# Video-guided Multimodal Machine Translation: A Survey of Models, Datasets, and Challenges

**Anonymous ACL submission**

## Abstract

In recent years, machine translation has evolved with the integration of multimodal information. Infusion of multi-modality into translation tasks decreases disambiguation and enhances translation scores. Common modalities include images, speech, and videos, which provide additional context alongside the text to be translated. While multimodal translation with images has been extensively studied, video-guided machine translation (VMT) has gained increasing attention, particularly since (Wang et al., 2019) first explored this task. In this paper, we provide a comprehensive overview of VMT, highlighting its unique challenges, methodologies, and recent advancements. Unlike previous surveys that primarily focus on image-based multimodal translation, this work explores the distinct complexities and opportunities introduced by video as a modality.

## 1 Introduction

Multimodal Machine Translation (MMT) improves translation by incorporating more context. This context can be in the form of images, audio and video. This infusion of extra context helps in disambiguation of translated text and make it more meaningful and accurate. MMT often mimics the way human translators annotated data. They take into account all information that emanates from all modalities while translating the sentence in source language to target language. While MMT mostly focuses on images being the additional modality to the source text sentence, Video-guided machine translation has been picking immense interest as compared to other MMT techniques due to its ability to provide richer, more dynamic contextual information than images.

VMT takes advantage of the temporal and multimodal nature of videos, which combine visual, auditory, and textual data into a single cohesive source of information. Unlike static images, videos capture sequences of events, actions, and interactions, offering a more comprehensive understanding of the context. This makes video-based MMT particularly effective for tasks such as translating instructional videos, movies, or multimedia content, where temporal alignment and multimodal fusion are critical. For example, in a cooking video, the translation of a spoken instruction (e.g., *"chop the onions"*) can be disambiguated by the visual demonstration of the action, ensuring the translation is both accurate and contextually appropriate. The importance of video-based MMT lies in its ability to address several limitations of traditional text-based and image-based translation systems. Videos provide temporal continuity, enabling models to capture the progression of events and actions over time. Second, the integration of multiple modalities (text, audio, and video) allows for more robust disambiguation of ambiguous terms or phrases. VMT has practical applications in real-world scenarios, such as cross-lingual video captioning, multimedia content localization, and assistive technologies for the hearing impaired.

In this paper, we provide a comprehensive survey of video-based multimodal machine translation, focusing on its methodologies, challenges, and advancements. Unlike previous surveys that primarily focus on image-based MMT, this work highlights the unique aspects of video-guided MMT and its growing importance in the field. We systematically categorize and analyze state-of-the-art approaches, datasets, and evaluation metrics, while also identifying key open problems and future research directions. By bridging the gap between traditional text-based translation and video-based MMT, this survey aims to serve as a valuable resource for researchers.

## 2 Background and Preliminaries

Multimodal Machine Translation (MMT) incorporates multiple modalities, such as images, speech, or videos, to improve translation quality. Image-guided machine translation (IMT), which uses visual information as an additional modality, gained momentum with the introduction of the Multi30K dataset by (Elliott et al., 2016). However, the scarcity of paired image-text datasets led to alternative approaches such as retrieval-based image machine translation (Fang and Feng, 2022; Tang et al., 2022; Zhang et al., 2020), which retrieves relevant images, and text-to-image-based machine translation (Calixto et al., 2019; Li et al., 2022a; Long et al., 2021; Yuasa et al., 2023; Guo et al., 2023), where synthetic images are generated from text. Beyond IMT, text-in-image machine translation (Chen et al., 2023; Lan et al., 2023; Ma et al., 2022, 2024, 2023) focuses on translating text embedded within images. Another development in MMT is simultaneous machine translation (SiMT) (Haralampieva et al., 2022; Imankulova et al., 2020; Ive et al., 2021), which generates translations before receiving the full input to reduce latency while maintaining quality. More recently, video-based machine translation has emerged, incorporating temporal information alongside visual and textual data for improved translation accuracy.

