Video Question Answering: Datasets, Algorithms and Challenges

Yaoyao Zhong[1,3*], **Junbin Xiao**[1,2*], **Wei Ji**[1,2*],
Yicong Li[1], **Weihong Deng**[3†], **Tat-Seng Chua**[1,2]
[1]National University of Singapore, Singapore [2]Sea-NExT Joint Lab, Singapore
[3]Beijing University of Posts and Telecommunications, Beijing, China
{zhongyaoyao,whdeng}@bupt.edu.cn, junbin@comp.nus.edu.sg,
liyicong@u.nus.edu, {jiwei,dcscts}@nus.edu.sg

Abstract

This survey aims to organize the recent advances in video question answering (VideoQA) and point towards future directions. We firstly categorize the datasets into: 1) normal VideoQA, multi-modal VideoQA and knowledge-based VideoQA, according to the modalities invoked in the question-answer pairs, and 2) factoid VideoQA and inference VideoQA, according to the technical challenges in comprehending the questions and deriving the correct answers. We then summarize the VideoQA techniques, including those mainly designed for Factoid QA (such as the early spatio-temporal attention-based methods and the recent Transformer-based ones) and those targeted at explicit relation and logic inference (such as neural modular networks, neural symbolic methods, and graph-structured methods). Aside from the backbone techniques, we also delve into specific models and derive some common and useful insights either for video modeling, question answering, or for cross-modal correspondence learning. Finally, we present the research trends of studying beyond factoid VideoQA to inference VideoQA, as well as towards the robustness and interpretability. Additionally, we maintain a repository, https://github.com/VRU-NExT/VideoQA, to keep trace of the latest VideoQA papers, datasets, and their open-source implementations if available. With these efforts, we strongly hope this survey could shed light on the follow-up VideoQA research.

1 Introduction

Recent years have witnessed a flourish of research in vision-language understanding (Xu et al., 2016; Chen et al., 2017; Antol et al., 2015; Chen et al., 2018; Jang et al., 2017), of which, video Question Answering (VideoQA) is one of the most prominent, given its promise to develop interactive AI to communicate with the dynamic visual world via natural languages. Despite the popularity, VideoQA remains one of the greatest challenges, because it demands the models to comprehensively understand the videos to correctly answer questions. The questions involve not only the recognition of visual objects, actions, activities and events, but also the inference of their semantic, spatial, temporal, and causal relationships (Xu et al., 2017; Jang et al., 2017; Shang et al., 2019, 2021; Yang et al., 2021b; Xiao et al., 2021, 2022a).

To tackle the challenges, techniques such as spatio-temporal attention (Jang et al., 2017), motion-appearance memory (Gao et al., 2018), and spatio-temporal or hierarchical graph models (Cherian et al., 2022; Xiao et al., 2022a) have been proposed and demonstrated their effectiveness on different VideoQA datasets. However, we find that the datasets, the defined challenges, and the corresponding algorithms are varied and a bit messy. There is a lack of a meaningful survey to categorize the datasets and to organize the technique developed, which seriously impedes the research.

Although a handful of recent works (Sun et al., 2021; Khurana and Deshpande, 2021; Patel et al., 2021) have tried to review VideoQA, they mostly follow an old-to-new fashion to summarize the literature and lack an effective taxonomy to classify them. In terms of the contents, these works focus merely on factoid questions and neglect the inference questions (see Fig. 1 for the difference). Furthermore, lots of recent new techniques (*e.g.*, pre-training and Transformer) are missing.

