

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 mod-

els 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 (Ruan and Jin, 2022) and **LLM-augmented architecture** (Tang et al., 2023b) (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 intelli-

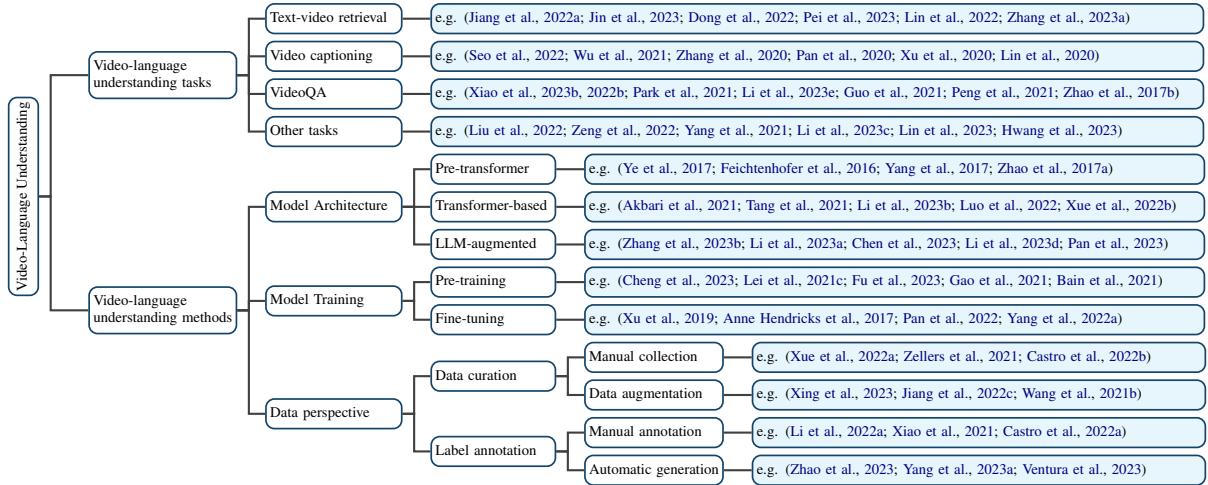


Figure 1: Taxonomy of Video-language Understanding

gence capable of versatile adaptation 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 full-fledged 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 of 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 training methods into Pre-training and Fine-tuning to adapt video-language representations to target downstream task. (3) *Data perspective*: we summarize ex-

isting 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

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). In practical applications, returning an entire video may not be desirable. Hence, video moment retrieval (VMR) has emerged with an aim to accurately locating relevant moments within a video based on user queries. VMR examines more nuanced and fine-grained understanding to capture different concepts and events in a video in order to pinpoint specific moments rather than capturing the overall theme in standard text-video retrieval.

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, or might also generate a paragraph as a more detailed summary.

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.

Connections among video-language understanding tasks. These tasks form the three fundamental testbeds for video-language understanding capacity (see Appendix B for their examples). In Figure 4 (Appendix A), we provide a hierarchy that describes the level-up of their video-language understanding degree. At the basic level, text-video retrieval globally associates a whole video with a textual content. In medium level, video captioning is more difficult than retrieval tasks since it needs to selectively maps entities and events within a video to the language modality. At the highest level, videoQA explores the relation of video and language content to produce the appropriate output. Each level of video-language understanding tasks is associated with a corresponding version that demands a more inferential or fine-grained understanding, *e.g.* inference videoQA (Xiao et al., 2021; Li et al., 2022a) with videoQA, dense video captioning (Zhou et al., 2018b) or video chapter generation (Yang et al., 2023b) with video captioning, and video moment retrieval (temporal grounding) with text-video retrieval. These more inferential or fine-grained tasks pose more challenges and play an increasingly significant role in current research heading towards the core of human intelligence (Fei-Fei and Krishna, 2022).

3 Challenges of Video-Language Understanding

The discussed video-language understanding tasks present unique challenges compared with image-language understanding, since a video incorporates an additional temporal channel. We summarize their important challenges 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 spa-

tial 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 modeling by necessitating the recognition of more objects and events in a higher number of video frames (Yu et al., 2020; Lin et al., 2022), and reasoning their long-term dependencies (Zhao et al., 2018). 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., 2023a).

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

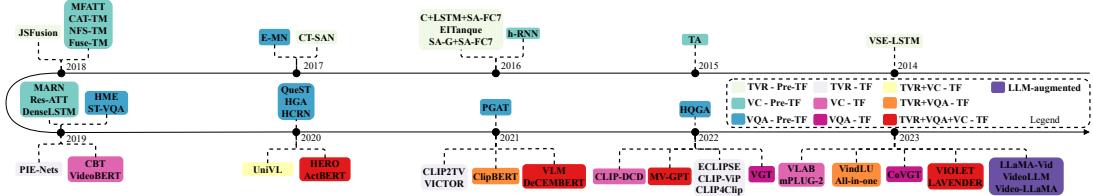


Figure 2: Timeline of the established video-language understanding methods (TVR: Text-video retrieval, VC: video captioning, VQA: video question answering, TF: Transformer, LLM: large language model). From left to right, our legend table follows the order: pre-Transformer (Pre-TF), task-specific Transformer, multi-task Transformer, and LLM-augmented architectures.

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.

Addressing challenges. These identified challenges encompass three critical perspectives: model architecture, model training, and data preparation in the field of video-language understanding. In general, there should be a synergistic relationship among these components. Specifically, model architecture should be designed to effectively capture video-language interactions. Concurrently, model training should be tailored to enable the architecture to adapt to target domains with their captured video-language interactions. Lastly, data preparation plays a pivotal role in shaping model training, which in turn significantly impacts the development of an efficacious model architecture.

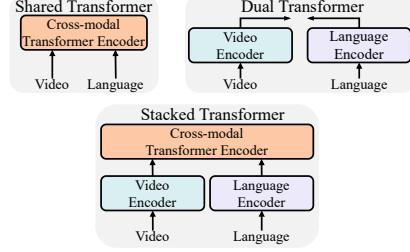


Figure 3: Illustration of video-language understanding Transformer-based architectures.

4 Model Architecture for Video-Language Understanding

Addressing the challenge of 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

325 CNN have also demonstrated effectiveness in ex-
326 tracting spatio-temporal representations (Tran et al.,
327 2017; Carreira and Zisserman, 2017). In addition
328 to CNN and RNN, Chen et al. (2018), Gay et al.
329 (2019), and Wei et al. (2017) also build graphs to
330 incorporate intra-modal relationships among video
331 entities such as video segments or visual objects.
332 These graph-structured works emphasize the rea-
333 soning ability of the model architecture.

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

341 **Cross-modal encoders.** Gao et al. (2017) and
342 Zeng et al. (2017) apply element-wise multiplication
343 to fuse the global video and question represen-
344 tations for video question answering. It demon-
345 strates the advantage of a simple operation for
346 video-language fusion. Attention has also been
347 used to model video-language relations, in order
348 to identify salient parts in video and language
349 sentence (Yuan et al., 2019), or to refine the represen-
350 tation of the video based on the language question
351 (Xu et al., 2017). Pre-transformer video-language
352 works have also combined attention with a wide
353 variety of techniques, including hierarchical learning
354 (Baraldi et al., 2017), memory networks (Fan et al.,
355 2019), and graph networks (Xiao et al., 2022a).

356 4.2 Transformer-based Architecture

357 Developed based on the self-attention mechanism,
358 which exhaustively correlates every pair of in-
359 put tokens with each other, Transformer-based
360 architecture has the capacity to capture long-
361 term dependencies and learn from web-scale data.
362 It has demonstrated remarkable performance in
363 many video-language tasks. Similar to the pre-
364 transformer architecture, the Transformer-based
365 framework also comprises unimodal encoders and
366 cross-modal encoders to model intra-modal and
367 cross-modal interactions, respectively. For uni-
368 modal encoders, several works find vision trans-
369 former for video encoding and BERT encoder for
370 language encoding outperform RNN- and CNN-
371 based encoding (Fu et al., 2021; Bain et al., 2021;
372 Seo et al., 2022). We then summarize fundamen-
373 tal types of Transformer-based architectures and
374 illustrate them in Figure 3.

