Video-Language Understanding: A Survey from Model Architecture, Model Training, and Data Perspectives

Anonymous ACL submission

Abstract

Humans use multiple senses to comprehend the environment. Vision and language are two of the most vital senses since they allow us to easily communicate our thoughts and perceive the world around us. There has been a lot of interest in creating video-language understanding systems with human-like senses since a video-language pair can mimic both our linguistic medium and visual environment with temporal dynamics. In this survey, we review the key tasks of these systems and highlight the associated challenges. Based on the challenges, we summarize their methods from model architecture, model training, and data perspectives. We also conduct performance comparison among the methods, and discuss promising directions for future research.

1 Introduction

Vision and language constitute fundamental components of our perception: vision allows us to perceive the physical world, while language enables us to describe and converse about it. However, the world is not merely a static image but exhibits dynamics in which objects move and interact across time. With the temporal dimension, videos are able to capture such temporal dynamics that characterize the physical world. Consequently, in pursuit of endowing artificial intelligence with human-like perceptual abilities, researchers have been developing Video-Language Understanding models that are capable of interpreting the spatio-temporal dynamics of videos and the semantics of language, dating back to the 1970s (Lazarus, 1973; McGurk and MacDonald, 1976). These models are distinctive from image-language understanding models, since they exhibit an additional ability to interpret the temporal dynamics (Li et al., 2020).

They have demonstrated impressive performance in various video-language understanding tasks. These tasks evaluate video-language models from coarse-grained to fine-grained understanding capacity. For example, for coarse-grained understanding, text-video retrieval task assesses the model's ability to holistically associate a language query with a whole video (Han et al., 2023). For more fine-grained understanding capacity, a video captioning model is required to understand the overall and detailed video content, then describe the content in concise language (Abdar et al., 2023). Fine-grained understanding in video questioning answering remains a difficult task, where a model needs to recognize minute visual objects or actions, and infers their semantic, spatial, temporal, and causal relationships (Xiao et al., 2021).

In order to effectively perform such video-language understanding tasks, there are three challenges that video-language understanding works have to explore. The first challenge lies in devising an appropriate neural architecture to model the interaction between video and language modalities. The second challenge is to design an effective strategy to train video-language understanding models in order to effectively adapt to multiple target tasks and domains. The third challenge is preparing high-quality video-language data that fuel the training of these models.

Although a handful of recent works have tried to review video-language understanding, they mostly focus on one challenge, for example, Transformer-based architecture (Ruan and Jin, 2022) (the 1st challenge), self-supervised learning (Schiappa et al., 2023) and pre-training (Cheng et al., 2023) (the 2nd challenge), and data augmentation (Zhou et al., 2024) (the 3rd challenge). Moreover, others also focus merely on one video-language understanding task, *e.g.* video question answering (Zhong et al., 2022), text-video retrieval (Zhu et al., 2023), and video captioning (Abdar et al., 2023). Such a narrow focus contradicts the growing consensus advocating for the development of artificial general intelligence capable of versatile adaptation

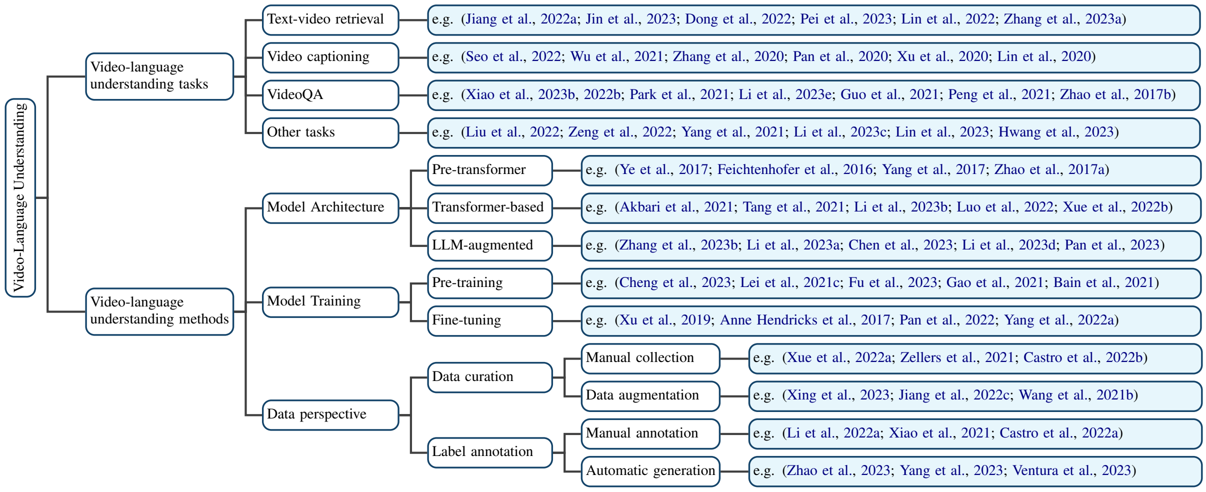

Figure 1: Taxonomy of Video-language Understanding

to a range of tasks and domains. Consider a human interaction scenario where an individual iteratively poses questions about a video, searches for a pertinent moment, and requests a summary. Such use case necessitates a broad capability to comprehend video and language content, without being bounded by a certain task. In addition, the development of a video-language understanding system often involves a multi-step process encompassing designing a model architecture, formulating a training method, and preparing data, rather than being a singular-step endeavor. Hence, this paper aims to present a more comprehensive and meaningful survey to connect the aspects of video-language understanding. Our contributions are as follows:

- We summarize the key tasks of video-language understanding and discuss their common challenges: intra-modal and cross-modal interaction, cross-domain adaptation, and data preparation.
- We provide a clear taxonomy to review video-language understanding works from three perspectives according to the three aforementioned challenges: (1) *Model architecture perspective:* we classify existing works into Pre-transformer, Transformer-based, and LLM-augmented architectures to model video-language relationship. In the latter category, we discuss recent efforts that utilize the advantages of LLMs to enhance video-language understanding. (2) *Model training perspective:* we categorize the training methods into Pre-training and Fine-tuning to adapt video-language representations to the target downstream task. (3) *Data perspective:* we also summarize existing approaches that curate video-language data and annotate them to fuel the training of video-language understanding models.
- Finally, we provide our prospects and propose potential directions for future research.

2 Video-Language Tasks

There exists a wide range of tasks that demand video-language understanding capacity. We illustrate typical examples of them in Figure 2.

2.1 Text-video retrieval

Text-video retrieval is the task to search for the corresponding video given a language query (text-to-video), or oppositely search for the language description given a video (video-to-text). At the moment, due to the popularity of social media platforms such as YouTube, Bilibili, and Netflix, where users want to find videos that suit their needs, there are more research works that concentrate on text-to-video retrieval than the video-to-text setting. The main evaluation metrics for text-video retrieval are recall at rank N ($R@N$), median rank ($MedR$), and mean rank (MnR) (Luo et al., 2022; Xue et al., 2022b). They assume a one-to-one correspondence between a pair of video and text. However, in practice, there might exist one-to-many matches for a query, to which these evaluation metrics may be unable to adapt (Fang et al., 2023a). Instead of extracting a complete video, there exists a variant of text-video retrieval, *i.e.* video moment retrieval which requires more fine-grained video-language understanding to extract relevant video moments for a textual query within a single video.

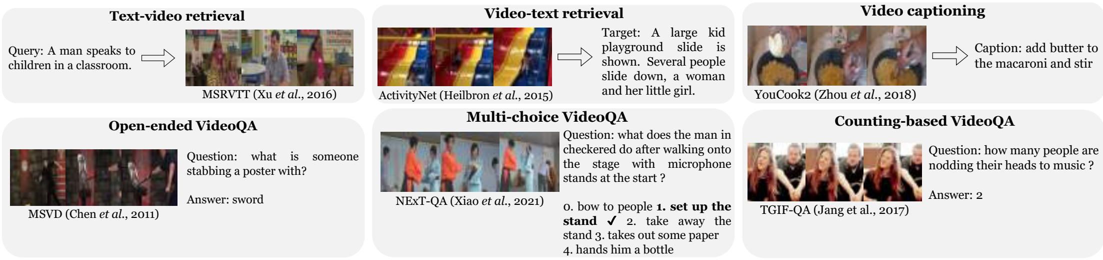

Figure 2: Illustration of video-language understanding tasks. For more examples, we refer reader to Appendix A.

2.2 Video captioning

Video captioning is the task to generate a concise language description for a video. A video captioning model receives as input a video and optionally a language transcript transcribed from the audio in the video. Typically, a model produces a sentence-level caption for the whole video. Krishna et al. (2017), Zhou et al. (2018b), and Yang et al. (2023) investigate generating a sentence caption or a title for each video segment in *dense video captioning* and *video chapter generation*. Moreover, Yu et al. (2021) also explores generating a paragraph-level caption to summarize a video in *multimodal abstractive summarization*.