This paper thus gives a more comprehensive and meaningful survey to VideoQA, in the hope of learning from the past and shaping the future. Our contributions are as follows. (1) We provide a clear taxonomy to VideoQA. We can either classify existing VideoQA tasks into Factoid VideoQA and Inference VideoQA according to the fundamental challenges embodied in QAs, or classify them into normal VideoQA, Multi-modal VideoQA, and

[*] The first three authors contribute equally to this work.
[†] Corresponding authors.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 6439 - 6455
December 7-11, 2022

Knowledge-based VideoQA according to the multi-modal information invoked in the QAs. (2) We categorize existing VideoQA techniques as Memory, Transformer, Graph, Neural Modular Network, and Neural-Symbolic method. Along with the techniques, some meaningful insights are also included: attention modeling, cross-modal pre-training, hierarchical learning, multi-granular ensemble, and progressive reasoning. (3) We analyze existing methods from the perspective of the challenges encountered in the various VideoQA tasks and provide our prospects for future research.

2 VideoQA Task and Datasets

2.1 Problem Formulation

VideoQA is a task to predict the correct answer a^* based on a question q and a video V. There are mainly two types of tasks in VideoQA: multi-choice QA and open-ended QA.

For **multi-choice** QA, the models are presented with several candidate answers $\mathcal{A}_{mc}$ for each question and are required to pick the correct one $a^* = \mathcal{F}(a|q, \mathcal{V}, \mathcal{A}_{mc})$. For **open-ended** QA, the problem can be classification (the most popular), generation (word-by-word) and regression (for counting) depending on the specific datasets. Specifically, open-ended QA is popularly set as a multi-class classification problem which requires the models to classify a video-question pair into a predefined global answer set $\mathcal{A}_{oe}$: $a^* = \mathcal{F}(a|q, \mathcal{V})$ where $a \in \mathcal{A}_{oe}$. Open-ended QA can also be formulated as a generation problem, which might have more practical use and receiving increasing attention. Usually the answer is denoted as $a = (a_1, a_2, ..., a_t, ..., a_M)$ of length M, where a_t is the t-th word; and the model is required to predict the next word a_t in the vocabulary set $\mathcal{W}$: $a_t^* = \mathcal{F}(a_t|q, \mathcal{V}, (a_1, a_2, ..., a_{t-1}))$, where $a_t \in \mathcal{W}$. For the counting task, which is defined as an open-ended question about counting the number of repetitions of an action (Jang et al., 2017), it is formulated as an regression problem, requiring the model to compute an integer-valued answer to be close to the ground truth.

Compared with open-ended QA, multi-choice QA is typically defined to study beyond factoid QA to inference QA (Xiao et al., 2021; Wu et al., 2021a), as it dispenses with the generation and evaluation of natural languages.

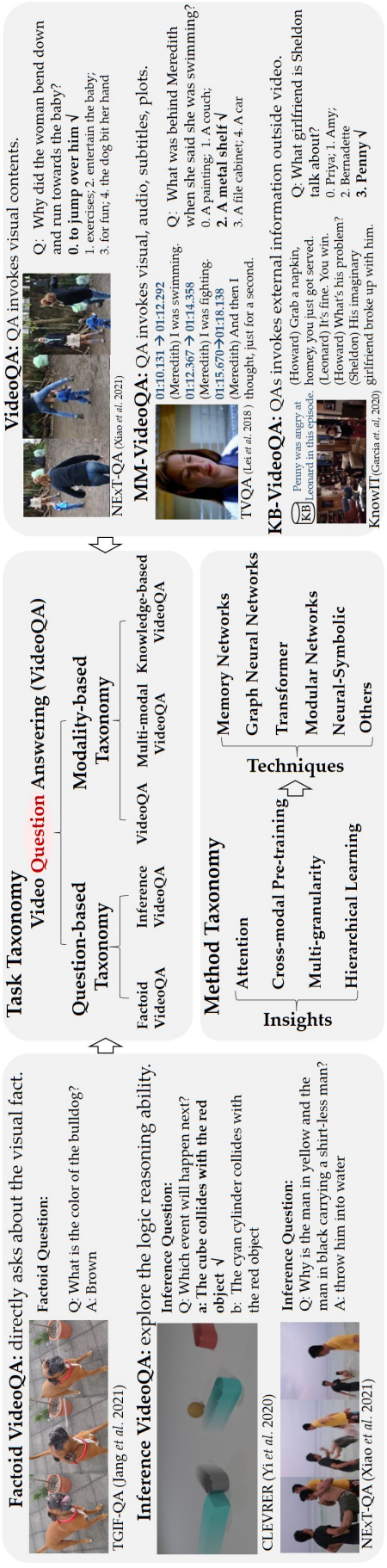

Figure 1: Illustration of the taxonomy. The taxonomy covers not only Factoid and Inference VideoQA in terms of understanding level, but also VideoQA, Multi-modal VideoQA, and Knowledge-based VideoQA in terms of the multi-modal information invoked in QAs to better analyze the challenges and help to uncover the future focus.