## 3 Problem Formulation

The task of {video-guided multimodal machine translation (VMT) involves generating accurate and contextually appropriate translations of source language text by leveraging additional modalities such as video and audio. Formally, given a source language text $S = \{s_1, s_2, \ldots, s_n\}$ and a corresponding video frame sequence $V = \{v_1, v_2, \ldots, v_m\}$ (which may include associated audio $A = \{a_1, a_2, \ldots, a_k\}$), the goal is to produce a target language translation $T = \{t_1, t_2, \ldots, t_p\}$ that is linguistically accurate and contextually aligned with the multimodal input. The objective of video-guided MT is to learn a mapping function $f$ that maximizes the likelihood of the target translation $T$ given the source text $S$, video $V$, and audio $A$, expressed as

$$f(S, V, A) = \arg\max_{T} P(T \mid S, V, A).$$

This involves optimizing model parameters to minimize the discrepancy between the predicted translation $\hat{T}$ and the ground truth $T$, typically using cross-entropy loss or other sequence-level objectives. The integration of video and audio modalities introduces unique challenges, such as temporal alignment, modality heterogeneity, and scalability, which distinguish video-based MT from traditional text-based or image-based MT and necessitate specialized approaches to effectively harness the rich, dynamic information provided by multimodal inputs.

## 4 Video Guided Machine Translation.

Video-guided multimodal MT leverages multiple modalities (text, video, and audio) to improve translation quality. The approaches can be broadly categorized based on how they handle **modality fusion**. Below and in Fig. 1, we present a taxonomy of these approaches, focusing on **Late Fusion**, **Early Fusion**, and **Hybrid Fusion**.

### 4.1 Late Fusion

The early approaches in VMT utilized separate encoders for video and text modalities and combined them at a later stage in the VMT pipeline.

**(Wang et al., 2019)** designed a multimodal sequence to sequence model with temporal attention and source attention for videos and text embeddings respectively.

**(Hirasawa et al., 2020)** introduce a novel approach to video representation in machine translation by incorporating positional encodings, making the model aware of the temporal order of frames. They further enhance the video representation by distinguishing between two types of features: action and appearance. The action features, captured by a dedicated video encoder, focus on motion information crucial for disambiguating verbs in the translation process. Conversely, appearance features, extracted by an image encoder, provide detailed information about objects and scenes within each frame, aiding in the disambiguation of nouns. This dual-feature approach allows the model to better align visual cues with textual elements.

**(Gu et al., 2021)** introduce a novel approach to video representation inspired by Hierarchical Attention Networks (HAN) (Miculicich et al., 2018). Their model divides video input processing into two distinct components: motion representation and spatial representation. For capturing motion dynamics, they employ a pretrained I3D (Carreira and Zisserman, 2017) network. The spatial aspect is handled by a specialized HAN, which constructs