2.3 Video question answering (videoQA)

Video question answering is the task to predict the correct answer based on a question q and a video v. There are two fundamental types of VideoQA, *i.e.* **multi-choice** VideoQA and **open-ended** VideoQA. In multi-choice VideoQA, a model is presented with a certain number of candidate answers and it will choose the correct answer among them. Open-ended VideoQA can be formulated as a classification problem, a generation problem, or a regression problem. Classification-based VideoQA associates a video-question pair with an answer from a pre-defined vocabulary set. Generation-based VideoQA is not restricted to a vocabulary set, in which a model can generate a sequence of tokens that represent the answer to a question. Regression-based VideoQA is often used for counting questions, *e.g.* counting the repetitions of an action or counting the number of an object in a video.

2.4 Connections among video-language understanding tasks

Apart from these three most popular groups of video-language understanding tasks, there are other tasks that have been widely studied in the literature such as action recognition, referring video object segmentation, etc. Although one may argue that these video-language understanding tasks possess distinct natures, research work has found that one foundation model can effectively tackle many of them (Wang et al., 2022). Li et al. (2023b) even unify text-video retrieval, video captioning, and videoQA as a single masked language modeling task and use the same set of parameter values to perform all of them. Additionally, Seo et al. (2022) find that a model designed for video captioning can effectively adapt to text-video retrieval, videoQA, and action recognition. Based on these works, we believe that even though different tasks exhibit different challenges due to their specific nature, their challenges can be summarized into common challenges of video-language understanding.

3 Challenges of Video-Language Understanding

Video-language understanding presents unique challenges compared with image-language understanding, since a video incorporates an additional temporal channel. We summarize important challenges of video-language understanding as follows:

Intra-modal and cross-modal interaction. While intra-modal interaction modeling within language can be directly taken from image-language understanding, intra-modal interaction modeling within video is different since it jointly consists of spatial interaction and temporal interaction. Spatial interaction delves into the relationships among pixels, patches, regions, or objects within an individual frame, whereas temporal interaction captures sequential dependencies among video frames or video segments. Longer video durations amplify the complexity of temporal interaction modeling by necessitating the recognition of a higher number of objects and events (Yu et al., 2020) and also computational demand to process more video frames (Lin et al., 2022). Particular video domains, such as egocentric videos, also complicate temporal interaction modeling, as objects undergo drastic appearance and disappearance dynamics over time, posing challenges in capturing their relationships (Bansal et al., 2022; Tang et al., 2023).

Given the larger semantic gap for video-language compared to image-language, cross-modal interaction plays a crucial role in video-language understanding. The interaction between visual and language features is pivotal for aligning the semantics of video and text query to associate them for text-video retrieval, or identifying relevant parts to answer the question and writing the caption in videoQA and video captioning, respectively. In addition, incorporating the interaction of motion and language features can mitigate the extraction of noisy information from videos (Ding et al., 2022). Lin et al. (2022) also discover that the interaction between audio and language features can compactly capture information related to objects, actions, and complex events, compensating for sparsely extracted video frames.

Cross-domain adaptation. Given the infinitude of online videos, that our video-language understanding model will encounter testing scenarios which are identically distributed to our training data is an impractical assumption. Moreover, with the advent of LLM-augmented models that can tackle a variety video-language understanding tasks (Li et al., 2023a,d), it is currently more advisable to train a model that can effectively adapt to multiple tasks and domains than to obtain a model which specializes in a specific understanding task. Furthermore, since a video can be considered as a sequence of images, training a model on video-text data is more computationally expensive than image-text data. Combined with the large-scale of recent video-language understanding models (Jiang et al., 2022a; Yang et al., 2022a), there is also a need to devise an efficient fine-tuning strategy to save the computational cost of fine-tuning these models.

Data preparation. Although Lei et al. (2021c) only use image-text data to train models for video-language understanding tasks, in essence, video-text data are crucial for the effectiveness of these models. In particular, compared with a static image, a video offers richer information with diverse spatial semantics with consistent temporal dynamics (Zhuang et al., 2023). As such, Cheng et al. (2023) find that training on videos outperforms training on images, but jointly training on both data achieves the best performance. As additional evidence, Yuan et al. (2023) shows that video-pretrained models outperform image-pretrained models in classifying motion-rich videos. However, video-text data takes up more storage cost than image-text data since a video comprises multiple images as video frames. Moreover, annotating a video is also more time-consuming and labor-intensive than annotating an image (Xing et al., 2023). Therefore, video-language understanding models have been limited by the small size of clean paired video-text corpora in contrast to billion-scale image-text datasets (Zhao et al., 2023). Various efforts (Zhao et al., 2023; Xing et al., 2023) have been put into devising efficient and economical methods to curate and label video-text data.

4 Model Architecture for Video-Language Understanding

Effective modeling intra-modal and cross-modal interaction is the key aim in designing video-language understanding model architectures, which can be divided into **Pre-transformer** and **Transformer-based architectures**. The advent of LLMs with remarkable zero-shot capability in addressing multiple tasks led to the design of **LLM-augmented architectures** that exhibit cross-domain adaptation ability to various video-language understanding tasks.

4.1 Pre-transformer architecture

Pre-transformer architectures typically comprise unimodal video and language encoders for implementing intra-modal interactions and cross-modal encoders for cross-modal interactions.

Unimodal encoders. A video encoder often encodes raw videos by extracting frame appearance and clip motion features as spatial and temporal representations, respectively. As each video frame can be considered as a single image, various works have utilized CNNs to extract spatial representations (Simonyan and Zisserman, 2014; Feichtenhofer et al., 2016; Zhao et al., 2017b). For temporal representations, the sequential nature of RNN makes it a popular choice in pre-transformer architectures (Yang et al., 2017; Zhao et al., 2017a; Venugopalan et al., 2015). Furthermore, 3D CNNs with an additional temporal channel inserted to 2D CNN have also demonstrated effectiveness in extracting spatio-temporal representations (Tran et al., 2017; Carreira and Zisserman, 2017). In addition to CNN and RNN, Chen et al. (2018), Gay et al. (2019), and Wei et al. (2017) also build graphs to incorporate intra-modal relationships among video

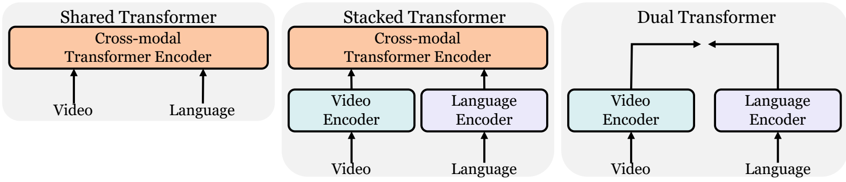

Figure 3: Illustration of video-language Transformer-based architectures.

entities such as video segments or visual objects. These graph-structured works emphasize the reasoning ability of the model architecture.

A common framework of language encoder is to extract pre-trained word embeddings such as word2vec (Kaufman et al., 2016; Yu et al., 2017) or GloVe (Torabi et al., 2016; Kiros et al., 2014), then proceed with RNN-based modules such as LSTM or GRU. Such framework is taken from language model architectures before the era of Transformer.

Cross-modal encoders. Gao et al. (2017) and Zeng et al. (2017) apply element-wise multiplication to fuse the global video and question representations for video question answering. It demonstrates the advantage of a simple operation for video-language fusion. Attention has also been used to model video-language relations, in order to identify salient parts in video and language sentence (Yuan et al., 2019), or to refine the representation of the video based on the language question (Xu et al., 2017). Pre-transformer video-language works have also combined attention with a wide variety of techniques, including hierarchical learning (Baraldi et al., 2017), multi-faceted representation (Long et al., 2018), memory networks (Fan et al., 2019), and graph networks (Xiao et al., 2022a).

4.2 Transformer-based architecture

Developed based on the self-attention mechanism, which exhaustively correlates every pair of input tokens with each other, Transformer-based architecture has the capacity to capture long-term dependencies and learn from web-scale data. It has demonstrated remarkable performance in many video-language tasks. Similar to the pre-transformer architecture, the Transformer-based framework also comprises unimodal encoders and cross-modal encoders to model intra-modal and cross-modal interactions, respectively. For unimodal encoders, several works find vision transformer for video encoding and BERT encoder for language encoding outperform RNN- and CNN-based encoding (Fu et al., 2021; Bain et al., 2021; Seo et al., 2022). We then summarize fundamental types of Transformer-based architectures and illustrate them in Figure 3.