2.2 Evaluation Metrics

Accuracy. For multi-choice QA and open-ended QA (classification), accuracy is defined based on the entire testing question set Q, given by:

$$acc = \frac{1}{|Q|} \sum_{q \in Q} \mathbf{I}\left[a^* = a\right], \tag{1}$$

where Q represents the number of QA pairs, and $\mathbf{I}[\cdot]$ is an indicator function (1 only if $a^* = a$ and 0 otherwise). Similarly, for open-ended QA (word-by-word generation) (Zhao et al., 2017b, 2018), accuracy is defined as:

$$acc = \frac{1}{|Q|} \sum_{q \in Q} \frac{1}{M} \sum_{i=1}^{L} \mathbf{I}\left[a_i^* = a_i\right], \tag{2}$$

where L denotes the length of the shorter answer.

WUPS. The WUPS is the soft measure of accuracy by taking into account word synonyms. It is based on the WUP score (Wu and Palmer, 1994) to evaluate the quality of the generated answer (Zhao et al., 2017b, 2018; Xiao et al., 2021). The WUP measures word similarity based on WordNet (Fellbaum, 1998). WUPS score with the threshold γ is defined as,

$$WUPS = \frac{1}{|Q|} \sum_{q \in Q} \min\{ \prod_{a_i \in A} \max_{a_j^* \in A^*} WUP_\gamma(a_i, a_j^*), \prod_{a_i^* \in A^*} \max_{a_j \in A} WUP_\gamma(a_i^*, a_j)\}, \tag{3}$$

where WUP score is given by,

$$WUP_\gamma(a_i, a_j^*) = \begin{cases} WUP(a_i, a_j^*) & WUP(a_i, a_j^*) \geqslant \gamma \\ 0.1 WUP(a_i, a_j^*) & WUP(a_i, a_j^*) < \gamma \end{cases}. \tag{4}$$

where the parameter γ is dataset-specific.

Mean $\mathcal{L}_2$ loss. For the repetition count task (Jang et al., 2017), the mean $\mathcal{L}_2$ loss is defined based on the entire testing question set Q:

$$\mathcal{L}_2 = \frac{1}{|Q|} \sum_{q \in Q} (a^* - a)^2, \tag{5}$$

in which a and a* are predicted and ground-truth numbers respectively.

The evaluation metrics mainly serve for different task settings, while there are also some novel and diagnostic ones (Gandhi et al., 2022; Li et al., 2022b; Castro et al., 2022a) that may be helpful for robustness and interpretation of VideoQA models.

2.3 Datasets

VideoQA can be understood from different perspectives, since the aim is to gain multi-view and multi-grained understanding of videos under the guidance of specific questions.

Modality-based Taxonomy. According to the data modality invoked in the questions and answers, VideoQA can be classified into normal VideoQA, multi-modal VideoQA (MM VideoQA), and knowledge VideoQA (KB VideoQA). Normal VideoQA only invokes visual resources to understand the question and to derive the correct answer. It emphasizes visual understanding of the video elements and reasoning of their relations. Usually, the videos are short and are typically user-generated on social platforms. Different from normal VideoQA, MM VideoQA often involves other resources aside from visual contents, such as subtitles/transcripts and text plots of movies (Tapaswi et al., 2016) and TV shows (Lei et al., 2018). MM VideoQA mainly challenges multi-modal information fusion and long video story understanding. Finally, KB VideoQA (Garcia et al., 2020) demands external knowledge distillation from explicit knowledge bases or commonsense reasoning (Fang et al., 2020). Different from MM VideoQA, KB VideoQA provides a global knowledge base for the whole dataset, instead of giving paired "knowledge" for each question. For better understanding of the three kinds of VideoQA, we show typical examples in Figure 1 (right).