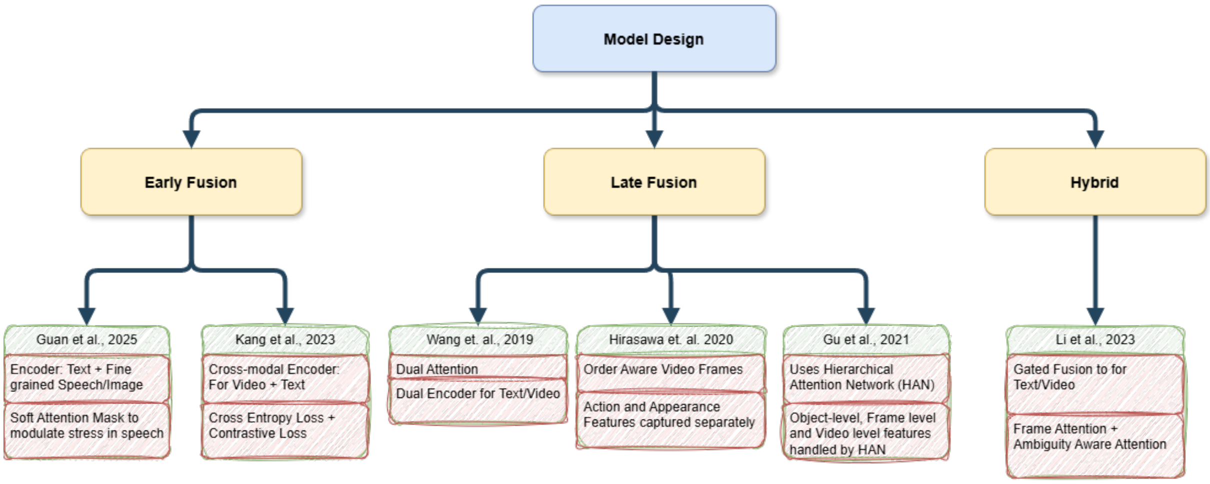


Figure 1: Taxonomy for Video Guided Machine Translation

a multi-level representation hierarchy: object-level, frame-level, and video-level. In this special HAN, each successive level of representation serves as a helper for the higher level, allowing for a progressively more comprehensive understanding of the video's spatial content. The object-level features inform the frame-level representation, which in turn contributes to the overall video-level understanding. This hierarchical approach enables the model to capture both fine-grained spatial details and broader contextual information. For generating the translated sentence, the authors utilize a GRU (Gated Recurrent Unit) (Chung et al., 2014) network as the decoder.

### 4.2 Early Fusion

Where different modalities are embedding tigether befre being passed on to a shared encoder:

**(Kang et al., 2023)**introduces a cross-modal encoder that jointly processes video and text representations. The model enhances video features with positional encodings to capture temporal information. This cross-modal architecture enables the model to focus on relevant parts of both text and video inputs, facilitating more effective multimodal understanding. The training process incorporates two key objectives: cross-entropy loss in the decoder for sequence generation, and a novel cross-modal contrastive learning (CTR) objective. The CTR objective is designed to learn shared semantics between video and text modalities, encouraging similar video-text pairs to have closer representations while pushing dissimilar pairs apart in the embedding space.

**(Guan et al., 2025)** introduces the FIAT architecture, a uni-modal encoder that integrates multiple fine-grained inputs for video-guided translation. The model incorporates various types of tags, including entities, audio sentiments, locations, expressions, and video captions, alongside source subtitles. This rich set of inputs enables a more comprehensive understanding of the video content. The cross-modal encoder processes these diverse inputs jointly, allowing for complex interactions between different modalities. To capture nuanced speech information, the architecture employs a soft attention mask that incorporates stress patterns from the audio. This attention mechanism helps the model focus on emphasized parts of speech, potentially improving the accuracy and naturalness of translations.

### 4.3 Hybrid Fusion

**(Li et al., 2023)** introduce SAFA (Selective Attention with Frame Attention), a novel approach for video-guided machine translation that integrates two key innovations: frame attention and selective attention. The frame attention mechanism, inspired by gated fusion techniques, encourages the model to focus on the most relevant video frames, particularly central frames where subtitles typically appear, implemented through a frame attention loss. The selective attention component dynamically determines when to leverage visual information for translation, especially useful for handling ambiguous text. To further enhance the model's ability to handle ambiguity, SAFA incorporates an ambiguity-aware loss, encouraging heavier reliance on video information for ambiguous text while prioritizing textual cues for non-ambiguous cases.

### 4.4 Datasets

Table 1 presents all the datasets used in Video-guided machine Translation. Detailed analysis of the dataset in given in Appendix A

| Dataset | Language | Domain | # Clip |
|---|---|---|---|
| How2 | En-Pt | instruction | 189K |
| VATEX | EN-Zh | caption | 41K |
| VISA | En-Ja | subtitle | 40K |
| EVA | En-Zh/Ja | subtitle | 1.4M |
| BigVideo | En-Zh | subtitle | 3.3M |
| MAD-VMT | En-Zh | caption | 193K |
| TriFine | En-Zh | subtitle | 2.4M |

Table 1: Dataset Information of Various VMT datsets as in (Guan et al., 2025)

## 5 Challenges and Future Directions

This section discusses about various challenges in VMT and also points towards possible future research directions

### 5.1 Challenges

**Information Redundancy and Computational Overhead** According to (Guan et al., 2025), VMT requires selecting multiple frames to extract coarse-grained visual features. However, not all frames contribute equally to translation quality, leading to increased computational overhead. The inclusion of redundant frames can also introduce regularization issues, impacting model performance.

