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 ambiguation 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-guided multimodal translation, this work explores the distinct complexities and opportunities introduced by video as a modality.

Source Subtitle: ……Number one **drive shot** requires smaller swing but more focus.
Target Subtitle: ······第一、抽球。挥杆幅度要小，但是要集中力量。
System w/o Video: ······第一个驾驶镜头需要较小的挥杆，但更多的焦点。
System w/ Video: ······第一，抽球需要更小的挥杆动作，但要集中注意力

Figure 1: A case. The phrases with semantic ambiguity are highlighted in red. The wrong translations are in blue and the correct translations are in yellow taken from Kang et al. (2023)

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 makes it more meaningful and accurate. MMT often mimics the way human translators annotate data. They take into account all the 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 VMT 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. In Fig. 1 the phrase "drive shot" is better translated by VMT system by understanding the meaning of *"shot"*.

The importance of video-guided MMT lies in its ability to address several limitations of traditional text-based and image-guided translation systems. Videos provide temporal continuity which enable 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-guided MMT, focusing on its methodologies, challenges, and advancements. Unlike previous surveys (Shen et al., 2024; Paul et al., 2024) that primarily focus on general aspects of MMT and image-guided MT, this work specifically 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 and datasets while also identifying key open problems and future research directions.

Our contributions are:

1. A novel taxonomy for video-guided multi-modal machine translation, which systematically categorizes existing VMT approaches. (Section 4)
2. Comprehensive comparisons of methods, datasets, and state-of-the-art systems provided. (Section 7)
3. Identifing key challenges and future research directions to guide further advancements in Video guided MT. (Section 9)

2 Background and Preliminaries

Machine translation involves translating texts from one language to another language. From statistical to neural MT has undergone pioneering transformations. We discuss below various stages of MT developments connecting it with VMT.

2.1 Neural Machine Translation

Neural Machine Translation (NMT) has evolved significantly through key innovations in neural architectures. Sutskever et al. (2014) pioneered sequence-to-sequence learning using LSTMs, demonstrating that reversing source sentences improved translation by shortening dependencies, achieving a BLEU score of 34.8 on English-French tasks. Bahdanau et al. (2016) introduced attention mechanisms, enabling dynamic focus on relevant source segments and addressing long-sequence limitations. Luong et al. (2015) refined this with global and local attention models. The transformer architecture (Vaswani et al., 2023) eliminated recurrence entirely, using self-attention for superior parallelization. Subword segmentation techniques like byte-pair encoding (Sennrich et al., 2016) improved rare-word handling through compositional translation units. Multilingual NMT systems achieved zero-shot translation via shared parameters and language tokens, revealing interlingual representations (Wu et al., 2016).

2.2 Image Guided Machine Translation

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., 2022a; Zhang et al., 2020), which retrieves relevant images, and text-to-image-guided 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.

2.3 Other Forms

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.

In all of the above cases videos are not a part of the modeling. Therefore video-guided machine translation has emerged which incorporates temporal information alongside visual and textual data for improved translation accuracy.

3 Problem Formulation

The task of VMT involves contextually appropriate translations of source language text by utilizing 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