Shared Transformer. Motivated by the success of Transformer in language modeling (Devlin et al., 2018), Akbari et al. (2021) and Wang et al. (2023a) construct a shared Transformer encoder for video-language understanding. Their encoder architectures receive the concatenation of visual patches and language tokens, then jointly calculate their interactions in a BERT-based manner. Akbari et al. (2021) additionally incorporate modality embeddings which comprise three values to denote three kinds of input modalities, *i.e.* (video, audio, text).

Stacked Transformer. Li et al. (2020) reveals that a shared Transformer encoder is weak in modeling temporal relations between videos and texts. To address this problem, they introduce a stacked Transformer architecture, with a hierarchical stack consisting of unimodal encoders to encode video and language inputs separately, and then a cross-modal Transformer to compute video-language interactions. A multitude of video-language understanding works follow such design to stack a cross-modal Transformer-based encoder above unimodal encoders (Fu et al., 2023; Li et al., 2023b; Lei et al., 2021c; Luo et al., 2022; Nie et al., 2022). To perform video captioning, Seo et al. (2022) and Luo et al. (2020) further insert a causal Transformer-based decoder that generates language tokens based on the encoded cross-modal representations.

Dual Transformer. Dual Transformer architectures have been favored for text-video retrieval (Luo et al., 2022; Bain et al., 2021, 2022; Lin et al., 2022; Xue et al., 2022b). These architectures use two Transformer encoders to encode video and language separately, yielding global representations for each input modality, then applying simple operations such as cosine similarity to compute cross-modal interaction. Such a separate encoding scheme enables them to mitigate the computational cost of computing pairwise interactions between every pair of video and language inputs. They have accomplished not only efficiency but also effectiveness in text-video retrieval problems.

4.3 LLM-augmented architecture

Large language models (LLMs) have achieved impressive results in simultaneously tackling multiple NLP tasks. Recent efforts have sought to apply LLMs for video-language understanding to extend its cross-domain adaptation ability to video-language settings (Chen et al., 2023; Li et al., 2023a). These efforts can be categorized into two approaches. The first approach employs LLM as a controller and video-language understanding models as helping tools. The controller will call the specific tool according to the language input instruction. The second approach utilizes LLM as the output generator and seeks to align video pretrained models to the LLM. For video-language understanding, since the second approach dominates the first one with a long list of recent works (Chen et al., 2023; Li et al., 2023a; Chen et al., 2023; Li et al., 2023d; Zhang et al., 2023b; Maaz et al., 2023), we review them as follows:

LLM as the output generator. The framework comprises a visual encoder, a semantic translator, and an LLM as the output generator. Regarding visual encoder, LLM-augmented architectures often use vision transformer and CNN models of the pre-Transformer and Transformer-based architectures (Chen et al., 2023). Since an LLM has never seen a video during its training, a semantic translator is needed to translate the visual semantics of a video to the LLM's semantics. For the translator, Video-LLaMA (Zhang et al., 2023b) and VideoChat (Li et al., 2023a) implement a Q-Former as a Transformer-based module that uses a sequence of query embeddings that interact with visual features of the video to extract informative video information. Instead of Q-Former, VideoLLM (Chen et al., 2023), Video-ChatGPT (Maaz et al., 2023), and LLaMA-Vid (Li et al., 2023d) find that a simple linear projection that projects visual features into the LLM's input dimension can achieve effective performance. Subsequently, these visual-based query embeddings or projected visual features are combined with the language instruction to become the input fed to the LLM to produce the final output.

4.4 Performance analysis

Among the Transformer-based architectures, dual Transformer is the most effective for the text-video retrieval task, as it excels at associating holistic language and video semantics. On the other hand, stacked Transformer architecture can deftly calculate intra-modal and inter-modal interactions with specialized unimodal and cross-modal encoders. Thus, it can extract meaningful video information with respect to the question for videoQA, and relate the currently generated language tokens to the video content for video captioning. Interestingly, recent LLM-augmented models significantly outperform the Transformer-based ones, proving themselves a promising architecture for video-language understanding. Due to the page limit, we defer our tables for performance comparison to Appendix B.

5 Model Training for Video-Language Understanding

5.1 Pre-training for Video-Language Understanding

Pre-trained language models have established outstanding performance in a broad range of NLP tasks. These models are trained upon a large corpus of text to gain valuable world knowledge that can be applied to multiple downstream tasks. Similar ideas have been adopted for video-language understanding. Various pre-training strategies have been devised to help a video-language understanding model obtain video and language contextual knowledge. We summarize them into three groups:

Language-based pre-training. The most popular language-based pre-training task is masked language modeling (MLM) (Lei et al., 2021c; Sun et al., 2019; Cheng et al., 2023), which randomly masks a portion of words in the language input and trains the model to predict the masked words based on unmasked language words and video entities. Instead of masking a portion of words, UniVL (Luo et al., 2020) and VICTOR (Lei et al., 2021a) discover that masking the whole language modality benefits video captioning task. MLM can be combined with other language-based pre-training task, *e.g.* masked sentence order modeling which is to classify the original order of the shuffled language sentences (Lei et al., 2021a).

Video-based pre-training. Video-based pre-training tasks help video-language models capture contextual information in the video modality. As a counterpart of MLM, masked video modeling (MVM) trains the model to predict the portion of masked video entities based upon the unmasked entities and language words. The continuous nature of videos leads to different choices of video entities, such as frame patches (Li et al., 2020) or video frames (Fu et al., 2021). In terms of the training objective, Li et al. (2020) use L2 regression loss to train the model to predict pre-trained features of the masked video frames extracted by ResNet and SlowFast models, while Fu et al. (2021) use cross-entropy loss to train the model to predict the masked visual tokens, which are quantized by a variational autoencoder from visual frame patches.

Video-text pre-training. Video-text pre-training is crucial for a model to capture video-language relation. Xue et al. (2022b), Gao et al. (2021), and Bain et al. (2021) utilize a framework of video-text contrastive learning to produce close representations for semantically similar video and language inputs. These works focus on creating a joint semantic space that aligns separate representations of video and language. Instead of separate representations, Tang et al. (2021), Fu et al. (2021), and Li et al. (2023b) enable video and textual representations to interact with each other and use a single token to represent the cross-modal input, which is forwarded to predict whether the video-text pair is matched or not. In these two pre-training frameworks, not only video-text data but also image-text data are utilized during pre-training, in which an image is considered as a video with a single frame.

Video-text contrastive learning has revealed promising results for text-video retrieval (Lin et al., 2022; Gao et al., 2021; Xue et al., 2022b). MLM has contributed to enhancing VideoQA since the task resembles MLM in predicting the language word given a video-language pair (the question is the language input in videoQA). Compared to these pre-training strategies, MVM does provide performance gain for video-language understanding but its gain is less significant (Cheng et al., 2023).

5.2 Fine-tuning for Video-Language Understanding

Task-specific fine-tuning is commonly used by pre-Transformer architectures to train from scratch since these models do not have sufficient parameter capacity to learn generalizable features through pre-training. It is also widely adopted by Transformer-based architectures to improve the performance for a specific downstream task. Moreover, LLM-augmented architectures also utilize instruction tuning as a variant of fine-tuning, to adapt from the visual and audio spaces to the LLM language space.

Fine-tuning strategies. Normally, all of the model parameters are updated during fine-tuning (Gao et al., 2017; Xu et al., 2019; Anne Hendricks et al., 2017). However, in cases computational resources or training data are limited, only adaptation layers such as low-rank adapters (Pan et al., 2022; Yang et al., 2022a) or learnable prompt vectors (Ju et al., 2022) are fine-tuned to reduce training cost or prevent overfitting. Such risks also apply for LLM-augmented architectures discussed in Section 4.3, since LLMs exhibit a billion scale of parameters, thus incurring excessively huge cost if full fine-tuning is conducted. For such models, Zhang et al. (2023b) and Li et al. (2023d) design a two-stage instruction tuning strategy which only fine-tunes the semantic translator. The first stage trains the model to generate the textual description based on the combined video and the language instruction, in order to align visual representations extracted by the visual encoder with the language space of LLM. The second stage is often performed on small-scale video-text pairs manually collected by the authors to further tailor the output features of the translator towards the target domains.

6 Data Perspective for Video-Language Understanding

6.1 Data curation

Manual collection. To construct video-language datasets, multiple works search for publicly available videos on the internet, which exhibit a wide diversity of content. As such, video-language datasets with online videos are mostly aimed for the purpose of pre-training models to learn generalizable knowledge. Instead of online videos, to collect videos satisfying a specific requirement, Xiao et al. (2021) inherit 6,000 videos from the video relation dataset VidOR since they want videos that describe scenes in daily life. Analogously, Causal-VidQA dataset (Li et al., 2022a) inherits 546,882 videos from the Kinetics-700 dataset, and FIBER dataset (Castro et al., 2022b) uses 41,250 video clips of the VaTeX dataset. Apart from making use of existing datasets, Goyal et al. (2017) and Damen et al. (2022) request human annotators to record videos by themselves.