375 **Shared Transformer.** Motivated by the success of
376 Transformer in language modeling (Devlin et al.,
377 2018), Akbari et al. (2021) and Wang et al. (2023a)
378 construct a shared Transformer encoder for video-
379 language understanding. Their encoder architec-
380 tures receive the concatenation of visual patches
381 and language tokens, then jointly calculate their
382 interactions in a BERT-based manner. Akbari et al.
383 (2021) additionally incorporate modality embed-
384 dings which comprise three values to denote three
385 kinds of input modalities, *i.e.* (video, audio, text).
Stacked Transformer. Li et al. (2020) reveals that
386 a shared Transformer encoder is weak in model-
387 ing temporal relations between videos and texts.
388 To address this problem, they introduce a stacked
389 Transformer architecture, with a hierarchical stack
390 consisting of unimodal encoders to encode video
391 and language inputs separately, and then a cross-
392 modal Transformer to compute video-language
393 interactions. A multitude of video-language under-
394 standing works follow such design to stack a cross-
395 modal Transformer-based encoder above unimodal
396 encoders (Fu et al., 2023; Li et al., 2023b; Lei et al.,
397 2021c; Luo et al., 2022; Nie et al., 2022). To per-
398 form video captioning, Seo et al. (2022) and Luo
399 et al. (2020) further insert a causal Transformer-
400 based decoder that generates language tokens based
401 on the encoded cross-modal representations.

402 **Dual Transformer.** Dual Transformer architec-
403 tures have been favored for text-video retrieval
404 (Luo et al., 2022; Bain et al., 2021, 2022; Lin et al.,
405 2022; Xue et al., 2022b). These architectures use
406 two Transformer encoders to encode video and
407 language separately, yielding global representa-
408 tions for each input modality, then applying simple
409 operations such as cosine similarity to compute
410 cross-modal interaction. Such a separate encoding
411 scheme enables them to mitigate the computational
412 cost of computing pairwise interactions between
413 every pair of video and language inputs. They have
414 accomplished not only efficiency but also effective-
415 ness in text-video retrieval problems.

417 4.3 LLM-Augmented Architecture

418 Large language models (LLMs) have achieved im-
419 pressive results in simultaneously tackling mul-
420 tiple NLP tasks. Recent efforts have sought to
421 apply LLMs for video-language understanding to
422 extend its cross-domain adaptation ability to video-
423 language settings (Chen et al., 2023; Li et al.,
424 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 pre-trained 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 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 Architecture Analysis

In Figure 2, we show the timeline of video-language understanding methodologies, categorized according to our defined architecture taxonomy and their affiliated downstream tasks. We also list details regarding their performance in Table 1, 2, and 3 (see Appendix C). The evolution of pre-transformer models aligns with our hierarchy of video-language understanding levels, i.e. models for video captioning generally appear subsequent to those for text-video retrieval, followed by the development of videoQA models. Owing to their impressive capacity, Transformer-based mod-

els capable of addressing multiple tasks have been introduced concurrently with task-specific Transformer frameworks. Recently, large language models (LLMs) have gained prominence for their superior in-context learning ability, enabling them to handle diverse tasks without fine-tuning. Consequently, new LLM-augmented architectures have emerged to utilize this capability to address multiple understanding tasks.

5 Model Training for Video-Language Understanding

Model training seeks to address the cross-domain adaptation ability of video-language understanding models. To achieve this goal, pre-training strategies have been devised to gain world knowledge that generalizes across multiple scenarios, then task-specific fine-tuning is conducted to specifically improve downstream task performance.

5.1 Pre-training for Video-Language Understanding

In this section, we mainly summarize pre-training strategies for video-language understanding models 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

524 to train the model to predict pre-trained features
525 of the masked video frames extracted by ResNet
526 and SlowFast models, while Fu et al. (2021) use
527 cross-entropy loss to train the model to predict the
528 masked visual tokens, which are quantized by a
529 variational autoencoder from visual frame patches.

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

544 Video-text contrastive learning has revealed
545 promising results for text-video retrieval (Lin et al.,
546 2022; Gao et al., 2021; Xue et al., 2022b). MLM
547 has contributed to enhancing VideoQA since the
548 task resembles MLM in predicting the language
549 word given a video-language pair (the question is
550 the language input in videoQA). Compared to these
551 pre-training strategies, MVM does provide performance
552 gain for video-language understanding but its gain is less significant. **For more details about**
553 **pre-training, please refer to** (Cheng et al., 2023).

554 **5.2 Fine-tuning for Video-Language 555 Understanding**

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

566 **Fine-tuning strategies.** Normally, all of the model
567 parameters are updated during fine-tuning (Gao
568 et al., 2017; Xu et al., 2019; Anne Hendricks et al.,
569 2017). However, in cases computational resources

570 or training data are limited, only adaptation layers
571 such as low-rank adapters (Pan et al., 2022;
572 Yang et al., 2022a) or learnable prompt vectors (Ju
573 et al., 2022) are fine-tuned to reduce training cost or
574 prevent overfitting. Such risks also apply for LLM-
575 augmented architectures discussed in Section 4.3,
576 since LLMs exhibit a billion scale of parameters,
577 thus incurring excessively huge cost if full fine-
578 tuning is conducted. For such models, Zhang et al.
579 (2023b) and Li et al. (2023d) design a two-stage
580 instruction tuning strategy which only fine-tunes
581 the semantic translator. The first stage trains the
582 model to generate the textual description based on
583 the combined video and the language instruction,
584 in order to align visual representations extracted by
585 the visual encoder with the language space of LLM.
586 The second stage is often performed on small-scale
587 video-text pairs manually collected by the authors
588 to further tailor the output features of the translator
589 towards the target domains.

590 **6 Data Perspective for Video-Language 591 Understanding**

592 In this section, we analyze data preparation ap-
593 proaches for video-language understanding models,
594 and provide details of the datasets in Appendix D.

595 **6.1 Data curation**

596 **Manual collection.** To curate video-language data,
597 multiple works search for publicly available online
598 videos, which exhibit a wide diversity of content.
599 Video-language datasets with online videos are
600 mostly aimed for pre-training models to learn gen-
601 eralizable knowledge, e.g. HowTo100M (Miech
602 et al., 2019) and YT-Temporal-180M (Zellers et al.,
603 2021), or they can also be used for fine-tuning, e.g.
604 MSRVTT (Xu et al., 2016) and YouCook2 (Zhou
605 et al., 2018a). To satisfy a certain requirement,
606 videos different from the online ones can be inher-
607 ited from existing datasets, e.g Xiao et al. (2021)
608 utilize 6,000 videos from VidOR dataset and (Li
609 et al., 2022a) inherit 546,882 videos from Kinetics-
610 700 since they describe scenes of daily life and real
611 world, respectively. Apart from making use of ex-
612 isting datasets' and online videos, videos can also
613 be recorded by human annotators to enable quality
614 control (Goyal et al., 2017; Damen et al., 2022).

615 **Data augmentation.** Rather than manually col-
616 lecting videos from external sources, Xing et al.
617 (2023) and Jiang et al. (2022c) explore data aug-
618 mentation techniques which are particularly de-

624 signed for videos. In detail, their TubeTokenMix
625 mixes two videos in which the mixing coefficient
626 is defined upon the temporal dimension, and their
627 temporal shift randomly shifts video frame features
628 backward or forward over the temporal dimension.
629 These techniques outperform standard augmentation
630 approaches for image data, such as CutMix
631 (Yun et al., 2019), Mixup (Zhang et al., 2017), and
632 PixMix (Hendrycks et al., 2022).

633 6.2 Label annotation

634 **Manual annotation.** Several works (Li et al.,
635 2022a; Lei et al., 2021b; Xiao et al., 2021) use hu-
636 man annotators since they provide high-quality la-
637 bels. However, such approach is expensive, partic-
638 ularly when dealing with video data. For example,
639 annotating QVHighlights dataset (Lei et al., 2021b)
640 costs approximately \$16,000 for 10K videos and
641 3 months to complete. Similarly, NExT-QA (Xiao
642 et al., 2021) needs 100 undergraduate students and
643 1 year to annotate only 5K videos.