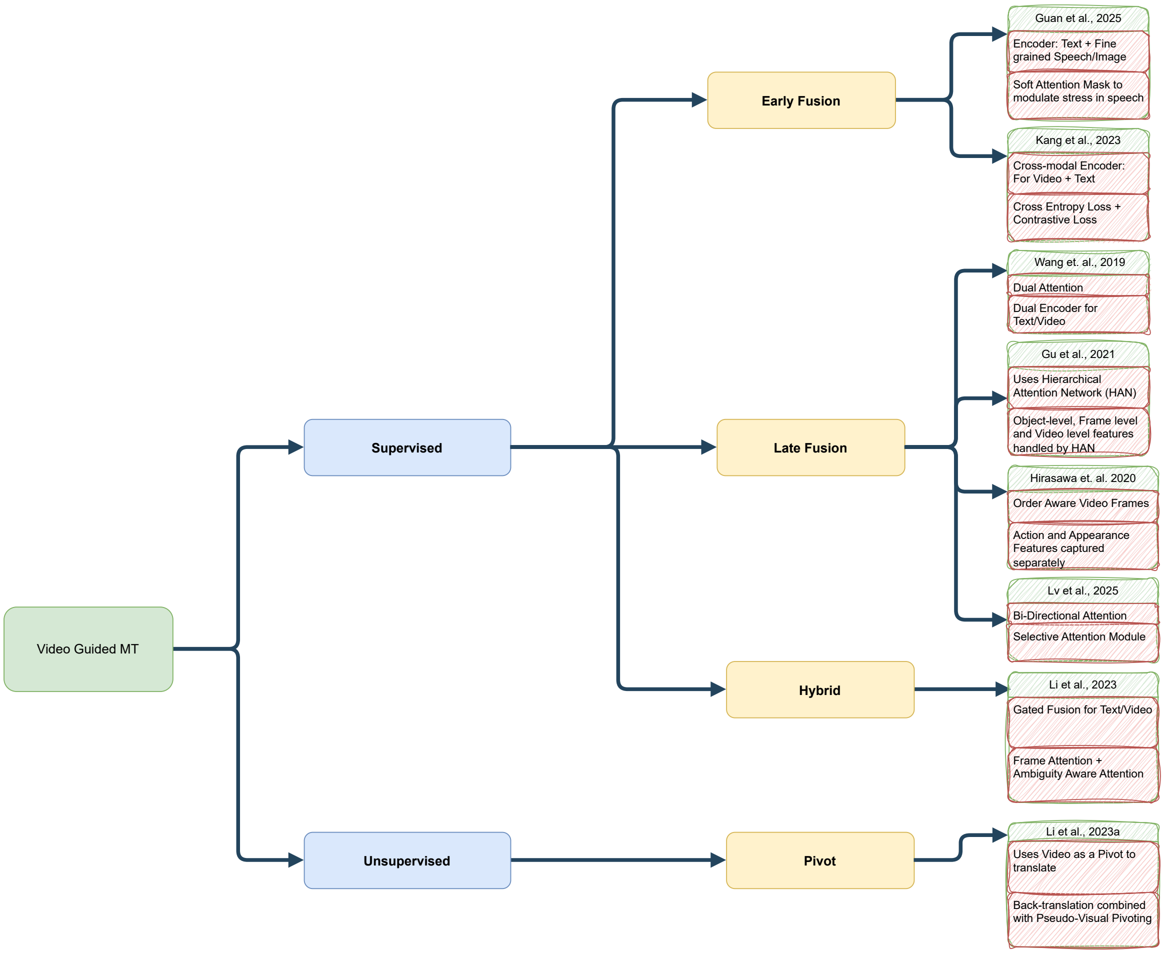

Figure 2: Taxonomy for Video Guided Machine Translation

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 and scalability, which distinguish video-guided MT from traditional text-based or image-guided MT and necessitate specialized approaches to effectively harness the rich, dynamic information provided by multimodal inputs. 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. 2, we present a taxonomy of these approaches, with supervised approaches focusing on **Late Fusion**, **Early Fusion**, **Hybrid Fusion** and unsupervised approaches focusing on **Video Pivoting**.

4 Video Guided Machine Translation.

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

Models	Datasets	Modelling Approaches	En-Zh	Zh-En
Wang et al. (2019)	VaTex	Dual Attention and Dual Encoder for Text/Video	29.1	26.4
Hirasawa et al. (2020)	VaTex	Order-aware video frames using positional embeddings.	35.4	-
Gu et al. (2021)	VaTex	Hierarchical Attention Network (HAN) applied at object, frame, and video levels.	35.9	-
Li et al. (2023b)	EVA	Introduces Frame Attention and Ambiguity-Aware Attention.	-	27.6
Li et al. (2023a)	Vatex	Uses Video as Pivot between languages	29.6	26.6
Kang et al. (2023)	VaTex BigVideo	Introduces additional contrastive loss.	37.6 44.8	-
Guan et al. (2025)	TriFine	Uses fine-grained speech features with soft attention masks.	38.06	25.51
Lv et al. (2025)	TopicVD	Uses selective attention and Bi-Attention on Text and Videos.	29.33	-

Table 1: Overview of Multimodal Translation Models, Approaches, and BLEU scores in En-Zh and Zh-En Directions

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 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.

Lv et al. (2025). integrates the selective attention module and the bidirectional attention module by taking inspiration from Li et al. (2021) and Tang et al. (2022b). Their architecture utilizes two encoders each for video and source text and fuses the obtained representations using a cross modal bidirectional attention mechanism. The fused representations are then decoded into target-language subtitles using an autoregressive transformer decoder. An empirical evaluation across multiple domains reveals that the model's performance notably diminishes in out-of-domain scenarios.

