078, -0.0018]	[-0.0271, -0.0160]
	Significant	Yes	Yes	Yes

Table 20: Paired bootstrap resampling comparing our VMT model (LoRA) against the text-only IndicTrans2 FT (LoRA) baseline on the unambiguous test sets. Significance is computed over sentence-level scores. CI denotes the 95% confidence interval.

MiniCPM-V-2.6	En–Hi			En–Mr			En–Od		
	BLEU	COMET	BERT	BLEU	COMET	BERT	BLEU	COMET	BERT
0-shot	4.69	0.41	0.72	2.80	0.33	0.67	1.65	0.34	0.69
0-shot with text description	3.71	0.40	0.70	2.04	0.32	0.64	1.80	0.37	0.76
1-shot with text description	5.01	0.4166	0.7095	2.15	0.31	0.63	1.97	0.39	0.88
3-shot with text description	5.02	0.40	0.70	3.17	0.30	0.62	2.54	0.39	0.88

Table 21: Zero-shot and few-shot performance of MiniCPM-V-2.6 on video-guided machine translation

	Unambiguous			Ambiguous		
Model	BLEU	chrF++	COMET	BLEU	chrF++	COMET
VaTex	47.03	69.09	81.12	4.94	27.08	47.38
Ours (LoRA)	59.07	79.07	90.26	14.27	42.64	82.41
Ours (LoRA) + CTR	58.96	78.82	90.31	15.14	42.75	83.60

Table 22: Comparisons of VMT Scores with the baseline VaTex (Wang et al., 2019) for En-Hi

	Unambiguous			Ambiguous		
Model	BLEU	chrF++	COMET	BLEU	chrF++	COMET
VaTex	6.31	49.12	30.74	2.65	36.62	20.58
Ours (LoRA)	23.18	53.90	62.40	9.98	42.17	61.17
Ours (LoRA) + CTR	23.08	54.07	62.54	9.69	42.08	61.08

Table 23: Comparisons of VMT Scores with the baseline VaTex (Wang et al., 2019) for En-Mr