Video-guided Machine Translation for Indian Languages

Anonymous ACL submission

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 training strategies for VMT using 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.

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, where the subtitles occur in the particular span or context of the corresponding video.

Figure 1: Task of VMT. Given a video clip and a source sentence in English, translation to Odia with and without the video is depicted here. Here, text in red represents the ambiguous part.

Prior works (Kang et al., 2023a; 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 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 develop effective training strategies for VMT using a limited number of triplet data by leveraging strong text-only pre-trained translation models. Further-

more, 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:

- **Training Strategies:** We propose a video-guided machine translation framework that integrates video representations into a pre-trained text translation model through LoRA-based fine-tuning, and introduce a symmetric contrastive loss to align video and text embeddings for improved semantic consistency. (Section 3.1)
- **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 will be released under 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)

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 translations 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.

2.2 Video-guided Machine Translation

However, to incorporate more visual information, the possibility of adding video as a modality to translation was first explored by Wang et al. (2019) by introducing a multilingual video description dataset to benchmark video-guided translation as a task. They used dual attention and dual encoders for text and video modality. Hirasawa et al. (2020) introduced order-aware frames using positional embeddings to better capture 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. (2023b) introduced additional contrastive loss-based training. Guan et al. (2025) use LLMs to extract fine-grained audio features from the videos in natural language and use them to train an Enc-Dec to enhance the audio awareness in VMT. Lv et al. (2025) integrates 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 (Sanabria et al., 2018), EVA (Li et al., 2023b), BIGVIDEO (Kang et al., 2023a), 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 with limited or no 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 consisting of the video features v_i of a video clip, 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.

3.1 Leveraging Pre-Trained Text-Only Models

State-of-the-art text-only translation models already capture robust linguistic knowledge. Instead of training a multimodal model from scratch, 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: 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.

Video Feature Extraction Given a video V 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:

$$V = \{I_1, I_2, \ldots, I_T\}, \quad I_t \in \mathbb{R}^{H \times W \times C}.$$

where 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 d of the translation model via a feed-forward network, and the final features are:

$$V = \{z_1, z_2, \ldots, z_T\} \in \mathbb{R}^{T \times d}$$

where $z_t = FFN(VE(I_t))$

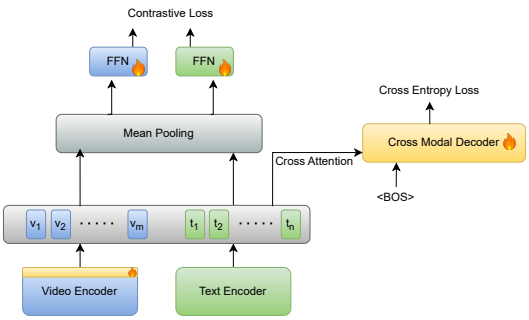


Figure 2: Our proposed VMT architecture.

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. LoRA (Low-rank Adaptation) (Hu et al., 2021) has also been used to fine-tune our model.

We use an additional symmetric Contrastive Loss for aligning the video and text embeddings. First, mean-pooled embeddings from the text and image 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 x and v denote text and video embeddings, respectively. Then:

$$\ell_{ij} = \frac{x_i^{\intercal} v_j}{\tau}$$

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

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 $(\mathcal{L}_{CE})$:

$$\mathcal{L}_{CE} = -\frac{1}{N} \sum_{n=1}^{N} \sum_{j} log(p(y_j^{(n)} | y_{<j}^{(n)}, x^{(n)}, v^{(n)}))$$

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

Where λ is a hyperparameter to control the weight given to the contrastive loss term. Diagrammatically, it can be seen in Figure 2.

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 to attend 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:

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

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

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

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 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.

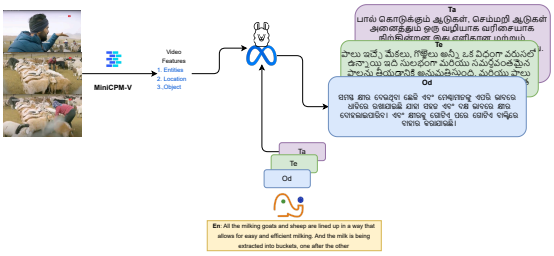


Figure 3: Synthetic data generation pipeline

As shown in Figure 3, our pipeline employs a Multimodal Large Language Model (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.

Since extending the VMT task to Indic languages requires triplet data in Indic languages, we make use of the synthetic data creation pipeline to ground the already available Indic language triplet data.