4.2 Early Fusion

This fusion occurs when different modalities are embedding together before 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. 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, improving the accuracy and naturalness of translations.

4.3 Hybrid Fusion

Li et al. (2023b) introduce SAFA (Selective Attention with Frame Attention) 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. This is 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 Unsupervised Methods

Li et al. (2023a) uses videos to serve as a "universal pivot" to bridge language pairs without parallel corpora, with spatial-temporal graphs providing fine-grained visual grounding for both close and distant language pairs. Video pivoting in MMT leverages visual content from videos as an intermediary to align source and target languages in unsupervised settings. This approach addresses the challenge of latent space alignment between languages by exploiting the shared visual-semantic information in videos, which provide richer spatial-temporal context than static images. The core mechanism involves multimodal back-translation combined with pseudo-visual pivoting, where models learn a shared multilingual embedding space.

Table 1 presents a comparison between all existing approaches.

5 Video Encoders

Recent advances in video encoding architectures have significantly expanded the toolkit for video understanding in VMT tasks moving beyond traditional 3D CNNs and ResNet-based approaches to specialized transformer architectures and cross-modal alignment strategies. Transformer-based models like VideoSwin Transformer (Liu et al., 2021) introduced locality-constrained spatiotemporal attention through shifted window mechanisms which reduced computational costs by 20× compared to 3D CNNs through hierarchical feature processing. Concurrently, ViViT (Arnab et al., 2021) demonstrated pure-transformer efficacy by factorizing spatial-temporal tokens and leveraging image-pretrained weights through temporal adaptation of vision transformers. Contrastive learning frameworks such as CLIP4Clip (Luo et al., 2021) adapted image-text pretrained CLIP models for video retrieval via parameter-free similarity calculation and temporal alignment modules and jointly optimized video-text embeddings. This paradigm was extended by VideoCLIP(Xu et al., 2021), which incorporated hard negative mining during contrastive pretraining to boost zero-shot performance on video QA and aslo enabled temporal localization without task-specific fine-tuning. Emerging foundational encoders like VideoPrism

(Zhao et al., 2024) unified global-local video understanding through hybrid contrastive and masked autoencoding pretraining. For multimodal integration, VideoGPT+ (Maaz et al., 2024b) introduced dual spatial-temporal pathways combining ViT-L/14 image encoders with TimeSformer (Bertasius et al., 2021) video models via adaptive pooling gates. The MERV (Chung et al., 2025) framework advanced specialized knowledge fusion by spatiotemporally aligning features from DINOv2 (Oquab et al., 2024), ViViT (Arnab et al., 2021)(temporal), and SigLIP (Zhai et al., 2023) encoders through cross-attentive mixing, boosting VideoLLM performances. These architectures collectively address VMT's core requirements - balancing spatial-temporal resolution, cross-modal alignment, and computational efficiency - while providing adaptable frameworks for integrating domain-specific visual knowledge into translation pipelines.

6 Analysis

This section presents a targeted analysis of recent multimodal translation models, focusing on three critical areas: cross-modal fusion strategies, the use of auxiliary loss functions, and the scaling of datasets and input modalities.

Cross-Modal Fusion: The design and depth of cross-modal fusion have a significant impact on translation quality. Early approaches, such as Wang et al. (2019), employed dual attention and dual encoders to handle video and text inputs separately. While foundational, these architectures lacked the capacity to model complex interactions between modalities. Subsequent models introduced more sophisticated fusion techniques—Hirasawa et al. (2020) encoded frame order through positional embeddings, and Gu et al. (2021) advanced this with hierarchical attention across object, frame, and scene levels. These enhancements led to notable improvements in En-Zh BLEU scores, underscoring the importance of structured and temporally-aware fusion mechanisms. Similarly, Lv et al. (2025) introduced discourse-level topic information via selective and bi-attention, though the impact was relatively modest—indicating that while more information can help, effective integration is crucial.