Question-based Taxonomy. According to the type of question (or the challenges posted in the questions), VideoQA can be classified into factoid VideoQA and inference VideoQA. A factoid question directly asks about the visual fact, such as the location (`where is`), objects/attributes (`who/what (color) is`), and invokes little relations to understand the questions and infer the correct answers. Factoid QA emphasizes the holistic understanding of the questions and the recognition of the visual elements. In contrast, inference VideoQA aims to explore the logic and knowledge reasoning ability in dynamic scenarios. It features various relationships between the visual facts. Though rich in relation types, VideoQA emphasizes temporal (`before/after`) and causal (`why/How/what if`) relationships that feature temporal dynamics, as emphasized by recent works (Zadeh et al., 2019; Yi et al., 2020; Xiao et al., 2021; Li et al., 2022b).

Datasets Analysis. The timeline of some established VideoQA datasets is shown in Figure 2. We categorize all the datasets according to our defined taxonomy in Table 1 and their details are

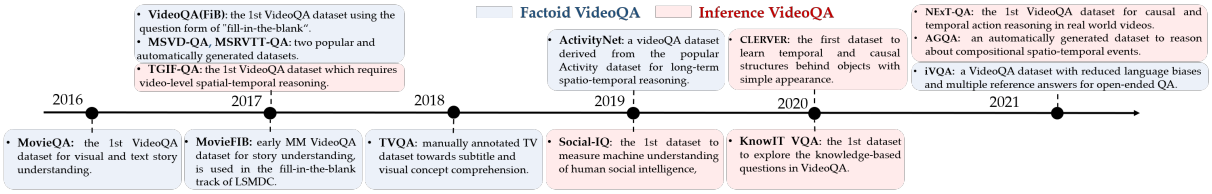

Figure 2: Timeline of established VideoQA datesets. The items above the timeline show the normal VideoQA datasets. The Multi-modal VideoQA and Knowledge-based VideoQA are listed below the timeline. Blue and red colors represent datasets focused on Factoid VideoQA and Inference VideoQA.

Table 1: VideoQA datasets in the literature.

	Factoid VideoQA	Inference VideoQA
VideoQA	VideoQA(FiB) (Zhu et al., 2017), VideoQA (Zeng et al., 2017), MSVD-QA (Xu et al., 2017), MSRVTT-QA (Xu et al., 2017), YouTube2Text-QA (Zhao et al., 2017a), MarioQA (Mun et al., 2017), EgoVQA (Fan, 2019), ActivityNet-QA (Yu et al., 2019), iVQA (Yang et al., 2021a), ASRL-QA (Sadhu et al., 2021), Charades-SRL-QA (Sadhu et al., 2021), FIBER (Castro et al., 2022b), WildQA (Castro et al., 2022a)	TGIF-QA (Jang et al., 2017), SVQA (Song et al., 2018), V2C-QA (Fang et al., 2020), CLEVRER (Yi et al., 2020), SUTD-TrafficQA (Xu et al., 2021), AGQA (Grunde-McLaughlin et al., 2021), AGQA 2.0 (Gandhi et al., 2022), VQuAD (Gupta et al., 2022), STAR (Wu et al., 2021a), NExT-QA (Xiao et al., 2021),Causal-VidQA (Li et al., 2022b)
MM VideoQA	MovieQA (Tapaswi et al., 2016), MovieFIB (Maharaj et al., 2017), PororoQA (Kim et al., 2017), TVQA (Lei et al., 2018), TVQA+ (Lei et al., 2020), LifeQA (Castro et al., 2020), How2QA (Li et al., 2020), Env-QA (Gao et al., 2021), Pano-AVQA (Yun et al., 2021), DramaQA (Choi et al., 2021), MUSIC-AVQA (Li et al., 2022a), AVQA (Yang et al., 2022b)	Social-IQ (Zadeh et al., 2019)
KB VideoQA	/	PsTuts-VQA (Zhao et al., 2020), KnowIT VQA (Garcia et al., 2020), KnowIT-X VQA (Wu et al., 2021b), NEWSKVQA (Gupta and Gupta, 2022)