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

656 7 Future Directions

657 **Fine-grained understanding.** Existing methods
658 excel at video-language understanding at a coarse-
659 grained level, enabling effective responses to ques-
660 tions like “*what is*” or the recognition of global
661 events without significant difficulty (Xiao et al.,
662 2021). Nevertheless, limiting comprehension to
663 this coarse level could hinder practical utility of
664 existing systems. In real-world scenarios, a user
665 might require a precise timestamp and location of
666 an object within a video (Jiang et al., 2022b), or
667 request the AI agent to forecast potential alterna-
668 tive events, which is a common need in predictive
669 analytics (Xiao et al., 2021; Li et al., 2022a). These
670 tasks necessitate an advanced understanding and
671 inference capability regarding the causal and tem-
672 poral relationships present in a video. At present,
673 models exhibit a constrained visio-linguistic ca-

674 pacity to engage in temporal reasoning, categoriz-
675 ing them as image-sequence-and-language models
676 rather than video-language models (Kesen et al.,
677 2023). Therefore, future research in this direction
678 deserves more attention and exploration.

679 **Long-form video-language understanding.** Cur-
680 rent understanding systems have demonstrated re-
681 markable performance on short video clips lasting
682 several seconds. However, they tend to struggle
683 when switching to long-form videos which last
684 several minutes or hours. To enhance the applica-
685 bility of these systems, it is essential to enhance
686 their capability of understanding long-form videos.
687 Current approaches mainly feature reducing com-
688 putational cost through architectures more efficient
689 than Transformer-based ones such as state space
690 models (Yang et al., 2024; Li et al., 2024), which
691 can be considered as linear RNN with specifically
692 designed fixed weights, or compensating sparsely
693 extracted video frames with additional information
694 (Lin et al., 2022). In general, how to effectively
695 model long-form videos and adapt them to the joint
696 context with language deserves more attention.

697 **Trustworthiness of video-language understand-
698 ing models.** Although modern video-language un-
699 derstanding systems have demonstrated remarkable
700 performance, their black-box nature undermines
701 our trust to deploy them. In particular, we still do
702 not precisely understand what part of the video a
703 videoQA model looks at to answer the question (Li
704 et al., 2022b), or how video and language seman-
705 tic information flows into the common representa-
706 tion space of the video retrieval model (Jia et al.,
707 2022). Furthermore, adversarial noise sensitivity
708 or hallucination of video-language understanding
709 models are also open problems. Future trustworthi-
710 ness benchmarks such as (Xiao et al., 2023a; Wang
711 et al., 2021a) for video-language understanding are
712 of great significance towards practical systems.

713 8 Conclusion

714 In this paper, we survey the broad research field
715 of video-language understanding. Particularly, we
716 categorize related video-language understanding
717 tasks and discuss meaningful insights from model
718 architecture, model training, and data perspectives.
719 We thoroughly analyze each perspective, and fi-
720 nally conclude with promising future directions.
721 We hope our survey can foster more research to-
722 wards constructing effective AI systems that can
723 comprehensively understand dynamic visual world
724 and meaningfully interact with humans.

725 9 Limitations

726 Although we have sought to comprehensively analyze
727 the literature of video-language understanding,
728 we might not fully cover all of the tasks, model ar-
729 chitectures, model training, and data perspectives.
730 Therefore, we complement the survey with a reposi-
731 tory¹. The repository comprises the latest video-
732 language understanding papers, datasets, and their
733 open-source implementations. We will periodically
734 update the repository to trace the progress of the
735 latest research.

736 References

- 737 Moloud Abdar, Meenakshi Kollati, Swaraja Kuraparthi,
738 Farhad Pourpanah, Daniel McDuff, Mohammad
739 Ghavamzadeh, Shuicheng Yan, Abdullah Mohamed,
740 Abbas Khosravi, Erik Cambria, et al. 2023. A review
741 of deep learning for video captioning. *arXiv preprint*
742 *arXiv:2304.11431*.
- 743 Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul
744 Natsev, George Toderici, Balakrishnan Varadarajan,
745 and Sudheendra Vijayanarasimhan. 2016. YouTube-
746 8M: A Large-Scale Video Classification Benchmark.
747 *arXiv preprint arXiv:1609.08675*.
- 748 Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong
749 Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong.
750 2021. Vatt: Transformers for multimodal self-
751 supervised learning from raw video, audio and text.
752 *Advances in Neural Information Processing Systems*,
753 34:24206–24221.
- 754 Lisa Anne Hendricks, Oliver Wang, Eli Shechtman,
755 Josef Sivic, Trevor Darrell, and Bryan Russell. 2017.
756 Localizing moments in video with natural language.
757 In *Proceedings of the IEEE international conference*
758 *on computer vision*, pages 5803–5812.
- 759 Anurag Arnab, Mostafa Dehghani, Georg Heigold,
760 Chen Sun, Mario Lučić, and Cordelia Schmid. 2021.
761 Vivit: A video vision transformer. In *Proceedings of*
762 *the IEEE/CVF international conference on computer*
763 *vision*, pages 6836–6846.
- 764 Max Bain, Arsha Nagrani, Gü̈l Varol, and Andrew Zis-
765 serman. 2021. Frozen in time: A joint video and
766 image encoder for end-to-end retrieval. In *Pro-
767 ceedings of the IEEE/CVF International Conference on*
768 *Computer Vision*, pages 1728–1738.
- 769 Max Bain, Arsha Nagrani, Gü̈l Varol, and Andrew Zis-
770 serman. 2022. A clip-hitchhiker’s guide to long video
771 retrieval. *arXiv preprint arXiv:2205.08508*.

¹Due 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.	772
My view is the best view: Procedure learning from egocentric videos. In <i>European Conference on Computer Vision</i> , pages 657–675. Springer.	773
Lorenzo Baraldi, Costantino Grana, and Rita Cucchiara.	774
2017. Hierarchical boundary-aware neural encoder for video captioning. In <i>Proceedings of the IEEE conference on computer vision and pattern recognition</i> , pages 1657–1666.	775
Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. 2018. A short note about kinetics-600. <i>arXiv preprint arXiv:1808.01340</i> .	781
Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. 2019. A short note on the kinetics-700 human action dataset. <i>arXiv preprint arXiv:1907.06987</i> .	782
Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In <i>proceedings of the IEEE Conference on Computer Vision and Pattern Recognition</i> , pages 6299–6308.	783
Santiago Castro, Naihao Deng, Pingxuan Huang, Mihai Burzo, and Rada Mihalcea. 2022a. In-the-wild video question answering. In <i>Proceedings of the 29th International Conference on Computational Linguistics</i> , pages 5613–5635.	784
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 <i>Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 2925–2940.	785
David Chen and William B Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In <i>Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies</i> , pages 190–200.	786
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. <i>arXiv preprint arXiv:2305.13292</i> .	787
Yuting Chen, Joseph Wang, Yannan Bai, Gregory Castañón, and Venkatesh Saligrama. 2018. Probabilistic semantic retrieval for surveillance videos with activity graphs. <i>IEEE Transactions on Multimedia</i> , 21(3):704–716.	788
Feng Cheng, Xizi Wang, Jie Lei, David Crandall, Mohit Bansal, and Gedas Bertasius. 2023. Vindlu: A recipe for effective video-and-language pretraining. In <i>Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition</i> , pages 10739–10750.	789