Auxiliary Loss Functions: The introduction of auxiliary learning objectives, particularly contrastive losses, has proven effective in strengthening cross-modal alignment. Kang et al. (2023) achieved the highest En-Zh BLEU score by combining a contrastive loss with standard translation objectives. This allowed the model to more effectively discriminate between semantically aligned and unaligned video-text pairs. Similarly, Li et al. (2023b) leveraged ambiguity-aware attention as a form of auxiliary supervision, yielding the highest Zh-En BLEU score. These results demonstrate that auxiliary objectives targeting representation quality and semantic clarity can lead to significant translation gains. (Li et al., 2023a) uses back translation for latent space alignment for videos and text with pseudo-visual pivoting.

Scaling of Data and Additional Features: Model performance has also benefited from scaling both data and modalities. The use of large-scale datasets, as in Kang et al. (2023), clearly contributes to better generalization and more robust cross-modal representations. Additionally, Guan et al. (2025) incorporated fine-grained speech features alongside video and text, achieving strong En-Zh performance. However, the asymmetry in Zh-En results suggests that the effectiveness of additional modalities such as audio depends on language direction or modality alignment quality.

7 Datasets

Table 2 presents all the datasets used in Video-guided machine Translation.

Vatex dataset introduced in (Wang et al., 2019) is one of the most widely used benchmarks for video-guided 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., 2023b)is a large-scale resource focused on subtitle ambiguity. It contains 852,000 Japanese-English and 520,000 Chinese-English parallel subtitle pairs, each aligned with corresponding video clips sourced from movies and TV episodes. EVA also features a specially curated evaluation set where subtitle ambiguity is guaranteed and the accompanying video is necessary for

Dataset	Language	Clips	Secs	Sen	Domain	Genre	AM	FT	S	A-S Alignment	TB
How2	En-Pt	186K	5.8	186K	Instruction	Short Video	×	×	✓	✓	×
VATEX	En-Zh	41K	10	129K	Captions	Short Video	×	×	✓	×	×
VISA	En-Ja	40K	10	40K	Subtitle	Film and Television	✓	×	×	×	×
EVA	En-Zh/Ja	1.4M	10	1.4M	Subtitle	Film and Television	✓	×	×	×	×
BigVideo	En-Zh	3.3M	8	4.5M	Subtitle	Short Video	✓	×	×	×	×
MAD-VMT	En-Zh	193K	-	193K	Caption	Movies	×	×	×	×	×
Trifine	En-Zh	2.4M	10	2.4M	Subtitle	Short Video	✓	✓	✓	✓	×
TopicVD	En-Zh	122K	8.4	122K	Subtitle	Documentary	×	×	×	✓	✓

Table 2: Overview of Video Guided Machine Translation Datasets. "Secs" denote the duration of each clip. "Sen" denote the number of sentences in the dataset. "AM" denote the availability of ambiguity-aware dataset. "FT" denotes the availability of fine-grained tags of the dataset. "S" denotes the availability of Audio. "A-S" alignment indicates whether the Audio-Video are aligned. "TB" denotes topic based segragation of the dataset.

disambiguation, directly addressing a major limitation of prior MMT datasets.

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.

TopicVD (Lv et al., 2025) is a topic-based dataset designed for VMT of documentaries, addressing the lack of large-scale, diverse video data in long-form videos. It consists of 256 documentaries spanning eight topics - Economy, Food, History, Figure, Military, Nature, Social, and Technology, comprising 285 hours of video and 122,930 Chinese-English parallel subtitle pairs, with contextual information for each video-subtitle pair. The dataset enables research on domain adaptation as experiments show that visual and contextual information significantly enhance translation performance, especially in in-domain scenarios.

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.

8 Previous Surveys

Shen et al. (2024) explores Multimodal Machine Translation in detail covering various aspects like Image-guided MT, In-Image MT, Video-guided MT and Chat Multimodal MT. It explores image-guided MT in utmost detail, underlining its modelling approaches and datasets in detail. It also touches upon various works which analyze the extent of the importance of images in improving the translations. However, Shen et al. (2024) doesn't explore the intricacies of video-guided MT by going into the depth of modeling and taxonomy of VMT. Similarly, Paul et al. (2024) surveys MMT papers related to Indian Languages with Image-guided MT in focus. Video-guided MT differs from general Multimodal MT which covers wide range of MMT tasks. Since video modality is information heavy, it demands its own analysis and dedicated survey.

9 Challenges and Future Directions

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

9.1 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. Table 2 shows that most video-guided MT datasets consist of English and Chinese data with no representation from other language families.