Data augmentation. Rather than manually collecting videos from external sources, Xing et al. (2023) and Jiang et al. (2022c) explore data augmentation techniques which are particularly designed for videos. In detail, their TubeTokenMix mixes two videos in which the mixing coefficient is defined upon the temporal dimension, and their temporal shift randomly shifts video frame features backward or forward over the temporal dimension. These techniques outperform standard augmentation approaches for image data, such as CutMix (Yun et al., 2019), Mixup (Zhang et al., 2017), and PixMix (Hendrycks et al., 2022).

6.2 Label annotation

Manual annotation. Several works (Li et al., 2022a; Lei et al., 2021b; Xiao et al., 2021) use human annotators since they provide high-quality labels. However, such approach is expensive, particularly when dealing with video data. For example, annotating QVHighlights dataset (Lei et al., 2021b) costs approximately $16,000 for 10K videos and 3 months to complete. Similarly, NExT-QA (Xiao et al., 2021) needs 100 undergraduate students and 1 year to annotate only 5K videos.

Automatic generation. Directly taking language transcripts of YouTube videos as textual labels could reduce annotation cost (Miech et al., 2019; Xue et al., 2022a; Zellers et al., 2021). However, these labels have been shown to be grammatically incorrect and temporally misalign with the video content (Tang et al., 2021). Motivated by the success of LLMs, Zhao et al. (2023) train a system consisting of a TimeSformer-L visual encoder and a GPT-2XL decoder to write dense captions for videos. Moreover, Li et al. (2023a) use GPT-4 to generate summaries for movie synopses.

7 Future Directions

Fine-grained video-language understanding. Existing methods excel at performing video-language understanding at a coarse-grained level. Thus, answering questions like "*what is*" or recognizing a global event is no longer a difficult problem (Xiao et al., 2021). Nevertheless, stopping at the coarse-grained understanding level can restrict practical applications of current systems. In practice, a user might search for the specific timestamp and the position of an object within a video (Jiang et al., 2022b). Moreover, he or she may ask the AI agent to predict alternative events, which is typical in predictive applications (Xiao et al., 2021; Li et al., 2022a). These circumstances require fine-grained understanding and inference ability about causal and temporal relationships within a video. Future research in this direction is needed to promote progress towards the core of human intelligence.

Long-form video-language understanding. Current video-language understanding systems have been trained exclusively upon short video clips (5-15 seconds in length) (Lin et al., 2022). Consequently, they struggle with real-world videos which may last several minutes or hours. The reasons that models are mostly trained on short video clips are two-fold: 1) training on long-range videos demands huge computational cost to process a high number of video frames, 2) many benchmarks contain spatial bias that enables a model to determine the answer based on short-term video cues (Lei et al., 2022). To address the first issue, existing work has sought to train a model on an additional modality while maintaining the number of their input frames in long videos (Lin et al., 2022). For the second issue, Mangalam et al. (2023) introduce a benchmark of authentically long-term video-language understanding. However, feeding a model with additional information may introduce noise and Mangalam et al. (2023)'s benchmark is restricted to the egocentric domain. Consequently, designing an efficient training framework for the model to capture spatial, temporal, and causal relationships in long videos deserves more attention.

Trustworthiness of video-language understanding models. Although modern video-language understanding systems have demonstrated remarkable performance, their black-box nature undermines our trust to deploy them. In particular, we still do not precisely understand what part of the video a videoQA model looks at to answer the question (Li et al., 2022b), or how video and language semantic information flows into the common representation space of the video retrieval model (Jia et al., 2022). Furthermore, adversarial noise sensitivity or hallucination of video-language understanding models are also open problems. Future trustworthiness benchmarks such as (Xiao et al., 2023a; Wang et al., 2021a) for video-language understanding are of great significance towards practical systems.

8 Conclusion

In this paper, we survey the broad research field of video-language understanding. Particularly, we categorize related video-language understanding tasks and discuss meaningful insights from model architecture, model training, and data perspectives. Moreover, we analyze performances of different video-language understanding methods, and finally conclude with promising future directions. We hope our survey can foster more research towards constructing effective AI systems that can comprehensively understand dynamic visual world and meaningfully interact with humans.

9 Limitations

Although we have sought to comprehensively analyze the literature of video-language understanding, we might not fully cover all of the tasks, model architectures, model training, and data perspectives. Therefore, we complement the survey with a repository[1]. The repository comprises the latest video-language understanding papers, datasets, and their open-source implementations. We will periodically update the repository to trace the progress of the latest research.

References

Moloud Abdar, Meenakshi Kollati, Swaraja Kuraparthi, Farhad Pourpanah, Daniel McDuff, Mohammad Ghavamzadeh, Shuicheng Yan, Abduallah Mohamed, Abbas Khosravi, Erik Cambria, et al. 2023. A review of deep learning for video captioning. *arXiv preprint arXiv:2304.11431*.

Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. 2016. YouTube-8M: A Large-Scale Video Classification Benchmark. *arXiv preprint arXiv:1609.08675*.

Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. 2021. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems*, 34:24206–24221.

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812.

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. 2021. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2022. A clip-hitchhiker's guide to long video retrieval. *arXiv preprint arXiv:2205.08508*.

[1]Due to the double-blind review, the repository can be found at https://anonymous.4open.science/r/survey-video-language-understanding, or in the submitted software package.

Siddhant Bansal, Chetan Arora, and CV Jawahar. 2022. My view is the best view: Procedure learning from egocentric videos. In *European Conference on Computer Vision*, pages 657–675. Springer.

Lorenzo Baraldi, Costantino Grana, and Rita Cucchiara. 2017. Hierarchical boundary-aware neural encoder for video captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1657–1666.

Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. 2018. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*.

Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. 2019. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*.

Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308.

Santiago Castro, Naihao Deng, Pingxuan Huang, Mihai Burzo, and Rada Mihalcea. 2022a. In-the-wild video question answering. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5613–5635.

Santiago Castro, Ruoyao Wang, Pingxuan Huang, Ian Stewart, Oana Ignat, Nan Liu, Jonathan Stroud, and Rada Mihalcea. 2022b. Fiber: Fill-in-the-blanks as a challenging video understanding evaluation framework. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2925–2940.

David Chen and William B Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 190–200.

Guo Chen, Yin-Dong Zheng, Jiahao Wang, Jilan Xu, Yifei Huang, Junting Pan, Yi Wang, Yali Wang, Yu Qiao, Tong Lu, et al. 2023. Videollm: Modeling video sequence with large language models. *arXiv preprint arXiv:2305.13292*.

Yuting Chen, Joseph Wang, Yannan Bai, Gregory Castañón, and Venkatesh Saligrama. 2018. Probabilistic semantic retrieval for surveillance videos with activity graphs. *IEEE Transactions on Multimedia*, 21(3):704–716.

Feng Cheng, Xizi Wang, Jie Lei, David Crandall, Mohit Bansal, and Gedas Bertasius. 2023. Vindlu: A recipe for effective video-and-language pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10739–10750.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. 2022. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, pages 1–23.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Zihan Ding, Tianrui Hui, Junshi Huang, Xiaoming Wei, Jizhong Han, and Si Liu. 2022. Language-bridged spatial-temporal interaction for referring video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4964–4973.

Jianfeng Dong, Yabing Wang, Xianke Chen, Xiaoye Qu, Xirong Li, Yuan He, and Xun Wang. 2022. Reading-strategy inspired visual representation learning for text-to-video retrieval. *IEEE transactions on circuits and systems for video technology*, 32(8):5680–5694.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Chenyou Fan, Xiaofan Zhang, Shu Zhang, et al. 2019. Heterogeneous memory enhanced multimodal attention model for video question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1999–2007.

Bo Fang, Chang Liu, Yu Zhou, Min Yang, Yuxin Song, Fu Li, Weiping Wang, Xiangyang Ji, Wanli Ouyang, et al. 2023a. Uatvr: Uncertainty-adaptive text-video retrieval. *arXiv preprint arXiv:2301.06309*.

Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. 2023b. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19358–19369.

Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. 2016. Convolutional Two-Stream Network Fusion for Video Action Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. 2021. Violet: End-to-end video-language transformers with masked visual-token modeling. *arXiv preprint arXiv:2111.12681*.

Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. 2023. An empirical study of end-to-end video-language transformers with masked visual modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22898–22909.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pages 5267–5275.