**Audio Integration in VMT** While VMT primarily relies on visual cues for translation, incorporating audio is crucial. Audio provides essential contextual information, such as speaker intent, tone, and background sounds, which significantly enhance translation accuracy. However, effectively fusing audio with video representations remains a challenge. (Guan et al., 2025) has only introduced a trimodal dataset with audio and fine grained tags.

**Data Scarcity in Low-Resource Languages** VMT models require triplet data—video, source text, and target text—for training. However, such datasets are scarce, particularly for low-resource languages and underrepresented language families. This data bottleneck limits the scalability and generalization of VMT models.

### 5.2 Future Directions

**Integrating World Knowledge** Enhancing VMT with external world knowledge, such as named entities (famous personalities, cultural references) and idiomatic expressions, could improve translation accuracy. Techniques like knowledge graph integration or retrieval-augmented generation could be explored.

**Leveraging Large Multimodal Models** Pre-trained large-scale multimodal models, trained on extensive text-image corpora, could be fine-tuned for VMT. These models inherently capture rich cross-modal representations, making them valuable for video-based translation tasks.

**High-Quality Multilingual and Domain-Specific Datasets** Developing large-scale, high-quality datasets across multiple language families and diverse domains is essential for improving VMT. This would address current data scarcity challenges and enhance translation performance in various contexts.

**Real-Time Translation with Low Latency** Achieving real-time video-based translation with minimal latency is a key goal. Optimizations such as efficient frame selection, lightweight transformer architectures, and parallelized inference pipelines could be explored to enable low-latency, high-accuracy translations.

## 6 Conclusion

In this paper, we provide a comprehensive overview of video-guided machine translation (VMT). We begin by discussing the background and evolution of multimodal machine translation (MMT) to VMT. Next, we present a taxonomy of various VMT approaches based on their model design. We then review the datasets commonly used for VMT research. Finally, we discuss the key challenges in VMT and explore potential future directions for advancing this task.

## Limitations

Since video-guided machine translation is an emerging field, any survey on this topic must be continuously updated to reflect new research developments. As new datasets, models, and approaches are introduced, the landscape of VMT evolves rapidly, making it challenging to maintain a comprehensive and up-to-date overview.

## 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.

João Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733.

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.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *Preprint*, arXiv:1412.3555.

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.

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.

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.

Veneta Haralampieva, Ozan Caglayan, and Lucia Specia. 2022. Supervised visual attention for simultaneous multimodal machine translation. *J. Artif. Intell. Res.*, 74:1059–1089.

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.

Aizhan Imankulova, Masahiro Kaneko, Tosho Hirasawa, and Mamoru Komachi. 2020. Towards multimodal simultaneous neural machine translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 594–603, Online. Association for Computational Linguistics.

Julia Ive, Andy Mingren Li, Yishu Miao, Ozan Caglayan, Pranava Madhyastha, and Lucia Specia. 2021. Exploiting multimodal reinforcement learning for simultaneous machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3222–3233, Online. 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.

Zhibin Lan, Jiawei Yu, Xiang Li, Wen Zhang, Jian Luan, Bin Wang, Degen Huang, and Jinsong Su. 2023. Exploring better text image translation with multimodal codebook. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3479–3491, Toronto, Canada. Association for Computational Linguistics.

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. 2023. 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.

Quanyu Long, Mingxuan Wang, and Lei Li. 2021. Generative imagination elevates machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5738–5748, Online. Association for Computational Linguistics.