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 Indian language subtitles, and for each extracted video ID, we downloaded English and Indian subtitle tracks. We restricted ourselves to only content creators uploaded subtitles to ensure high-quality subtitles. The subtitles are aligned using a timestamp-based alignment process that matches English and Indian 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, we randomly sample 1000 sentences and ask expert annotators in Hindi, Marathi and Odia to first identify sentences which require 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. Annotation instructions and detailed procedure are provided in Appendix A.2

We chose Od (Odia) as the low-resource language for investigating the synthetic data pipeline

to train the VMT task in En-Od direction. This is because Od was found to have no parallel subtitle data available in English, which qualifies it as a low-resource language for the VMT task

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

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

To ensure data quality and semantic coherence, Language-agnostic BERT Sentence Embeddings (LaBSE) (Feng et al., 2022) filtering was applied as our final quality control step. LaBSE provides robust cross-lingual sentence representations that enable accurate semantic similarity assessment between English-Indic sentence pairs. We computed cosine similarity scores between the multilingual embeddings of aligned subtitle pairs and retained only those exceeding a predetermined similarity threshold ($= 0.9$ for Hindi, $= 0.8$ for Marathi, and $= 0.8$ for Odia). This filtering step effectively removed mistranslations, temporal misalignments, and semantically divergent pairs, resulting in high-quality parallel translations with 46K sentence pairs for Hindi, 35K for Marathi, and 25K for Odia, with corresponding video segments for each pair.

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 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. Finally, we use llama3.1-8b (Grattafiori et al., 2024) to combine video descriptions and machine translated outputs to get the visually-grounded translations for Odia.

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.

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. [1]

2. CLIP (Contrastive Language-Image Pre-training) (Radford et al., 2021) takes a different approach: it 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. [2]

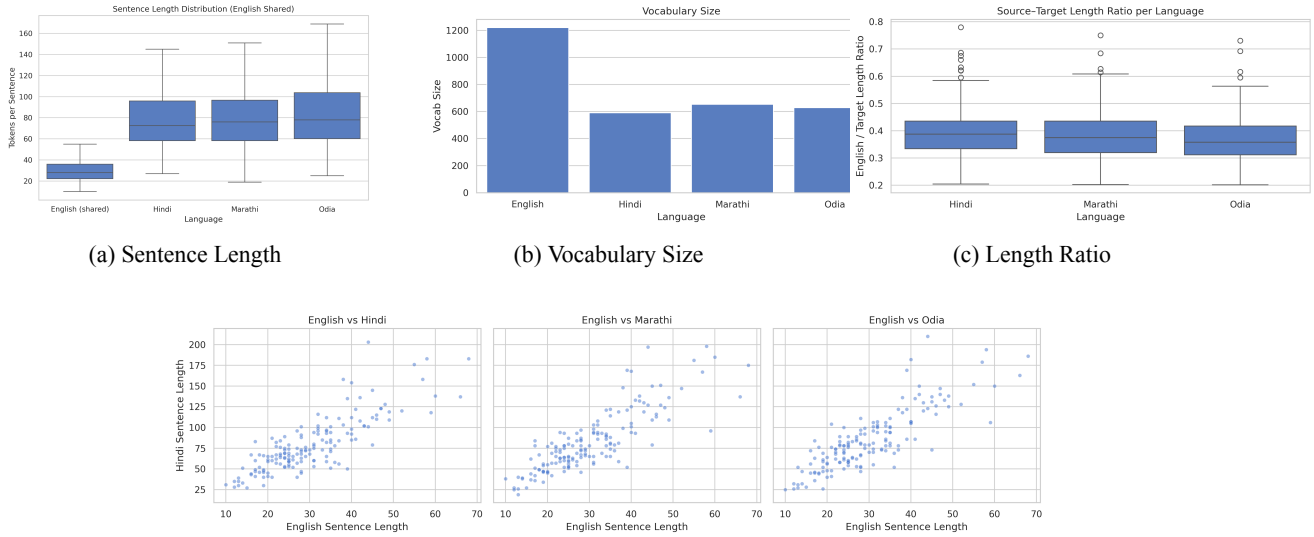
The hidden dimensions for these are 768 and 512, respectively, for ViT and CLIP. These are projected onto the dimension of the base translation model via linear layers.

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.

Evaluations 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 on the video features for each data point. This set is denoted by **Out-of-domain**.

For English-Marathi and English-Chinese directions, the test set used is a random split from our filtered dataset.

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



(a) Sentence Length (b) Vocabulary Size (c) Length Ratio

(d) Src vs Tgt Length

Figure 4: Overview of the VMT evaluation test set across all target languages. Top row: sentence length, source–target correlation, and length ratio. Bottom: vocabulary size comparison.

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 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.

5 Results and Analysis

We evaluate our proposed VMT systems across English–Hindi and English–Marathi 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 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 VGMT model with LoRA and CLIP embeddings achieves the best performance (BLEU 59.07 on the In-domain test set, COMET 90.26), outperforming the text-only fine-tuned baseline by +1.75 BLEU and +0.14 COMET. 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 and +0.12 COMET. This suggests that video modality provides stronger disambiguation when the test set differs more from training data distributions.