Zijian Gao, Jingyu Liu, Weiqi Sun, Sheng Chen, Dedan Chang, and Lili Zhao. 2021. Clip2tv: Align, match and distill for video-text retrieval. *arXiv preprint arXiv:2111.05610*.

Paul Gay, James Stuart, and Alessio Del Bue. 2019. Visual graphs from motion (vgfm): Scene understanding with object geometry reasoning. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14*, pages 330–346. Springer.

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. 2017. The" Something Something" Video Database for Learning and Evaluating Visual Common Sense. In *The IEEE International Conference on Computer Vision (ICCV)*.

Zhicheng Guo, Jiaxuan Zhao, Licheng Jiao, Xu Liu, and Lingling Li. 2021. Multi-scale progressive attention network for video question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 973–978.

Ning Han, Yawen Zeng, Chuhao Shi, Guangyi Xiao, Hao Chen, and Jingjing Chen. 2023. Bic-net: Learning efficient spatio-temporal relation for text-video retrieval. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(3):1–21.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xingjian He, Sihan Chen, Fan Ma, Zhicheng Huang, Xiaojie Jin, Zikang Liu, Dongmei Fu, Yi Yang, Jing Liu, and Jiashi Feng. 2023. Vlab: Enhancing video language pre-training by feature adapting and blending. *arXiv preprint arXiv:2305.13167*.

Dan Hendrycks, Andy Zou, Mantas Mazeika, Leonard Tang, Bo Li, Dawn Song, and Jacob Steinhardt. 2022. Pixmix: Dreamlike pictures comprehensively improve safety measures. In *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16783–16792.

Gabriel Huang, Bo Pang, Zhenhai Zhu, Clara Rivera, and Radu Soricut. 2020. Multimodal pretraining for dense video captioning. *arXiv preprint arXiv:2011.11760*.

Minyoung Hwang, Jaeyeon Jeong, Minsoo Kim, Yoonseon Oh, and Songhwai Oh. 2023. Meta-explore: Exploratory hierarchical vision-and-language navigation using scene object spectrum grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6683–6693.

Yunseok Jang, Yale Song, Chris Dongjoo Kim, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 2019. Video question answering with spatio-temporal reasoning. *International Journal of Computer Vision*, 127(10):1385–1412.

Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 2017. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2758–2766.

Mohan Jia, Zhongjian Dai, Yaping Dai, and Zhiyang Jia. 2022. An adversarial video moment retrieval algorithm. In *2022 41st Chinese Control Conference (CCC)*, pages 6689–6694. IEEE.

Haojun Jiang, Jianke Zhang, Rui Huang, Chunjiang Ge, Zanlin Ni, Jiwen Lu, Jie Zhou, Shiji Song, and Gao Huang. 2022a. Cross-modal adapter for text-video retrieval. *arXiv preprint arXiv:2211.09623*.

Ji Jiang, Meng Cao, Tengtao Song, and Yuexian Zou. 2022b. Video referring expression comprehension via transformer with content-aware query. *arXiv preprint arXiv:2210.02953*.

Jianwen Jiang, Ziqiang Chen, Haojie Lin, Xibin Zhao, and Yue Gao. 2020. Divide and conquer: Question-guided spatio-temporal contextual attention for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11101–11108.

Pin Jiang and Yahong Han. 2020. Reasoning with heterogeneous graph alignment for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11109–11116.

Xun Jiang, Xing Xu, Jingran Zhang, Fumin Shen, Zuo Cao, and Heng Tao Shen. 2022c. Semi-supervised video paragraph grounding with contrastive encoder. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2466–2475.

Peng Jin, Hao Li, Zesen Cheng, Kehan Li, Xiangyang Ji, Chang Liu, Li Yuan, and Jie Chen. 2023. Diffusionret: Generative text-video retrieval with diffusion model. *arXiv preprint arXiv:2303.09867*.

Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. 2022. Prompting visual-language models for efficient video understanding. In *European Conference on Computer Vision*, pages 105–124. Springer.

Dotan Kaufman, Gil Levi, Tal Hassner, and Lior Wolf. 2016. Temporal tessellation for video annotation and summarization. *arXiv preprint arXiv:1612.06950*, 3.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. 2017. The Kinetics Human Action Video Dataset. *arXiv preprint arXiv:1705.06950*.

Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. 2014. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715.

Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. 2011. HMDB: A Large Video Database for Human Motion Recognition. In *The IEEE International Conference on Computer Vision (ICCV)*.

Arnold A Lazarus. 1973. Multimodal behavior therapy: Treating the "basic id". *The Journal of nervous and mental disease*, 156(6):404–411.

Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. 2020. Hierarchical conditional relation networks for video question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9972–9981.

Chenyi Lei, Shixian Luo, Yong Liu, Wanggui He, Jiamang Wang, Guoxin Wang, Haihong Tang, Chunyan Miao, and Houqiang Li. 2021a. Understanding chinese video and language via contrastive multimodal pre-training. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2567–2576.

Jie Lei, Tamara L Berg, and Mohit Bansal. 2021b. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems*, 34:11846–11858.

Jie Lei, Tamara L Berg, and Mohit Bansal. 2022. Revealing single frame bias for video-and-language learning. *arXiv preprint arXiv:2206.03428*.

Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. 2021c. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7331–7341.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. 2018. Tvqa: Localized, compositional video question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1369–1379.

Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. 2020. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 447–463. Springer.

Jiangtong Li, Li Niu, and Liqing Zhang. 2022a. From representation to reasoning: Towards both evidence and commonsense reasoning for video question-answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21273–21282.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023a. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*.

Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. 2020. Hero: Hierarchical encoder for video+ language omni-representation pre-training. *arXiv preprint arXiv:2005.00200*.

Linjie Li, Zhe Gan, Kevin Lin, Chung-Ching Lin, Zicheng Liu, Ce Liu, and Lijuan Wang. 2023b. Lavender: Unifying video-language understanding as masked language modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23119–23129.

Pandeng Li, Chen-Wei Xie, Hongtao Xie, Liming Zhao, Lei Zhang, Yun Zheng, Deli Zhao, and Yongdong Zhang. 2023c. Momentdiff: Generative video moment retrieval from random to real. *arXiv preprint arXiv:2307.02869*.

Yanwei Li, Chengyao Wang, and Jiaya Jia. 2023d. Llama-vid: An image is worth 2 tokens in large language models. *arXiv preprint arXiv:2311.17043*.

Yicong Li, Xiang Wang, Junbin Xiao, and Tat-Seng Chua. 2022b. Equivariant and invariant grounding for video question answering. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4714–4722.

Yicong Li, Junbin Xiao, Chun Feng, Xiang Wang, and Tat-Seng Chua. 2023e. Discovering spatio-temporal rationales for video question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13869–13878.

Ke Lin, Zhuoxin Gan, and Liwei Wang. 2020. Semi-supervised learning for video captioning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1096–1106.

Kunyang Lin, Peihao Chen, Diwei Huang, Thomas H Li, Mingkui Tan, and Chuang Gan. 2023. Learning vision-and-language navigation from youtube videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8317–8326.

Yan-Bo Lin, Jie Lei, Mohit Bansal, and Gedas Bertasius. 2022. Eclipse: Efficient long-range video retrieval using sight and sound. In *European Conference on Computer Vision*, pages 413–430. Springer.

Ye Liu, Siyuan Li, Yang Wu, Chang-Wen Chen, Ying Shan, and Xiaohu Qie. 2022. Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3042–3051.

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

Xiang Long, Chuang Gan, and Gerard De Melo. 2018. Video captioning with multi-faceted attention. *Transactions of the Association for Computational Linguistics*, 6:173–184.

Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. 2020. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*.

Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. 2022. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2023. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. 2023. Egoschema: A diagnostic benchmark for very long-form video language understanding. *arXiv preprint arXiv:2308.09126*.

Harry McGurk and John MacDonald. 1976. Hearing lips and seeing voices. *Nature*, 264(5588):746–748.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2630–2640.

Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, et al. 2019. Moments in time dataset: one million videos for event understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(2):502–508.

Arsha Nagrani, Paul Hongsuck Seo, Bryan Seybold, Anja Hauth, Santiago Manen, Chen Sun, and Cordelia Schmid. 2022. Learning audio-video modalities from image captions. In *European Conference on Computer Vision*, pages 407–426. Springer.

Liqiang Nie, Leigang Qu, Dai Meng, Min Zhang, Qi Tian, and Alberto Del Bimbo. 2022. Search-oriented micro-video captioning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3234–3243.

Boxiao Pan, Haoye Cai, De-An Huang, Kuan-Hui Lee, Adrien Gaidon, Ehsan Adeli, and Juan Carlos Niebles. 2020. Spatio-temporal graph for video captioning with knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10870–10879.