listed in Table A1 (see Appendix). VideoQA and MM VideoQA almost appear simultaneously, and have been studied separately by the community. Despite the unique challenges of MM VideoQA in reasoning on multiple modalities (Kim et al., 2020), algorithms targeting VideoQA and MM VideoQA share similar spirits. Modality-based taxonomy stems from research preference for video domains. While question-based taxonomy is affected more by the methodological considerations, since the recently proposed Inference VideoQA brings new technical challenges, which is driving artificial intelligence towards new heights, not just limited to learning the correlations in data.

2.4 Main Framework

As shown in Figure 3, a common framework comprises four parts: video encoder, question encoder, cross-modal interaction, and answer decoder. The video encoder often encodes raw videos by jointly extracting frame appearance and clip motion features. Recent works also show that object-level visual and semantic features (*e.g.*, category and attribute labels) are important. These features are usually extracted with pre-trained 2D or 3D neural networks, as summarized in Table 2. Question encoder extracts token-level representation, such as GloVe and BERT features (Kenton and Toutanova, 2019). Then, the sequential data of vision and language can be further processed by sequential models (*e.g.*, RNN, CNN, and Transformer) for the convenience of cross-modal interaction, which will be detailed further. For multi-choice QA, the answer decoder can be a 1-way classifier to select the correct answer from the provided multiple choices. For open-ended QA, it can be either an n-way classifier to select an answer from a pre-defined global answer set, or a language generator to generate an answer word by word. The video and language encoders can be pre-trained or more recently end-to-end fine-tuned (Lei et al., 2021).

2.5 Challenges and Meaningful Insights

Unique Challenges. Compared with ImageQA (Lu et al., 2016; Anderson et al., 2018), VideoQA is much more challenging because of the spatio-temporal nature of videos (Xiao et al., 2020, 2021). Thus, a simple extension of existing ImageQA techniques to answer queries of videos will lead to sub-optimal results. Compared with other video tasks, question-answering requires a comprehensive understanding of videos in different aspects and granularity, such as from fine-grained to coarse-grained in both temporal and spatial domains, and from factoid questions to inference questions. To tackle the challenges, a lot of research efforts have been developed on cross-modal interaction, which aims to gain understanding of videos under the guidance of questions. We summarize some common and meaningful insights as follows:

Attention. Attention is a human-inspired mechanism that locates the important part of the input

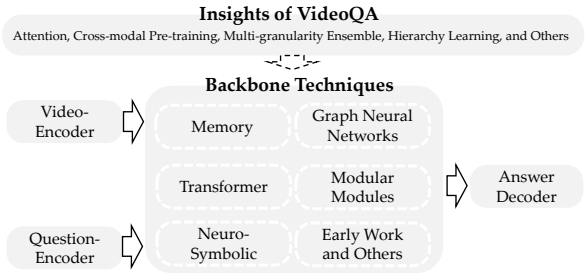

Figure 3: A common solution framework for VideoQA. It includes: a video encoder, a question encoder, a cross-modal interaction, and an answer decoder. Some common insights are also involved in the model design.

and selectively focuses on useful information. In VideoQA, to attend to a specific part of videos in both spatial and temporal dimensions, *temporal attention* and *spatial attention* are widely used. *Self-attention* has a good ability to model long-range dependencies, and can be used in intra-modal modeling, such as temporal information in the video and global dependencies of questions. *Co-attention* (*Cross-modal attention*) can attend to both relevant and critical multi-modal information, such as the question-guided video representation and video-guided question representation.