For English–Marathi, the addition of video features similarly improves translation, with our model achieving BLEU 30.52 and COMET 66.52 using ViT embeddings,

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), the contrastive loss slightly reduces BLEU, suggesting sensitivity to the choice of video encoder and dataset scale.

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

Model	Video Features	In-domain		Out-of-domain	
		BLEU	COMET	BLEU	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 3: Model performance on unambiguous dataset for En–Mr and En–Od

Model	Video Features	En–Mr		En–Od	
		BLEU	COMET	BLEU	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 4: Model performance on ambiguous dataset for En-Hi with ViT features

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 5: Model performance on ambiguous dataset for En-Mr with ViT features

Model	En-Mr	
	BLEU	COMET
indictrans2 FT (LoRA)	**10.0**	**61.36**
Ours (LoRA)	9.98	61.17
Ours (LoRA) + CTR	9.69	61.08

English-Chinese Translation We also evaluate our approach on the English–Chinese language pair. As can be seen from table 6, 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.

Table 6: Model performance for En-Zh with CLIP features

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

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. These findings confirm that video guidance is most beneficial in scenarios where textual information alone is insufficient. The qualitative examples are discussed in Appendix-A.3

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 8.

Two-stage Training In this setting, we first train the model with the source text heavily masked such that the decoder is forced to rely primarily on video features to generate translations. In the second stage, we introduce both full source text and video features to the model. This curriculum 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.

Table 7: Comparison of different training strategies

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

Adapters vs LoRA Finally, comparing adaptation strategies (Table 8), 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 have experimented with adapters of sizes 128, 256, and 512. The results are shown for adapters of size 256, which gave the best results of the three.

Table 8: Comparison of different Adaptation techniques

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

7 Conclusion

In this paper, we introduced a VMT an annotated evaluation set for three Indian languages: English–Hindi, English–Marathi, and English–Odia. We proposed efficient training strategies that leverage strong pre-trained text-only translation models, eliminating the need to train VMT models from scratch. Our experiments demonstrate that these approaches can also be extended to other language pairs, such as English–Chinese. Additionally, we presented a synthetic data generation framework that enables the creation of visually grounded translations, to enable the extension of VMT to low-resource languages like Odia.

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.

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. *Trans. Multi.*, 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.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30k: Multilingual english-german image descriptions.

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.

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.

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.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

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. 2023a. Bigvideo: A large-scale video subtitle translation dataset for multimodal machine translation.

Liyan Kang, Luyang Huang, Ningxin Peng, Peihao Zhu, Zewei Sun, Shanbo Cheng, Mingxuan Wang, Degen Huang, and Jinsong Su. 2023b. 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.

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.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. *arXiv preprint arXiv:2009.09025*.

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.

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 Description of Prompts

MiniCPM-V

```
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.
```

```
def generate_video_description(frames):
    msgs = [{"role": "user", "content":
        frames + [DESCRIPTION_PROMPT]}]
    params = {"use_image_id": False, "
        max_slice_nums": 2}
    return model.chat(image=None, msgs=
        msgs, tokenizer=tokenizer, **
        params)
```

A.2 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, and the average age of the annotators was 40 years. The annotators have consented for this data to be publicly available.

Annotation Guidelines 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।"

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।" **Hindi (better):** "वह रोने लगी।"

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

Speaker Identity & Gender

> **English:** ”I am tired.”
> **Video:** A woman speaking.
> **Hindi:** "मैं थक गई हूँ।"

Setting / Location

> **English:** ”It’s hot in here.”
> **Video:** People are in a kitchen.
> **Marathi:** "स्वयंपाकघरात खूप गरम आहे।"

Emotional Expressions & Tone

> **English:** ”Oh, that’s just great!” (said sarcastically)
> **Odia:** "ବାହ! ବହୁତ ଭଲ!" (sarcastic tone)

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.”
> **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.”
> **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।"
> **Extra Info Added:** Emotional tone included.
> **Notes:** Video context indicates emotional breakdown.

A.3 Qualitative Examples

In Table 9, we present qualitative examples comparing translations generated by our model trained using video+text versus text-only inputs. Each source word is shown with the same sentence context, highlighting how the addition of visual information can influence the target-language output.

Source	Modality	Predicted	Sentence / Context
round	Video + Text	दौर	Now, I'm doing a round again.
	Text Only	राउंड	*Ref:* अब, मैं फिर से एक दौर कर रहा हूँ
bead	Video + Text	मनका	I'll take a bead now.
	Text Only	बीड	*Ref:* मैं अब एक मनका लूंगा।
case	Video + Text	मामले	Install the power button in the case.
	Text Only	केस	*Ref:* मामले में पावर बटन स्थापित किया।

Table 9: Illustrative translation examples comparing text-only and video+text outputs. The sentence/context is shared for both modalities.