Junting Pan, Ziyi Lin, Yuying Ge, Xiatian Zhu, Renrui Zhang, Yi Wang, Yu Qiao, and Hongsheng Li. 2023. Retrieving-to-answer: Zero-shot video question answering with frozen large language models. *arXiv preprint arXiv:2306.11732*.

Junting Pan, Ziyi Lin, Xiatian Zhu, Jing Shao, and Hongsheng Li. 2022. St-adapter: Parameter-efficient image-to-video transfer learning. *Advances in Neural Information Processing Systems*, 35:26462–26477.

Jungin Park, Jiyoung Lee, and Kwanghoon Sohn. 2021. Bridge to answer: Structure-aware graph interaction network for video question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15526–15535.

Renjing Pei, Jianzhuang Liu, Weimian Li, Bin Shao, Songcen Xu, Peng Dai, Juwei Lu, and Youliang Yan. 2023. Clipping: Distilling clip-based models with a student base for video-language retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18983–18992.

Liang Peng, Shuangji Yang, Yi Bin, and Guoqing Wang. 2021. Progressive graph attention network for video question answering. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2871–2879.

Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. 2013. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, 1:25–36.

Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. 2015. A dataset for movie description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3202–3212.

Marcus Rohrbach, Sikandar Amin, Mykhaylo Andriluka, and Bernt Schiele. 2012. A database for fine grained activity detection of cooking activities. In *2012 IEEE conference on computer vision and pattern recognition*, pages 1194–1201. IEEE.

Ludan Ruan and Qin Jin. 2022. Survey: Transformer based video-language pre-training. *AI Open*, 3:1–13.

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 preprint arXiv:1811.00347*.

Madeline C Schiappa, Yogesh S Rawat, and Mubarak Shah. 2023. Self-supervised learning for videos: A survey. *ACM Computing Surveys*, 55(13s):1–37.

Paul Hongsuck Seo, Arsha Nagrani, Anurag Arnab, and Cordelia Schmid. 2022. End-to-end generative pre-training for multimodal video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17959–17968.

Xindi Shang, Donglin Di, Junbin Xiao, Yu Cao, Xun Yang, and Tat-Seng Chua. 2019. Annotating objects and relations in user-generated videos. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, pages 279–287.

Karen Simonyan and Andrew Zisserman. 2014. Two-Stream Convolutional Networks for Action Recognition in Videos. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. 2015. Tvsum: Summarizing web videos using titles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5179–5187.

Jonathan C Stroud, Zhichao Lu, Chen Sun, Jia Deng, Rahul Sukthankar, Cordelia Schmid, and David A Ross. 2020. Learning video representations from textual web supervision. *arXiv preprint arXiv:2007.14937*.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7464–7473.

Min Sun, Ali Farhadi, and Steve Seitz. 2014. Ranking domain-specific highlights by analyzing edited videos. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 787–802. Springer.

Hao Tang, Kevin Liang, Kristen Grauman, Matt Feiszli, and Weiyao Wang. 2023. Egotracks: A long-term egocentric visual object tracking dataset. *arXiv preprint arXiv:2301.03213*.

Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. 2019. Coin: A large-scale dataset for comprehensive instructional video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1207–1216.

Zineng Tang, Jie Lei, and Mohit Bansal. 2021. Decembert: Learning from noisy instructional videos via dense captions and entropy minimization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2415–2426.

Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73.

Atousa Torabi, Niket Tandon, and Leonid Sigal. 2016. Learning language-visual embedding for movie understanding with natural-language. *arXiv preprint arXiv:1609.08124*.

Du Tran, Jamie Ray, Zheng Shou, Shih-Fu Chang, and Manohar Paluri. 2017. Convnet architecture search for spatiotemporal feature learning. *arXiv preprint arXiv:1708.05038*.

Lucas Ventura, Antoine Yang, Cordelia Schmid, and Gül Varol. 2023. Covr: Learning composed video retrieval from web video captions. *arXiv preprint arXiv:2308.14746*.

Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. 2015. Sequence to sequence-video to text. In *Proceedings of the IEEE international conference on computer vision*, pages 4534–4542.

Jinpeng Wang, Yixiao Ge, Rui Yan, Yuying Ge, Kevin Qinghong Lin, Satoshi Tsutsui, Xudong Lin, Guanyu Cai, Jianping Wu, Ying Shan, et al. 2023a. All in one: Exploring unified video-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6598–6608.

Junke Wang, Dongdong Chen, Zuxuan Wu, Chong Luo, Luowei Zhou, Yucheng Zhao, Yujia Xie, Ce Liu, Yu-Gang Jiang, and Lu Yuan. 2022. Omnivl: One foundation model for image-language and video-language tasks. *Advances in neural information processing systems*, 35:5696–5710.

Lijie Wang, Hao Liu, Shuyuan Peng, Hongxuan Tang, Xinyan Xiao, Ying Chen, Hua Wu, and Haifeng Wang. 2021a. Dutrust: A sentiment analysis dataset for trustworthiness evaluation. *arXiv preprint arXiv:2108.13140*.

Xiang Wang, Shiwei Zhang, Zhiwu Qing, Yuanjie Shao, Changxin Gao, and Nong Sang. 2021b. Self-supervised learning for semi-supervised temporal action proposal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1905–1914.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4581–4591.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, et al. 2023b. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*.

Lina Wei, Fangfang Wang, Xi Li, Fei Wu, and Jun Xiao. 2017. Graph-theoretic spatiotemporal context modeling for video saliency detection. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 4197–4201. IEEE.

Bofeng Wu, Guocheng Niu, Jun Yu, Xinyan Xiao, Jian Zhang, and Hua Wu. 2021. Weakly supervised dense video captioning via jointly usage of knowledge distillation and cross-modal matching. *arXiv preprint arXiv:2105.08252*.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9777–9786.

Junbin Xiao, Angela Yao, Yicong Li, and Tat Seng Chua. 2023a. Can i trust your answer? visually grounded video question answering. *arXiv preprint arXiv:2309.01327*.

Junbin Xiao, Angela Yao, Zhiyuan Liu, Yicong Li, Wei Ji, and Tat-Seng Chua. 2022a. Video as conditional graph hierarchy for multi-granular question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2804–2812.

Junbin Xiao, Pan Zhou, Tat-Seng Chua, and Shuicheng Yan. 2022b. Video graph transformer for video question answering. In *European Conference on Computer Vision*, pages 39–58. Springer.

Junbin Xiao, Pan Zhou, Angela Yao, Yicong Li, Richang Hong, Shuicheng Yan, and Tat-Seng Chua. 2023b. Contrastive video question answering via video graph transformer. *arXiv preprint arXiv:2302.13668*.

Saining Xie, Ross Girshick, Piotr Dollar, Zhuowen Tu, and Kaiming He. 2017. Aggregated Residual Transformations for Deep Neural Networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhen Xing, Qi Dai, Han Hu, Jingjing Chen, Zuxuan Wu, and Yu-Gang Jiang. 2023. Svformer: Semi-supervised video transformer for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18816–18826.

Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. 2017.

Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653.

Haiyang Xu, Qinghao Ye, Ming Yan, Yaya Shi, Jiabo Ye, Yuanhong Xu, Chenliang Li, Bin Bi, Qi Qian, Wei Wang, et al. 2023. mplug-2: A modularized multi-modal foundation model across text, image and video. *arXiv preprint arXiv:2302.00402*.

Huijuan Xu, Kun He, Bryan A Plummer, Leonid Sigal, Stan Sclaroff, and Kate Saenko. 2019. Multilevel language and vision integration for text-to-clip retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9062–9069.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296.

Wanru Xu, Jian Yu, Zhenjiang Miao, Lili Wan, Yi Tian, and Qiang Ji. 2020. Deep reinforcement polishing network for video captioning. *IEEE Transactions on Multimedia*, 23:1772–1784.

Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. 2022a. Advancing high-resolution video-language representation with large-scale video transcriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5036–5045.

Hongwei Xue, Yuchong Sun, Bei Liu, Jianlong Fu, Ruihua Song, Houqiang Li, and Jiebo Luo. 2022b. Clip-vip: Adapting pre-trained image-text model to video-language representation alignment. *arXiv preprint arXiv:2209.06430*.

Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. 2022a. Zero-shot video question answering via frozen bidirectional language models. *Advances in Neural Information Processing Systems*, 35:124–141.

Antoine Yang, Arsha Nagrani, Ivan Laptev, Josef Sivic, and Cordelia Schmid. 2023. Vidchapters-7m: Video chapters at scale. *arXiv preprint arXiv:2309.13952*.