Cross-modal Pre-training and Fine-tuning. With the development of unified network architectures (*e.g.*, Transformer (Vaswani et al., 2017)) that can well handle visual and linguistic data, cross-modal pre-training can make full use of the semantic information from noisy but web-scale vision-text data (Radford et al., 2021). The learned model can be transferred to downstream vision-language tasks by fine-tuning on small-scale manually annotated datasets with strong supervision. Currently, the research in this paradigm lies in four aspects: large-scale data collection, proxy task definition, Transformer-style model design, and downstream adaption. We recommend the readers to read the latest survey (Chen et al., 2022) for details.

Multi-Granularity Ensemble. Questions are diverse and unconstrained, and may demand video information of different granularities for answers (Xiao et al., 2022a). To gain rich information and answer the varied questions, the multi-granularity ensemble is essential. Specifically, the multi-granularity ensemble exists in both the text domain and the vision domain of both spatial and temporal dimensions. In the text domain, word-, phrase- and sentence-level feature representations are coordinated to achieve both fine- and coarse-grained information modeling. In the vision domain, region-, trajectory-, frame- and clip-level feature representations can complement each other to achieve comprehensive video understanding.

Hierarchical Learning. Considering that the video elements and their textual correspondences in QA pairs are in different abstraction levels, hierarchical learning aims to organize multi-modal representation from low-level to high-level, and from local to global (Le et al., 2020; Dang et al., 2021; Xiao et al., 2022a). Specifically, linguistic concepts are analyzed from word to sentence. Similarly, video elements are processed from objects to actions, activities, and global events. Compared with the multi-granularity ensemble, hierarchical learning processes the multi-granular information progressively. It gradually reasons and aggregates the low-level, local visual information into the high-level, global video representation. Thus, hierarchical learning can better reflect the structure and relationship of video elements and accomplish question answering hierarchically.

Others. Aside from the above, multi-step reasoning (Wang et al., 2021; Mao et al., 2022) and causal discovery (Li et al., 2022d) also demonstrate the effectiveness. Most importantly, these insights are not mutually exclusive; they can be coordinated in a single model for good performance.

3 Algorithms

3.1 Methods

Early Attention-based Works. (Zeng et al., 2017) try to directly apply element-wise multiplication to fuse the global video and question representations for answer prediction. Additionally, it demonstrates the advantage of a simple temporal attention. Attention is also explored in more complex scenarios in conjunction with various other ideas, such as multi-granularity ensemble (Xu et al., 2017) and hierarchical learning (Zhao et al., 2017a). In particular, (Jang et al., 2017) propose a dual-LSTM based approach with spatial and temporal attention mechanisms, which can focus better on critical frames in a video and critical regions in a frame. (Xu et al., 2017) refine attention over both frame-level and clip-level visual features, conditioned with both the coarse-grained question feature and fine-grained word feature. (Zhao et al., 2017a) propose hierarchical dual-level attention networks (DLAN) to learn the question-aware video representations with word-level and question-level attention based on appearance and motion.

Despite the ability to attend to video frames and clips, these works rely on RNN for history information modeling, which has later been shown to be weak in capturing long-term dependency.

Memory Networks. Memory networks can cache sequential inputs in memory slots and explicitly utilize even far early information. Memory especially receives attention in *long video story understanding*, such as movies and TV-Shows. Because the QAs in these VideoQA tasks not only involve the understanding of visual contents, but also the long stories they convey.

(Tapaswi et al., 2016) first incorporate and modify the memory network (Sukhbaatar et al., 2015) into VideoQA, to store video and subtitle features in the memory bank. To enable memory read and write operations with high capacity and flexibility, (Na et al., 2017) design a memory network with multiple convolution layers. Considering dual-modal information in the movie story, (Kim et al., 2019) introduce a progressive attention mechanism to progressively prune out irrelevant temporal parts in the memory bank for each modality, and adaptively integrate outputs of each memory.