9.2 Future Directions

Integrating World Knowledge using Video LLMs 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. Pretrained large-scale multimodal models, trained on extensive text-image corpora, could be fine-tuned for VMT. Video LLMs like Maaz et al. (2024b), Cheng et al. (2024) and Maaz et al. (2024a) inherently capture rich cross-modal representations and have instruction following ability making them valuable for video-guided translation tasks which may involve reasoning. Video-LLMs also provide an able ground to explore agentic machine translation setups as seen in Wu et al. (2025) which can further the reasoning and generalization ability of translation systems.

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. Only Lv et al. (2025) currently has domain specific segregation of data in English and Chinese.

Real-Time Translation with Low Latency Achieving real-time video-guided 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. Recently Chen et al. (2024) attempted to cruch stream video using Video LLMs. However, they lose out on better representation for spatial and temporal features.

10 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

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. 2021. Vivit: A video vision transformer. *Preprint*, arXiv:2103.15691.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. Neural machine translation by jointly learning to align and translate. *Preprint*, arXiv:1409.0473.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. 2021. Is space-time attention all you need for video understanding? *Preprint*, arXiv:2102.05095.

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.

Joya Chen, Zhaoyang Lv, Shiwei Wu, Kevin Qinghong Lin, Chenan Song, Difei Gao, Jia-Wei Liu, Ziteng Gao, Dongxing Mao, and Mike Zheng Shou. 2024. Videollm-online: Online video large language model for streaming video. In *CVPR*.

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.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. 2024. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *Preprint*, arXiv:2406.07476.

Jihoon Chung, Tyler Zhu, Max Gonzalez Saez-Diez, Juan Carlos Niebles, Honglu Zhou, and Olga Russakovsky. 2025. Unifying specialized visual encoders for video language models. *Preprint*, arXiv:2501.01426.

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.

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.

Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. 2021. Video swin transformer. *arXiv preprint arXiv:2106.13230*.

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.

Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. 2021. CLIP4Clip: An empirical study of clip for end to end video clip retrieval. *arXiv preprint arXiv:2104.08860*.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. 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.

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.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2024a. Video-chatgpt: Towards detailed video understanding via large vision and language models. *Preprint*, arXiv:2306.05424.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2024b. Videogpt+: Integrating image and video encoders for enhanced video understanding. *arxiv*.

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.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. 2024. Dinov2: Learning robust visual features without supervision. *Preprint*, arXiv:2304.07193.

Binnu Paul, Dwijen Rudrapal, Kunal Chakma, and Anupam Jamatia. 2024. Multimodal machine translation approaches for indian languages: A comprehensive survey. *JUCS - Journal of Universal Computer Science*, 30:694–717.

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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. *Preprint*, arXiv:1508.07909.

Huangjun Shen, Liangying Shao, Wenbo Li, Zhibin Lan, Zhanyu Liu, and Jinsong Su. 2024. A survey on multi-modal machine translation: Tasks, methods and challenges. *Preprint*, arXiv:2405.12669.

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.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 3104–3112, Cambridge, MA, USA. MIT Press.

ZhenHao Tang, XiaoBing Zhang, Zi Long, and XiangHua Fu. 2022a. 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.

ZhenHao Tang, XiaoBing Zhang, Zi Long, and XiangHua Fu. 2022b. 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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. *Preprint*, arXiv:1706.03762.

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.

Minghao Wu, Jiahao Xu, Yulin Yuan, Gholamreza Haffari, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2025. (perhaps) beyond human translation: Harnessing multi-agent collaboration for translating ultra-long literary texts. *Preprint*, arXiv:2405.11804.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *Preprint*, arXiv:1609.08144.

Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. 2021. Video-clip: Contrastive pre-training for zero-shot video-text understanding. *Preprint*, arXiv:2109.14084.

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.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. *Preprint*, arXiv:2303.15343.

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*.

Long Zhao, Nitesh B. Gundavarapu, Liangzhe Yuan, Hao Zhou, Shen Yan, Jennifer J. Sun, Luke Friedman, Rui Qian, Tobias Weyand, Yue Zhao, Rachel Hornung, Florian Schroff, Ming-Hsuan Yang, David A. Ross, Huisheng Wang, Hartwig Adam, Mikhail Sirotenko, Ting Liu, and Boqing Gong. 2024. Videoprism: A foundational visual encoder for video understanding. *Preprint*, arXiv:2402.13217.