Cong Ma, Xu Han, Linghui Wu, Yaping Zhang, Yang Zhao, Yu Zhou, and Chengqing Zong. 2024. Modal contrastive learning based end-to-end text image machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2153–2165.

Cong Ma, Yaping Zhang, Mei Tu, Xu Han, Linghui Wu, Yang Zhao, and Yu Zhou. 2022. Improving end-to-end text image translation from the auxiliary text translation task. *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 1664–1670.

Cong Ma, Yaping Zhang, Mei Tu, Yang Zhao, Yu Zhou, and Chengqing Zong. 2023. E2timt: Efficient and effective modal adapter for text image machine translation.

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.

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.

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.

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.

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.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5288–5296.

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*.

## A Dataset Details

**Vatex** datset introduced in (Wang et al., 2019) is one of the most widely used benchmarks for video-based multimodal machine translation. It consists of multilingual video descriptions and is designed to facilitate research in video captioning and translation. The dataset contains over 41,000 videos collected from the MSR-VTT (Xu et al., 2016) dataset, with each video annotated with 10 English descriptions and their corresponding translations in Mandarin Chinese. The videos cover a diverse range of topics, including sports, music, and everyday activities, making it a robust resource for training and evaluating multimodal MT models.

**EVA** (Li et al., 2023) is a more recent dataset focused on educational videos, designed to support research in translating instructional content. It includes videos from online educational platforms, annotated with {source text (English) and **target translations (multiple languages)**. EVA is particularly useful for studying the translation of domain-specific content, such as lectures, tutorials, and demonstrations. Key features of EVA include its focus on educational content, multilingual translations for diverse target languages, and high-quality audio and visual data. However, it poses challenges such as the need for domain-specific knowledge to handle technical terminology and complex sentence structures. EVA is widely used for translating instructional and educational videos, as well as for domain adaptation in multimodal MT.

**How2** (Sanabria et al., 2018) was one of the first datasets addressing multimodal language understanding. It contains 79,114 instructional videos along with English subtitles and aligned Portuguese subtitles. All the clips contain the summary of the event occurring in the clip.

**VISA** (Li et al., 2022b) contains clips from movies and TV along with parallel subtitles in English and Japanese. All subtitles are ambiguous and fall into either the "Polysemy" or "Ambiguous" category. Hence, any translation task involving these subtitles must rely on the corresponding video clip for context.

**BigVideo** (Kang et al., 2023)is a large-scale dataset specifically focusing on video subtitle translation. It contains 4.5 million English-Chinese sentence pairs aligned with 156,000 unique videos, totaling 9,981 hours of content. It is currently the largest video-guided machine translation dataset available. BigVideo contains two specially annotated test sets: Ambiguous and Unambiguous. The Ambiguous set contains source inputs that require video context for accurate translation, while the Unambiguous set includes self-contained text suitable for translation without visual cues.

The **MAD-VMT** (Shurtz et al., 2024) (Movie Audio Descriptions for Video-guided Machine Translation) dataset is derived from the MAD dataset, which contains transcribed audio descriptions of movies typically used for visually impaired audiences. To create MAD-VMT, the English transcriptions from MAD were machine-translated into Chinese using Google Translate. This approach was adopted to increase the amount and lexical diversity of both source and target language pretraining data for video-guided machine translation tasks. The dataset underwent quality control using the COMET-QE metric, resulting in approximately 193,130 sentence pairs (about 69% of the original size) after filtering. Unlike the original MAD dataset, MAD-VMT includes character names in the training set instead of replacing them with generic tokens, making it more suitable for translation tasks where the source text can provide context for character names.

**Trifine** (Guan et al., 2025) is a comprehensive tri modal dataset designed for vision-audio-subtitle analysis and translation tasks. It features a parallel corpus of English-Chinese subtitles, complemented by fine-grained audio labels such as audio sentiment and stress, as well as video labels including location, entities, expressions, and actions.