826	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. <i>International Journal of Computer Vision</i> , pages 1–23.	880
827		881
828		882
829		883
830		884
831		885
832		886
833	Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. <i>arXiv preprint arXiv:1810.04805</i> .	887
834		888
835		889
836		890
837	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 <i>Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition</i> , pages 4964–4973.	891
838		892
839		893
840		894
841		895
842		896
843	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. <i>IEEE transactions on circuits and systems for video technology</i> , 32(8):5680–5694.	897
844		898
845		899
846		900
847		901
848	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. <i>arXiv preprint arXiv:2010.11929</i> .	902
849		903
850		904
851		905
852		906
853		907
854		908
855	Chenyou Fan, Xiaofan Zhang, Shu Zhang, et al. 2019. Heterogeneous memory enhanced multimodal attention model for video question answering. In <i>Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition</i> , pages 1999–2007.	909
856		910
857		911
858		912
859		913
860	Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. 2023. Eva: Exploring the limits of masked visual representation learning at scale. In <i>Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition</i> , pages 19358–19369.	914
861		915
862		916
863		917
864		918
865		919
866		920
867	Li Fei-Fei and Ranjay Krishna. 2022. Searching for computer vision north stars. <i>Journal of the American Academy of Arts & Sciences</i> , page 85.	921
868		922
869		923
870	Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. 2016. Convolutional Two-Stream Network Fusion for Video Action Recognition. In <i>The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)</i> .	924
871		925
872		926
873		927
874		928
875	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. <i>arXiv preprint arXiv:2111.12681</i> .	929
876		930
877		931
878		932
879		933
880	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 <i>Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition</i> , pages 22898–22909.	934
881		935
882		936
883		937
884		938
885		939
886		940
887	Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. Tall: Temporal activity localization via language query. In <i>Proceedings of the IEEE international conference on computer vision</i> , pages 5267–5275.	941
888		942
889		943
890		944
891		945
892	Zijian Gao, Jingyu Liu, Weiqi Sun, Sheng Chen, Dedan Chang, and Lili Zhao. 2021. Clip2tv: Align, match and distill for video-text retrieval. <i>arXiv preprint arXiv:2111.05610</i> .	946
893		947
894		948
895		949
896	Paul Gay, James Stuart, and Alessio Del Bue. 2019. Visual graphs from motion (vgfm): Scene understanding with object geometry reasoning. In <i>Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14</i> , pages 330–346. Springer.	950
897		951
898		952
899		953
900		954
901		955
902		956
903	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 <i>The IEEE International Conference on Computer Vision (ICCV)</i> .	957
904		958
905		959
906		960
907		961
908		962
909		963
910		964
911	Zhicheng Guo, Jiaxuan Zhao, Licheng Jiao, Xu Liu, and Lingling Li. 2021. Multi-scale progressive attention network for video question answering. In <i>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)</i> , pages 973–978.	965
912		966
913		967
914		968
915		969
916		970
917		971
918	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. <i>ACM Transactions on Multimedia Computing, Communications and Applications</i> , 20(3):1–21.	972
919		973
920		974
921		975
922		976
923	Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In <i>The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)</i> .	977
924		978
925		979
926		980
927	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. <i>arXiv preprint arXiv:2305.13167</i> .	981
928		982
929		983
930		984
931		985
932	Dan Hendrycks, Andy Zou, Mantas Mazeika, Leonard Tang, Bo Li, Dawn Song, and Jacob Steinhardt. 2022. Pixmix: Dreamlike pictures comprehensively improve safety measures. In <i>Proceedings of the</i>	986
933		987
934		988
935		989

936	<i>IEEE/CVF Conference on Computer Vision and Pattern Recognition</i> , pages 16783–16792.	Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. 2022. Prompting visual-language models for efficient video understanding. In <i>European Conference on Computer Vision</i> , pages 105–124. Springer.	991 992 993 994 995
937			
938	Gabriel Huang, Bo Pang, Zhenhai Zhu, Clara Rivera, and Radu Soricut. 2020. Multimodal pretraining for dense video captioning. <i>arXiv preprint arXiv:2011.11760</i> .	Dotan Kaufman, Gil Levi, Tal Hassner, and Lior Wolf. 2016. Temporal tessellation for video annotation and summarization. <i>arXiv preprint arXiv:1612.06950</i> , 3.	996 997 998
939			
940			
941			
942	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 <i>Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition</i> , pages 6683–6693.	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. <i>arXiv preprint arXiv:1705.06950</i> .	999 1000 1001 1002 1003 1004
943			
944			
945			
946			
947			
948	Yunseok Jang, Yale Song, Chris Dongjoo Kim, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 2019. Video question answering with spatio-temporal reasoning. <i>International Journal of Computer Vision</i> , 127(10):1385–1412.	Ilker Kesenci, Andrea Pedrotti, Mustafa Dogan, Michele Cafagna, Emre Can Acikgoz, Letitia Parcalabescu, Iacer Calixto, Anette Frank, Albert Gatt, Aykut Erdem, et al. 2023. Vilma: A zero-shot benchmark for linguistic and temporal grounding in video-language models. <i>arXiv preprint arXiv:2311.07022</i> .	1005 1006 1007 1008 1009 1010
949			
950			
951			
952			
953	Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 2017. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In <i>Proceedings of the IEEE conference on computer vision and pattern recognition</i> , pages 2758–2766.	Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. 2014. Unifying visual-semantic embeddings with multimodal neural language models. <i>arXiv preprint arXiv:1411.2539</i> .	1011 1012 1013 1014
954			
955			
956			
957			
958	Mohan Jia, Zhongjian Dai, Yaping Dai, and Zhiyang Jia. 2022. An adversarial video moment retrieval algorithm. In <i>2022 41st Chinese Control Conference (CCC)</i> , pages 6689–6694. IEEE.	Hildegard Kuehne, Hueihan Jhuang, Estibaliz Garrote, Tomaso Poggio, and Thomas Serre. 2011. HMDB: A Large Video Database for Human Motion Recognition. In <i>The IEEE International Conference on Computer Vision (ICCV)</i> .	1015 1016 1017 1018 1019
959			
960			
961			
962	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. <i>arXiv preprint arXiv:2211.09623</i> .	Arnold A Lazarus. 1973. Multimodal behavior therapy: Treating the “basic id”. <i>The Journal of nervous and mental disease</i> , 156(6):404–411.	1020 1021 1022
963			
964			
965			
966	Ji Jiang, Meng Cao, Tengtao Song, and Yuxian Zou. 2022b. Video referring expression comprehension via transformer with content-aware query. <i>arXiv preprint arXiv:2210.02953</i> .	Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. 2020. Hierarchical conditional relation networks for video question answering. In <i>Proceedings of the IEEE/CVF conference on computer vision and pattern recognition</i> , pages 9972–9981.	1023 1024 1025 1026 1027
967			
968			
969			
970	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 <i>Proceedings of the AAAI Conference on Artificial Intelligence</i> , volume 34, pages 11101–11108.	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 <i>Proceedings of the 29th ACM International Conference on Multimedia</i> , pages 2567–2576.	1028 1029 1030 1031 1032 1033 1034
971			
972			
973			
974			
975			
976	Pin Jiang and Yahong Han. 2020. Reasoning with heterogeneous graph alignment for video question answering. In <i>Proceedings of the AAAI Conference on Artificial Intelligence</i> , volume 34, pages 11109–11116.	Jie Lei, Tamara L Berg, and Mohit Bansal. 2021b. Detecting moments and highlights in videos via natural language queries. <i>Advances in Neural Information Processing Systems</i> , 34:11846–11858.	1035 1036 1037 1038
977			
978			
979			
980			
981	Xun Jiang, Xing Xu, Jingran Zhang, Fumin Shen, Zuo Cao, and Heng Tao Shen. 2022c. Semi-supervised video paragraph grounding with contrastive encoder. In <i>Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition</i> , pages 2466–2475.	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 <i>Proceedings of the IEEE/CVF conference on computer vision and pattern recognition</i> , pages 7331–7341.	1039 1040 1041 1042 1043 1044
982			
983			
984			
985			
986			
987	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. <i>arXiv preprint arXiv:2303.09867</i> .		
988			
989			
990			