Bang Yang, Tong Zhang, and Yuexian Zou. 2022b. Clip meets video captioning: Concept-aware representation learning does matter. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 368–381. Springer.

Xun Yang, Fuli Feng, Wei Ji, Meng Wang, and Tat-Seng Chua. 2021. Deconfounded video moment retrieval with causal intervention. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1–10.

Yinchong Yang, Denis Krompass, and Volker Tresp. 2017. Tensor-train recurrent neural networks for video classification. In *International Conference on Machine Learning*, pages 3891–3900. PMLR.

Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville. 2015. Describing videos by exploiting temporal structure. In *Proceedings of the IEEE international conference on computer vision*, pages 4507–4515.

Yunan Ye, Zhou Zhao, Yimeng Li, Long Chen, Jun Xiao, and Yueting Zhuang. 2017. Video question answering via attribute-augmented attention network learning. In *Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 829–832.

Haonan Yu, Jiang Wang, Zhiheng Huang, Yi Yang, and Wei Xu. 2016. Video paragraph captioning using hierarchical recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4584–4593.

Tiezheng Yu, Wenliang Dai, Zihan Liu, and Pascale Fung. 2021. Vision guided generative pre-trained language models for multimodal abstractive summarization. *arXiv preprint arXiv:2109.02401*.

Ting Yu, Jun Yu, Zhou Yu, Qingming Huang, and Qi Tian. 2020. Long-term video question answering via multimodal hierarchical memory attentive networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(3):931–944.

Youngjae Yu, Jongseok Kim, and Gunhee Kim. 2018. A joint sequence fusion model for video question answering and retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 471–487.

Youngjae Yu, Hyungjin Ko, Jongwook Choi, and Gunhee Kim. 2017. End-to-end concept word detection for video captioning, retrieval, and question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3165–3173.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. 2019. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9127–9134.

Liangzhe Yuan, Nitesh Bharadwaj Gundavarapu, Long Zhao, Hao Zhou, Yin Cui, Lu Jiang, Xuan Yang, Menglin Jia, Tobias Weyand, Luke Friedman, et al. 2023. Videoglue: Video general understanding evaluation of foundation models. *arXiv preprint arXiv:2307.03166*.

Yitian Yuan, Tao Mei, and Wenwu Zhu. 2019. To find where you talk: Temporal sentence localization in video with attention based location regression. In

Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, pages 9159–9166.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. 2019. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032.

Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. 2021. Merlot: Multimodal neural script knowledge models. *Advances in Neural Information Processing Systems*, 34:23634–23651.

Kuo-Hao Zeng, Tseng-Hung Chen, Ching-Yao Chuang, Yuan-Hong Liao, Juan Carlos Niebles, and Min Sun. 2017. Leveraging video descriptions to learn video question answering. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Yawen Zeng, Da Cao, Shaofei Lu, Hanling Zhang, Jiao Xu, and Zheng Qin. 2022. Moment is important: Language-based video moment retrieval via adversarial learning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 18(2):1–21.

Bowen Zhang, Xiaojie Jin, Weibo Gong, Kai Xu, Zhao Zhang, Peng Wang, Xiaohui Shen, and Jiashi Feng. 2023a. Multimodal video adapter for parameter efficient video text retrieval. *arXiv preprint arXiv:2301.07868*.

Hang Zhang, Xin Li, and Lidong Bing. 2023b. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.

Ziqi Zhang, Yaya Shi, Chunfeng Yuan, Bing Li, Peijin Wang, Weiming Hu, and Zheng-Jun Zha. 2020. Object relational graph with teacher-recommended learning for video captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13278–13288.

Rui Zhao, Haider Ali, and Patrick Van der Smagt. 2017a. Two-stream rnn/cnn for action recognition in 3d videos. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4260–4267. IEEE.

Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. 2023. Learning video representations from large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6586–6597.

Zhou Zhao, Jinghao Lin, Xinghua Jiang, Deng Cai, Xiaofei He, and Yueting Zhuang. 2017b. Video question answering via hierarchical dual-level attention network learning. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1050–1058.

Yaoyao Zhong, Junbin Xiao, Wei Ji, Yicong Li, Weihong Deng, and Tat-Seng Chua. 2022. Video question answering: Datasets, algorithms and challenges. *arXiv preprint arXiv:2203.01225*.

Luowei Zhou, Chenliang Xu, and Jason Corso. 2018a. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. 2018b. End-to-end dense video captioning with masked transformer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8739–8748.

Yue Zhou, Chenlu Guo, Xu Wang, Yi Chang, and Yuan Wu. 2024. A survey on data augmentation in large model era. *arXiv preprint arXiv:2401.15422*.

Cunjuan Zhu, Qi Jia, Wei Chen, Yanming Guo, and Yu Liu. 2023. Deep learning for video-text retrieval: a review. *International Journal of Multimedia Information Retrieval*, 12(1):3.

Jiafan Zhuang, Zilei Wang, and Junjie Li. 2023. Video semantic segmentation with inter-frame feature fusion and inner-frame feature refinement. *arXiv preprint arXiv:2301.03832*.

Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, and Josef Sivic. 2019. Cross-task weakly supervised learning from instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3537–3545.

Appendix

A More Examples of Video-Language Understanding tasks

Due to limited space, further examples of video-language understanding tasks are provided in Figure 4.

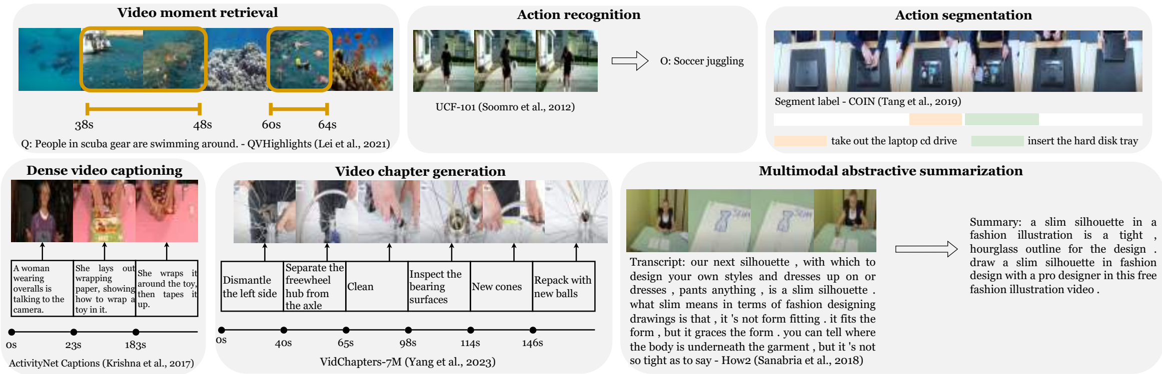

Figure 4: More examples of video-language understanding tasks.

B Details of Video-Language Understanding performance

Due to page limit, full details of performance in text-video retrieval, video captioning, and videoQA tasks are listed in Table 1, 2, and 3, respectively.

C Details of Video-Language Understanding datasets

Due to page limit, details of the datasets for video-language understanding tasks are listed in Table 4.

Methods	**Model architecture**	**Video**	**Text**	**R@1**	**R@5**	**R@10**
JSFusion (Yu et al., 2018)	Pre-TF	RN	GloVe-LSTM	10.2	31.2	43.2
C+LSTM+SA-FC7 (Torabi et al., 2016)		VGG	GloVe-LSTM	4.2	12.9	19.9
VSE-LSTM (Kiros et al., 2014)		ConvNet/OxfordNet	GloVe-LSTM	3.8	12.7	17.1
EITanque (Kaufman et al., 2016)		VGG	word2vec-LSTM	4.7	16.6	24.1
SA-G+SA-FC7 (Torabi et al., 2016)		VGG	GloVe	3.1	9.0	13.4
CT-SAN (Yu et al., 2017)		RN	word2vec-LSTM	4.4	16.6	22.3
All-in-one (Wang et al., 2023a)	Shared TF	ViT	BT	37.9	68.1	77.1
VindLU (Cheng et al., 2023)	Stacked TF	ViT	BT	48.8	72.4	82.2
HERO (Li et al., 2020)	Stacked TF	RN+SlowFast	BT	16.8	43.4	57.7
MV-GPT (Seo et al., 2022)	Stacked TF	ViViT	BT	37.3	65.5	75.1
CLIP-ViP (Xue et al., 2022a)	Dual TF	ViT	CLIP-text	49.6	74.5	84.8
CLIP4Clip (Luo et al., 2022)	Dual TF	ViT	CLIP-text	44.5	71.4	81.6

Table 1: Performance on text-video retrieval. (Pre-TF: Pre-transformer, Shared TF: Shared Transformer, Stack TF: Stack Transformer, Dual TF: Dual Transformer, RN: ResNet/ResNeXt (He et al., 2016; Xie et al., 2017), ViT: Vision Transformer (Dosovitskiy et al., 2020), BT: BERT (Devlin et al., 2018), ViViT: Video Vision Transformer (Arnab et al., 2021)). We report recall at rank 1 (R@1), 5 (R@5), and 10 (R@10). We choose MSRVTT as one of the most popular datasets for text-video retrieval.

Methods	Model architecture	Video	BLEU-4	METEOR
MFATT (Long et al., 2018)	Pre-TF	Video: RN+C3D	39.1	26.7
TA (Yao et al., 2015)		Video: 3D-CNN	36.5	25.7
h-RNN (Yu et al., 2016)		Video: VGG	36.8	25.9
CAT-TM (Long et al., 2018)		Video: RN+C3D	36.6	25.6
NFS-TM (Long et al., 2018)		Video: RN+C3D	37.0	25.9
Fuse-TM (Long et al., 2018)		Video: RN+C3D	37.5	25.9
VLAB (He et al., 2023)	Stacked TF	EVA-G	54.6	33.4
UniVL (Luo et al., 2020)		S3D	41.8	28.9
MV-GPT (Seo et al., 2022)		ViViT	48.9	38.7
CLIP-DCD (Yang et al., 2022b)		ViT	48.2	30.9
DeCEMBERT (Tang et al., 2021)		RN	45.2	29.7
mPLUG-2 (Xu et al., 2023)		ViT	57.8	34.9

Table 2: Performance on video captioning. (Pre-TF: Pre-transformer, Stacked TF: Stacked Transformer, RN: ResNet/ResNeXt (He et al., 2016; Xie et al., 2017), ViViT: Video Vision Transformer (Arnab et al., 2021), EVA-G: Fang et al. (2023b)). We report BLEU-4 and METEOR, which are two popular metrics for language generation. We choose MSRVTT as one of the most popular datasets for video captioning.

Methods	Architecture	Video	Text	Dataset	
				MSRVTT	MSVD
QueST (Jiang et al., 2020)	Pre-TF	RN + C3D	GloVe-LSTM	40.0	-
HME (Fan et al., 2019)		RN/VGG + C3D	GloVe-GRU	34.6	36.1
HGA (Jiang and Han, 2020)		RN/VGG + C3D	GloVe-GRU	33.0	33.7
ST-VQA (Jang et al., 2019)		RN+C3D	GloVe-LSTM	35.5	34.7
PGAT (Peng et al., 2021)		Faster-RCNN	GloVe-LSTM	38.1	39.0
HCRN (Le et al., 2020)		RN	GloVe-LSTM	35.6	36.1
HQGA (Xiao et al., 2022a)		Faster-RCNN	BERT-LSTM	38.6	41.2
All in one (Wang et al., 2023a)	Shared TF	ViT	BT	44.3	47.9
LAVENDER (Li et al., 2023b)	Stacked TF	VS-TF	BT	45.0	56.6
VIOLET (Fu et al., 2023)	Stacked TF	VS-TF	BT	44.5	54.7
ClipBERT (Lei et al., 2021c)	Stacked TF	CLIP-text	BT	37.4	-
VGT (Xiao et al., 2022b)	Dual TF	Faster-RCNN	BT	39.7	-
CoVGT (Xiao et al., 2023b)	Dual TF	Faster-RCNN	BT	40.0	-
LLaMA-Vid (Li et al., 2023d)	LLM-Augmented	EVA-G	Vicuna	58.9	70.0

Table 3: Performance on videoQA. (Pre-TF: Pre-transformer, Dual TF: Dual Transformer, RN: ResNet/ResNeXt (He et al., 2016; Xie et al., 2017), BT: BERT (Devlin et al., 2018), VS-TF: Video Swin Transformer (Liu et al., 2021), EVA-G: Fang et al. (2023b)). We report accuracy of the methods. We choose MSRVTT and MSVD as two of the most popular datasets for videoQA.

Dataset	Video source	Annotation	Tasks	#Videos/#Routes
MSVD (Chen and Dolan, 2011)	YouTube videos	Manual	TVR, VC, VideoQA	1.9K
MSRVTT (Xu et al., 2016)	Web videos	Manual	TVR, VC, VideoQA	7.2K
ActivityNet (Yu et al., 2019)	YouTube videos	Manual	AL, TVR, VC, VMR	5.8K
FIBER (Castro et al., 2022b)	VaTeX (Wang et al., 2019)	Manual	VC, VideoQA	28K
WildQA (Castro et al., 2022a)	YouTube videos	Manual	VideoQA	0.4K
NExT-QA (Xiao et al., 2021)	VidOR (Shang et al., 2019)	Manual	VideoQA	5.4K
CausalVid-QA (Li et al., 2022a)	Kinetics-700 (Carreira et al., 2019)	Manual	VideoQA	26K
HowTo100M (Miech et al., 2019)	YouTube videos	Auto	PT	1.2M
HD-VILA-100M (Xue et al., 2022a)	YouTube videos	Auto	PT	3.3M
YT-Temporal-180M (Zellers et al., 2021)	YouTube videos	Auto	PT	6M
TGIF-QA (Jang et al., 2017)	Animated GIFs	Manual	VideoQA	71K
TGIF-QA-R (Peng et al., 2021)	TGIF-QA (Jang et al., 2017)	Manual, Auto	VideoQA	71K
DiDeMo (Anne Hendricks et al., 2017)	YFCC100M (Thomee et al., 2016)	Manual	TVR	11K
YouCook2 (Zhou et al., 2018a)	YouTube videos	Manual	TVR, VC	2K
HMDB-51 (Kuehne et al., 2011)	Web videos	Manual	TVR, AR	6.8K
Kinetics-400 (Kay et al., 2017)	YouTube videos	Manual	AR	306K
Kinetics-600 (Carreira et al., 2018)	Kinetics-400 (Kay et al., 2017)	Manual	AR, VG	480K
Kinetics-700 (Carreira et al., 2019)	Kinetics-600 (Carreira et al., 2018)	Manual	AR	650K
VaTeX (Wang et al., 2019)	Kinetics-600 (Carreira et al., 2018)	Manual	TVR, VC	41K
TVR (Lei et al., 2020)	TVQA (Lei et al., 2018)	Manual	VMR	22K
How2R (Li et al., 2020)	HowTo100M (Miech et al., 2019)	Manual	VMR	22K
How2QA (Li et al., 2020)	HowTo100M (Miech et al., 2019)	Manual	VideoQA	22K
YouTube Highlights (Sun et al., 2014)	YouTube videos	Manual	VMR	0.6K
TACoS (Regneri et al., 2013)	MPII Composites (Rohrbach et al., 2012)	Manual	VMR	0.1K
QVHighlights (Lei et al., 2021b)	YouTube vlogs	Manual	VMR	10K
TVSum (Song et al., 2015)	YouTube videos	Manual	VMR	50
ViTT (Huang et al., 2020)	YouTube-8M (Abu-El-Haija et al., 2016)	Manual	VMR	5.8K
VidChapters-7M (Yang et al., 2023)	YT-Temporal-180M (Zellers et al., 2021)	Auto	VC, VMR	817K
VideoCC3M (Nagrani et al., 2022)	Web videos	Auto	PT	6.3M
WebVid-10M (Bain et al., 2021)	Web videos	Auto	PT	10.7M
COIN (Tang et al., 2019)	YouTube videos	Manual	AS	12K
CrossTask (Zhukov et al., 2019)	YouTube videos	Manual	AR	4.7K
Alivol-10M (Lei et al., 2021a)	E-commerce videos	Auto	PT	10M
LSMDC (Rohrbach et al., 2015)	British movies	Manual	TVR	72
EK-100 (Damen et al., 2022)	Manual	Manual	AR, AL	7K
SSV1 (Goyal et al., 2017)	Manual	Manual	AR	108K
SSV2 (Goyal et al., 2017)	Manual	Manual	AR	221K
Moments in Time (Monfort et al., 2019)	Web videos	Manual	AR	1M
InternVid (Wang et al., 2023b)	YouTube videos	Auto	PT	7.1M
How2 (Sanabria et al., 2018)	YouTube videos	Auto	VC	13.2K
WTS70M (Stroud et al., 2020)	YouTube videos	Auto	PT	70M
Charades (Gao et al., 2017)	Manual	Manual	AR, VMR, VideoQA	10K

Table 4: Video understanding datasets in the literature. (VMR: Video moment retrieval, TVR: text-video retrieval, VC: video captioning, AL: action localization, AR: action recognition, AS: action segmentation, VG: video generation, PT: pre-training).