1045	Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg.	Ke Lin, Zhuoxin Gan, and Liwei Wang. 2020. Semi-supervised learning for video captioning. In <i>Findings of the Association for Computational Linguistics: EMNLP 2020</i> , pages 1096–1106.	1101
1046	2018. Tvana: Localized, compositional video question answering. In <i>Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing</i> , pages 1369–1379.		1102
1047			1103
1048			1104
1049			
1050	Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal.	Kunyang Lin, Peihao Chen, Diwei Huang, Thomas H Li, Mingkui Tan, and Chuang Gan. 2023. Learning vision-and-language navigation from youtube videos. In <i>Proceedings of the IEEE/CVF International Conference on Computer Vision</i> , pages 8317–8326.	1105
1051	2020. Tvr: A large-scale dataset for video-subtitle moment retrieval. In <i>Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16</i> , pages 447–463. Springer.		1106
1052			1107
1053			1108
1054			1109
1055			
1056	Jiangtong Li, Li Niu, and Liqing Zhang. 2022a. From representation to reasoning: Towards both evidence and commonsense reasoning for video question-answering. In <i>Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition</i> , pages 21273–21282.	Yan-Bo Lin, Jie Lei, Mohit Bansal, and Gedas Bertasius. 2022. Eclipse: Efficient long-range video retrieval using sight and sound. In <i>European Conference on Computer Vision</i> , pages 413–430. Springer.	1110
1057			1111
1058			1112
1059			1113
1060			
1061			
1062	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. <i>arXiv preprint arXiv:2305.06355</i> .	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 <i>Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition</i> , pages 3042–3051.	1114
1063			1115
1064			1116
1065			1117
1066	Kunchang Li, Xinhao Li, Yi Wang, Yinan He, Yali Wang, Limin Wang, and Yu Qiao. 2024. Videomamba: State space model for efficient video understanding. <i>arXiv preprint arXiv:2403.06977</i> .	Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. 2021. Video swin transformer. <i>arXiv preprint arXiv:2106.13230</i> .	1118
1067			1119
1068			
1069			
1070	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. <i>arXiv preprint arXiv:2005.00200</i> .	Xiang Long, Chuang Gan, and Gerard De Melo. 2018. Video captioning with multi-faceted attention. <i>Transactions of the Association for Computational Linguistics</i> , 6:173–184.	1120
1071			1121
1072			1122
1073			
1074	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 <i>Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition</i> , pages 23119–23129.	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. <i>arXiv preprint arXiv:2002.06353</i> .	1123
1075			1124
1076			1125
1077			1126
1078			
1079			
1080	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. <i>arXiv preprint arXiv:2307.02869</i> .	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. <i>Neurocomputing</i> , 508:293–304.	1127
1081			1128
1082			1129
1083			1130
1084			1131
1085	Xiangpeng Li, Zhilong Zhou, Lijiang Chen, and Lianli Gao. 2019. Residual attention-based lstm for video captioning. <i>World Wide Web</i> , 22:621–636.	Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2023. Video-chatgpt: Towards detailed video understanding via large vision and language models. <i>arXiv preprint arXiv:2306.05424</i> .	1132
1086			1133
1087			1134
1088			1135
1089			1136
1090			
1091	Yanwei Li, Chengyao Wang, and Jiaya Jia. 2023d. Llama-vid: An image is worth 2 tokens in large language models. <i>arXiv preprint arXiv:2311.17043</i> .	Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. 2023. Egoschema: A diagnostic benchmark for very long-form video language understanding. <i>arXiv preprint arXiv:2308.09126</i> .	1137
1092			1138
1093			1139
1094			1140
1095			1141
1096	Yicong Li, Xiang Wang, Junbin Xiao, and Tat-Seng Chua. 2022b. Equivariant and invariant grounding for video question answering. In <i>Proceedings of the 30th ACM International Conference on Multimedia</i> , pages 4714–4722.	Harry McGurk and John MacDonald. 1976. Hearing lips and seeing voices. <i>Nature</i> , 264(5588):746–748.	1142
1097			1143
1098			1144
1099			1145
1100	Yicong Li, Junbin Xiao, Chun Feng, Xiang Wang, and Tat-Seng Chua. 2023e. Discovering spatio-temporal rationales for video question answering. In <i>Proceedings of the IEEE/CVF International Conference on Computer Vision</i> , pages 13869–13878.	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 <i>Proceedings of the IEEE/CVF international conference on computer vision</i> , pages 2630–2640.	1146
1101			1147

1154		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. <i>IEEE transactions on pattern analysis and machine intelligence</i> , 42(2):502–508.	1208
1155		Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. 2013. Grounding action descriptions in videos. <i>Transactions of the Association for Computational Linguistics</i> , 1:25–36.	1209
1156			1210
1157			1211
1158			1212
1159		Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. 2015. A dataset for movie description. In <i>Proceedings of the IEEE conference on computer vision and pattern recognition</i> , pages 3202–3212.	1213
1160			1214
1161		Marcus Rohrbach, Sikandar Amin, Mykhaylo Andriluka, and Bernt Schiele. 2012. A database for fine grained activity detection of cooking activities. In <i>2012 IEEE conference on computer vision and pattern recognition</i> , pages 1194–1201. IEEE.	1215
1162			1216
1163		Ludan Ruan and Qin Jin. 2022. Survey: Transformer based video-language pre-training. <i>AI Open</i> , 3:1–13.	1217
1164			1218
1165		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. <i>arXiv preprint arXiv:1811.00347</i> .	1219
1166			1220
1167		Madeline C Schiappa, Yogesh S Rawat, and Mubarak Shah. 2023. Self-supervised learning for videos: A survey. <i>ACM Computing Surveys</i> , 55(13s):1–37.	1221
1168			1222
1169		Paul Hongsuck Seo, Arsha Nagrani, Anurag Arnab, and Cordelia Schmid. 2022. End-to-end generative pre-training for multimodal video captioning. In <i>Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition</i> , pages 17959–17968.	1223
1170			1224
1171		Xindi Shang, Donglin Di, Junbin Xiao, Yu Cao, Xun Yang, and Tat-Seng Chua. 2019. Annotating objects and relations in user-generated videos. In <i>Proceedings of the 2019 on International Conference on Multimedia Retrieval</i> , pages 279–287.	1225
1172			1226
1173		Karen Simonyan and Andrew Zisserman. 2014. Two-Stream Convolutional Networks for Action Recognition in Videos. In <i>Advances in Neural Information Processing Systems (NeurIPS)</i> .	1227
1174			1228
1175		Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. 2015. Tvsun: Summarizing web videos using titles. In <i>Proceedings of the IEEE conference on computer vision and pattern recognition</i> , pages 5179–5187.	1229
1176			1230
1177		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. <i>arXiv preprint arXiv:2007.14937</i> .	1231
1178			1232
1179		Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. Videobert: A joint model for video and language representation learning. In <i>Proceedings of the IEEE/CVF international conference on computer vision</i> , pages 7464–7473.	1233
1180			1234
1181			1235
1182			1236
1183			1237
1184			1238
1185			1239
1186			1240
1187			1241
1188			1242
1189			1243
1190			1244
1191			1245
1192			1246
1193			1247
1194			1248
1195			1249
1196			1250
1197			1251
1198			1252
1199			1253
1200			1254
1201			1255
1202			1256
1203			1257
1204			1258
1205			1259
1206			1260
1207			1260

1261	Min Sun, Ali Farhadi, and Steve Seitz. 2014.	Ranking domain-specific highlights by analyzing edited videos. In <i>Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part I 13</i> , pages 787–802. Springer.	1318
1262	Hao Tang, Kevin Liang, Kristen Grauman, Matt Feiszi, and Weiyao Wang. 2023a.	Egotracks: A long-term egocentric visual object tracking dataset. <i>arXiv preprint arXiv:2301.03213</i> .	1319
1263	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 <i>Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition</i> , pages 1207–1216.	1320
1264	Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, et al. 2023b.	Video understanding with large language models: A survey. <i>arXiv preprint arXiv:2312.17432</i> .	1321
1265	Zineng Tang, Jie Lei, and Mohit Bansal. 2021.	Decembert: Learning from noisy instructional videos via dense captions and entropy minimization. In <i>Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies</i> , pages 2415–2426.	1322
1266	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. <i>Communications of the ACM</i> , 59(2):64–73.	1323
1267	Atousa Torabi, Niket Tandon, and Leonid Sigal. 2016.	Learning language-visual embedding for movie understanding with natural-language. <i>arXiv preprint arXiv:1609.08124</i> .	1324
1268	Du Tran, Jamie Ray, Zheng Shou, Shih-Fu Chang, and Manohar Paluri. 2017.	Convnet architecture search for spatiotemporal feature learning. <i>arXiv preprint arXiv:1708.05038</i> .	1325
1269	Lucas Ventura, Antoine Yang, Cordelia Schmid, and GÜl Varol. 2023.	Covr: Learning composed video retrieval from web video captions. <i>arXiv preprint arXiv:2308.14746</i> .	1326
1270	Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. 2015.	Sequence to sequence-video to text. In <i>Proceedings of the IEEE international conference on computer vision</i> , pages 4534–4542.	1327
1271	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 <i>Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition</i> , pages 6598–6608.	1328
1272	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. <i>arXiv preprint arXiv:2108.13140</i> .	1329
1273	Xiang Wang, Shiwei Zhang, Zhiwu Qing, Yuanjie Shao, Changxin Gao, and Nong Sang. 2021b.	Self-supervised learning for semi-supervised temporal action proposal. In <i>Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition</i> , pages 1905–1914.	1330
1274	Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuanfang Wang, and William Yang Wang. 2019.	Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In <i>Proceedings of the IEEE/CVF International Conference on Computer Vision</i> , pages 4581–4591.	1331
1275	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. <i>arXiv preprint arXiv:2307.06942</i> .	1332
1276	Lina Wei, Fangfang Wang, Xi Li, Fei Wu, and Jun Xiao. 2017.	Graph-theoretic spatiotemporal context modeling for video saliency detection. In <i>2017 IEEE International Conference on Image Processing (ICIP)</i> , pages 4197–4201. IEEE.	1333
1277	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. <i>arXiv preprint arXiv:2105.08252</i> .	1334
1278	Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021.	Next-qa: Next phase of question-answering to explaining temporal actions. In <i>Proceedings of the IEEE/CVF conference on computer vision and pattern recognition</i> , pages 9777–9786.	1335
1279	Junbin Xiao, Angela Yao, Yicong Li, and Tat Seng Chua. 2023a.	Can i trust your answer? visually grounded video question answering. <i>arXiv preprint arXiv:2309.01327</i> .	1336
1280	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 <i>Proceedings of the AAAI Conference on Artificial Intelligence</i> , volume 36, pages 2804–2812.	1337
1281	Junbin Xiao, Pan Zhou, Tat-Seng Chua, and Shuicheng Yan. 2022b.	Video graph transformer for video question answering. In <i>European Conference on Computer Vision</i> , pages 39–58. Springer.	1338
1282	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. <i>arXiv preprint arXiv:2302.13668</i> .	1339

1373	Saining Xie, Ross Girshick, Piotr Dollar, Zhuowen Tu, and Kaiming He. 2017. Aggregated Residual Transformations for Deep Neural Networks. In <i>The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)</i> .	1426
1374		1427
1375		1428
1376		1429
1377		1430
1378	Zhen Xing, Qi Dai, Han Hu, Jingjing Chen, Zuxuan Wu, and Yu-Gang Jiang. 2023. Svformer: Semi-supervised video transformer for action recognition. In <i>Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition</i> , pages 18816–18826.	1431
1379		1432
1380		1433
1381		
1382		
1383		
1384	Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yuetong Zhuang. 2017. Video question answering via gradually refined attention over appearance and motion. In <i>Proceedings of the 25th ACM international conference on Multimedia</i> , pages 1645–1653.	1434
1385		1435
1386		1436
1387		1437
1388		1438
1389		1439
1390	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. <i>arXiv preprint arXiv:2302.00402</i> .	1440
1391		1441
1392		1442
1393		1443
1394		1444
1395	Hu Xu, Gargi Ghosh, Po-Yao Huang, Prahal Arora, Mousumeh Aminzadeh, Christoph Feichtenhofer, Florian Metze, and Luke Zettlemoyer. 2021. Vlm: Task-agnostic video-language model pre-training for video understanding. <i>arXiv preprint arXiv:2105.09996</i> .	1445
1396		1446
1397		1447
1398		1448
1399		1449
1400	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 <i>Proceedings of the AAAI Conference on Artificial Intelligence</i> , volume 33, pages 9062–9069.	1450
1401		1451
1402		
1403		
1404		
1405	Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msrvtt: A large video description dataset for bridging video and language. In <i>Proceedings of the IEEE conference on computer vision and pattern recognition</i> , pages 5288–5296.	1452
1406		1453
1407		1454
1408		
1409		
1410	Wanru Xu, Jian Yu, Zhenjiang Miao, Lili Wan, Yi Tian, and Qiang Ji. 2020. Deep reinforcement polishing network for video captioning. <i>IEEE Transactions on Multimedia</i> , 23:1772–1784.	1455
1411		1456
1412		1457
1413		1458
1414	Hongwei Xue, Tiansai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baineng Guo. 2022a. Advancing high-resolution video-language representation with large-scale video transcriptions. In <i>Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition</i> , pages 5036–5045.	1459
1415		1460
1416		1461
1417		1462
1418		1463
1419		1464
1420		
1421	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. <i>arXiv preprint arXiv:2209.06430</i> .	1465
1422		1466
1423		1467
1424		1468
1425		1469
1426	Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. 2022a. Zero-shot video question answering via frozen bidirectional language models. <i>Advances in Neural Information Processing Systems</i> , 35:124–141.	1470
1427		
1428		
1429		
1430		
1431	Antoine Yang, Arsha Nagrani, Ivan Laptev, Josef Sivic, and Cordelia Schmid. 2023a. Vidchapters-7m: Video chapters at scale. <i>arXiv preprint arXiv:2309.13952</i> .	1471
1432		1472
1433		1473
1434	Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, and Cordelia Schmid. 2023b. Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. In <i>Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition</i> , pages 10714–10726.	1474
1435		1475
1436		1476
1437		1477
1438		1478
1439		1479
1440		1480
1441	Bang Yang, Tong Zhang, and Yuexian Zou. 2022b. Clip meets video captioning: Concept-aware representation learning does matter. In <i>Chinese Conference on Pattern Recognition and Computer Vision (PRCV)</i> , pages 368–381. Springer.	1481
1442		1482
1443		1483
1444		1484
1445		1485
1446	Xun Yang, Fuli Feng, Wei Ji, Meng Wang, and Tat-Seng Chua. 2021. Deconfounded video moment retrieval with causal intervention. In <i>Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval</i> , pages 1–10.	1486
1447		1487
1448		1488
1449		1489
1450		1490
1451		1491
1452	Yijun Yang, Zhaohu Xing, and Lei Zhu. 2024. Vivim: a video vision mamba for medical video object segmentation. <i>arXiv preprint arXiv:2401.14168</i> .	1492
1453		1493
1454		1494
1455	Yinchong Yang, Denis Krompass, and Volker Tresp. 2017. Tensor-train recurrent neural networks for video classification. In <i>International Conference on Machine Learning</i> , pages 3891–3900. PMLR.	1495
1456		1496
1457		1497
1458		1498
1459	Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville. 2015. Describing videos by exploiting temporal structure. In <i>Proceedings of the IEEE international conference on computer vision</i> , pages 4507–4515.	1499
1460		1500
1461		1501
1462		1502
1463		1503
1464		1504
1465	Yunan Ye, Zhou Zhao, Yimeng Li, Long Chen, Jun Xiao, and Yuetong Zhuang. 2017. Video question answering via attribute-augmented attention network learning. In <i>Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval</i> , pages 829–832.	1505
1466		1506
1467		1507
1468		1508
1469		1509
1470		1510
1471	Haonan Yu, Jiang Wang, Zhiheng Huang, Yi Yang, and Wei Xu. 2016. Video paragraph captioning using hierarchical recurrent neural networks. In <i>Proceedings of the IEEE conference on computer vision and pattern recognition</i> , pages 4584–4593.	1511
1472		1512
1473		1513
1474		1514
1475		1515
1476	Ting Yu, Jun Yu, Zhou Yu, Qingming Huang, and Qi Tian. 2020. Long-term video question answering via multimodal hierarchical memory attentive networks. <i>IEEE Transactions on Circuits and Systems for Video Technology</i> , 31(3):931–944.	1516
1477		1517
1478		1518
1479		1519
1480		1520

1481	Youngjae Yu, Jongseok Kim, and Gunhee Kim. 2018.	Hang Zhang, Xin Li, and Lidong Bing. 2023b.	1536
1482	A joint sequence fusion model for video question	Videllama: An instruction-tuned audio-visual language	1537
1483	answering and retrieval. In <i>Proceedings of the European Conference on Computer Vision (ECCV)</i> , pages	model for video understanding. <i>arXiv preprint arXiv:2306.02858</i> .	1538
1484	471–487.		1539
1485			
1486	Youngjae Yu, Hyungjin Ko, Jongwook Choi, and Gunhee Kim. 2017.	Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2017.	1540
1487	End-to-end concept word detection	mixup: Beyond empirical risk minimization. <i>arXiv preprint arXiv:1710.09412</i> .	1541
1488	for video captioning, retrieval, and question answering.		1542
1489	In <i>Proceedings of the IEEE conference on computer vision and pattern recognition</i> , pages	Ziqi Zhang, Yaya Shi, Chunfeng Yuan, Bing Li, Peijin Wang, Weiming Hu, and Zheng-Jun Zha. 2020.	1543
1490	3165–3173.	Object relational graph with teacher-recommended learning for video captioning. In <i>Proceedings of the IEEE/CVF conference on computer vision and pattern recognition</i> , pages 13278–13288.	1544
1491			1545
1492	Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yuet-ing Zhuang, and Dacheng Tao. 2019.	Rui Zhao, Haider Ali, and Patrick Van der Smagt.	1546
1493	Activitynet-qa: A dataset for understanding complex web videos via	2017a. Two-stream rnn/cnn for action recognition in 3d videos. In <i>2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)</i> , pages	1547
1494	question answering. In <i>Proceedings of the AAAI Conference on Artificial Intelligence</i> , volume 33, pages	4260–4267. IEEE.	1548
1495	9127–9134.		
1496			
1497			
1498	Liangzhe Yuan, Nitesh Bharadwaj Gundavarapu, Long Zhao, Hao Zhou, Yin Cui, Lu Jiang, Xuan Yang, Menglin Jia, Tobias Weyand, Luke Friedman, et al. 2023.	Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. 2023.	1549
1499	Videoglot: Video general understanding evaluation of foundation models. <i>arXiv preprint arXiv:2307.03166</i> .	Learning video representations from large language models. In <i>Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition</i> , pages 6586–6597.	1550
1500			1551
1501			1552
1502			1553
1503			
1504	Yitian Yuan, Tao Mei, and Wenwu Zhu. 2019.	Zhou Zhao, Jinghao Lin, Xinghua Jiang, Deng Cai, Xiaofei He, and Yueling Zhuang. 2017b.	1554
1505	To find where you talk: Temporal sentence localization in	Video question answering via hierarchical dual-level attention network learning. In <i>Proceedings of the 25th ACM international conference on Multimedia</i> , pages 1050–1058.	1555
1506	video with attention based location regression. In		1556
1507	<i>Proceedings of the AAAI Conference on Artificial Intelligence</i> , volume 33, pages 9159–9166.		1557
1508			1558
1509	Sangdoo Yun, Dongyoon Han, Seong Joon Oh,	Zhou Zhao, Zhu Zhang, Shuwen Xiao, Zhou Yu, Jun Yu, Deng Cai, Fei Wu, and Yueling Zhuang. 2018.	1559
1510	Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo.	Open-ended long-form video question answering via adaptive hierarchical reinforced networks. In <i>IJCAI</i> , volume 2, page 8.	1560
1511	2019.		1561
1512	Cutmix: Regularization strategy to train strong		1562
1513	classifiers with localizable features. In <i>Proceedings of the IEEE/CVF international conference on computer vision</i> , pages 6023–6032.		1563
1514			1564
1515	Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu,	Zhou Zhao, Jinghao Lin, Xinghua Jiang, Deng Cai, Xiaofei He, and Yueling Zhuang. 2017b.	1565
1516	Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi.	Video question answering via hierarchical dual-level attention network learning. In <i>Proceedings of the 25th ACM international conference on Multimedia</i> , pages 1050–1058.	1566
1517	2021.		1567
1518	Merlot: Multimodal neural script knowledge		1568
1519	models. <i>Advances in Neural Information Processing Systems</i> , 34:23634–23651.		1569
1520	Kuo-Hao Zeng, Tseng-Hung Chen, Ching-Yao Chuang,	Yaoyao Zhong, Junbin Xiao, Wei Ji, Yicong Li, Weihong Deng, and Tat-Seng Chua. 2022.	1570
1521	Yuan-Hong Liao, Juan Carlos Niebles, and Min Sun.	Video question answering: Datasets, algorithms and challenges. <i>arXiv preprint arXiv:2203.01225</i> .	1571
1522	2017.		1572
1523	Leveraging video descriptions to learn video		1573
1524	question answering. In <i>Thirty-First AAAI Conference on Artificial Intelligence</i> .		
1525	Yawen Zeng, Da Cao, Shaofei Lu, Hanling Zhang, Jiao Xu, and Zheng Qin. 2022.	Luowei Zhou, Chenliang Xu, and Jason Corso. 2018a.	1574
1526	Moment is important: Language-based video moment retrieval via adversarial learning. <i>ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)</i> ,	Towards automatic learning of procedures from web instructional videos. In <i>Proceedings of the AAAI Conference on Artificial Intelligence</i> , volume 32.	1575
1527	18(2):1–21.		1576
1528			1577
1529			
1530			
1531	Bowen Zhang, Xiaojie Jin, Weibo Gong, Kai Xu, Zhao Zhang, Peng Wang, Xiaohui Shen, and Jiashi Feng. 2023a.	Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. 2018b.	1578
1532	Multimodal video adapter for parameter efficient video text retrieval. <i>arXiv preprint arXiv:2301.07868</i> .	End-to-end dense video captioning with masked transformer. In <i>Proceedings of the IEEE conference on computer vision and pattern recognition</i> , pages 8739–8748.	1579
1533			1580
1534			1581
1535			1582

1590 Linchao Zhu and Yi Yang. 2020. Actbert: Learning
1591 global-local video-text representations. In *Proced-
1592 ings of the IEEE/CVF conference on computer vision
1593 and pattern recognition*, pages 8746–8755.

1594 Yongqing Zhu and Shuqiang Jiang. 2019. Attention-
1595 based densely connected lstm for video captioning.
1596 In *Proceedings of the 27th ACM international con-
1597 ference on multimedia*, pages 802–810.

1598 Jiafan Zhuang, Zilei Wang, and Junjie Li. 2023. Video
1599 semantic segmentation with inter-frame feature fu-
1600 sion and inner-frame feature refinement. *arXiv
1601 preprint arXiv:2301.03832*.

1602 Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gok-
1603 berk Cinbis, David Fouhey, Ivan Laptev, and Josef
1604 Sivic. 2019. Cross-task weakly supervised learn-
1605 ing from instructional videos. In *Proceedings of the
1606 IEEE/CVF Conference on Computer Vision and Pat-
1607 tern Recognition*, pages 3537–3545.

Appendix

A Levels of Video-Language Understanding

Due to limited space, in this appendix, we provide a hierarchy which denotes the level of understanding within fundamental video-language tasks, *i.e.* text-video retrieval, video captioning, and videoQA.

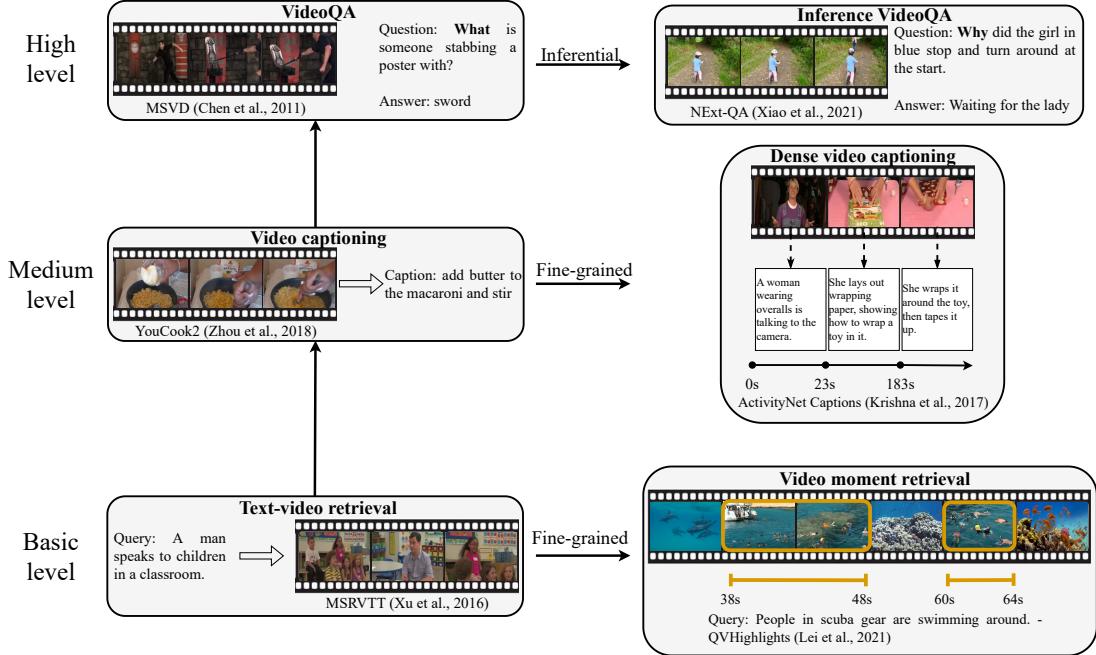


Figure 4: Level hierarchy of video-language understanding tasks.

B Examples of Video-Language Understanding tasks

In this appendix, we provide examples of video-language understanding tasks in Figure 5 and 6.

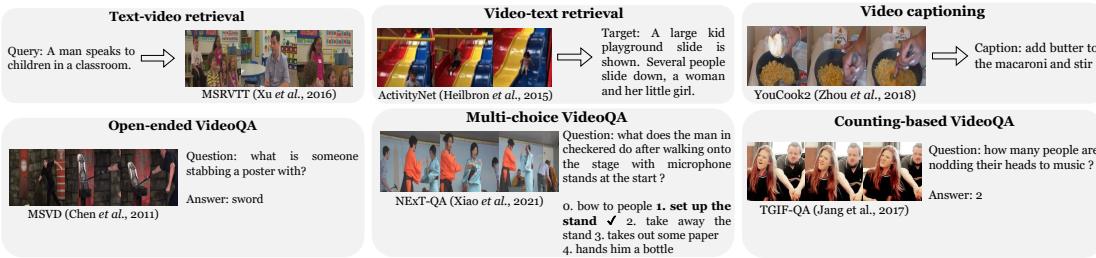


Figure 5: Illustration of text-video retrieval, video captioning, and video question answer (videoQA) tasks.

C Performance Analysis of Video-Language Understanding Architectures

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. Among Transformer-based architectures, the dual Transformer stands out as the most effective for text-video retrieval, adeptly associating global semantics of video and language modality. On the other hand, the stacked Transformer architecture excels at facilitating intra-modal and cross-modal interactions through its specialized unimodal and cross-modal encoders. These encoders are particularly efficient at correlating video content with the question in videoQA. Additionally, for video captioning, cross-modal encoder plays a crucial role in translating video content into textual

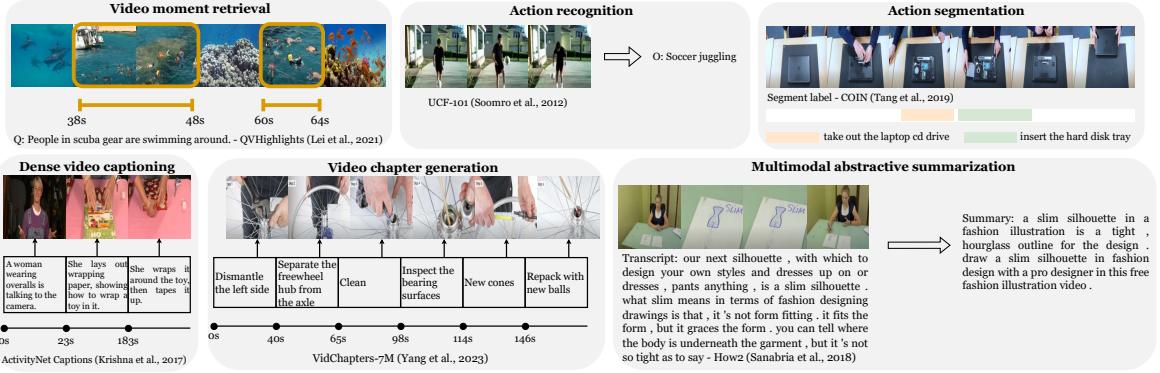


Figure 6: More illustration of video moment retrieval, action recognition, action segmentation, dense video captioning, video chapter generation, and multimodal abstractive summarization tasks.

descriptions. Recent LLM-augmented models have begun to outperform Transformer-based architectures in videoQA, signalling their potential as the next frontier in video-language understanding research.

D Analysis of Video-Language Understanding datasets

Due to page limit, details of the datasets for video-language understanding tasks are listed in Table 4. We categorize all the datasets according to their tasks that they support. While datasets for downstream tasks and fine-tuning have been consistently developed, those for pre-training emerge subsequently to the introduction of the Transformer architecture. Although pre-training and downstream video-language understanding datasets pursue different goals, they predominantly originate from the Internet. Regarding downstream datasets, more recent ones aim to present new technical challenges, such as evaluating reasoning and inference abilities (Xiao et al., 2021; Li et al., 2022a), or examining long-form modeling capacity of video-language understanding models (Mangalam et al., 2023).

Methods	Model architecture	Video	Text	R@1	R@5	R@10
VSE-LSTM (Kiros et al., 2014)		ConvNet/OxfordNet	GloVe-LSTM	3.8	12.7	17.1
C+LSTM+SA-FC7 (Torabi et al., 2016)		VGG	GloVe-LSTM	4.2	12.9	19.9
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
JSFusion (Yu et al., 2018)		RN	GloVe-LSTM	10.2	31.2	43.2
All-in-one (Wang et al., 2023a)	Shared TF	Linear	BT	37.9	68.1	77.1
VLM (Xu et al., 2021)	Shared TF	S3D	BT	28.1	55.5	67.4
DeCEMBERT (Tang et al., 2021)	Shared TF	RN	BT	17.5	44.3	58.6
ActBERT (Zhu and Yang, 2020)	Stacked TF	Faster-RCNN	BT	16.3	42.8	56.9
VIOLET (Fu et al., 2023)	Stacked TF	VS-TF	BT	37.2	64.8	75.8
VindLU (Cheng et al., 2023)	Stacked TF	ViT	BT	<u>48.8</u>	<u>72.4</u>	<u>82.2</u>
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
CLIP2TV (Gao et al., 2021)	Dual TF	ViT	CLIP-text	32.4	58.2	68.6
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	CIDEr
TA (Yao et al., 2015)	Pre-TF	Video: 3D-CNN	36.5	25.7	-
h-RNN (Yu et al., 2016)		Video: VGG	36.8	25.9	-
MFATT (Long et al., 2018)		Video: RN+C3D	39.1	26.7	-
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	-
MARN (Pei et al., 2019)		Video: RN	-	-	46.8
Res-ATT (Li et al., 2019)		Video: RN	37.0	26.9	40.7
DenseLSTM (Zhu and Jiang, 2019)		Video: VGG	38.1	27.2	42.8
VIOLET (Fu et al., 2023)		VS-TF	-	-	58.0
LAVENDER (Li et al., 2023b)	Stacked TF	VS-TF	-	-	57.4
VLAB (He et al., 2023)		EVA-G	54.6	33.4	74.9
UniVL (Luo et al., 2020)		S3D	41.8	28.9	50.0
MV-GPT (Seo et al., 2022)		ViViT	48.9	38.7	60.0
CLIP-DCD (Yang et al., 2022b)		ViT	48.2	30.9	64.8
DeCEMBERT (Tang et al., 2021)		RN	45.2	29.7	52.3
mPLUG-2 (Xu et al., 2023)		ViT	57.8	34.9	80.3

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. (2023)). 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
E-MN (Xu et al., 2017)	Pre-TF	VGG + C3D	GloVe-LSTM	30.4	26.7
QueST (Jiang et al., 2020)		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)		Linear	BT	44.3	47.9
LAVENDER (Li et al., 2023b)	Shared TF	VS-TF	BT	45.0	56.6
DeCEMBERT (Tang et al., 2021)	Stacked TF	RN	BT	37.4	-
VindLU (Cheng et al., 2023)	Stacked TF	ViT	BT	44.6	-
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. (2023)). 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
VIT (Huang et al., 2020)	YouTube-8M (Abu-El-Haija et al., 2016)	Manual	VMR	5.8K
VidChapters-7M (Yang et al., 2023a)	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
Alivolt-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).