Memory has also been explored in *normal VideoQA*. (Gao et al., 2018) propose a two-stream framework (CoMem) to deal with motion and appearance information with a co-memory attention module, introducing multi-level contextual information and producing dynamic fact ensembles for diverse questions. Considering that CoMem synchronizes the attentions detected by appearance and motion features, it could thus generate incorrect attention, (Fan et al., 2019) further introduce a heterogeneous external memory module (HME) with attentional read and write operations to integrate the motion and appearance features and learn the spatio-temporal attention simultaneously.

Transformer. Transformer (Vaswani et al., 2017) has a good ability to model long-term relationships and has demonstrated promising performance for modeling multi-modal vision-language tasks such as VideoQA, with pre-training on large-scale datasets (Zhu and Yang, 2020). Motivated by the success of Transformer, (Li et al., 2019) first introduce the architecture of Transformer *without pre-training* to VideoQA (PSAC), which consists of two positional self-attention blocks to replace LSTM, and a video-question co-attention block to simultaneously attend both visual and textual information. (Yang et al., 2020) and (Urooj et al., 2020) incorporate the pre-trained *language-based* Transformer (BERT) (Kenton and Toutanova, 2019) to movie and story understanding, which requires more modeling on languages like subtitles and dialogues. Both works process each of the input modalities such as video and subtitles, with question and candidate answer, respectively, and lately fuse several streams for the final answer.

More recently, (Lei et al., 2021) apply the *image-text pre-trained Transformer* for cross-modal pre-training and fine-tune it for downstream video-text tasks, such as VideoQA. (Yang et al., 2021a) train a VideoQA model, based on a large-scale dataset, with 69M video-question-answer triplets, using contrastive learning between a multi-modal video-question Transformer and an answer Transformer. This *video-text pre-trained Transformer* can be further fine-tuned on other downstream VideoQA tasks, which shows the benefits of task-specific pre-training for the target VideoQA task. Furthermore, (Zellers et al., 2021) train a cross-modal Transformer (MERLOT) in a label-free, self-supervised manner, based on 180M video segments with image frames and words. Similar to MERLOT, VIOLET (Fu et al., 2021) is another end-to-end *video-text pre-trained* Transformer model but with more advanced video encoder and proxy tasks.

While the aforementioned Transformer-style models have demonstrated strong performances on popular Factoid VideoQA datasets (refer to our analysis in Sec. 3.2), recent works (Buch et al., 2022; Xiao et al., 2022b) reveal that their performance are weak in answering questions that emphasize visual relation reasoning, especially the temporal and causal relations which feature video dynamics. Furthermore, their demands on large-scale video data for pre-training and the lack of explanability largely prevent their popularity. Such weaknesses call for more future efforts in developing foundation models for fine-grained video reasoning, and simultaneously, with less computation resources and better interpretability.

Graph Neural Networks. Graph-structured techniques (Kipf and Welling, 2017; Zhang et al., 2022) are recently more favoured for improving the reasoning ability of VideoQA models, especially when Inference VideoQA draws attention to the community (Jang et al., 2017; Xiao et al., 2021). HGA (Jiang and Han, 2020), and more recent works, B2A (Park et al., 2021) and DualVGR (Wang et al., 2021) build the graphs based

on coarse-grained video segments. Yet, they incorporate both intra- and inter-modal relationship learning and achieve good performances. To gain object-level information, (Huang et al., 2020) build the graph (LGCN) based on objects represented by their appearance and location features. They model the interaction between objects related to questions with GNN (Kipf and Welling, 2017).

Considering that the video elements are hierarchical in semantic space, (Liu et al., 2021a), (Peng et al., 2021) and (Xiao et al., 2022a) incorporate hierarchical learning into graph networks. Specifically, (Liu et al., 2021a) propose a graph memory mechanism (HAIR), to perform relational vision-semantic reasoning from object level to frame level; (Peng et al., 2021) concatenate different-level graphs, that is, object-level, frame-level, and clip-level, progressively to learn the visual relations (PGAT). (Xiao et al., 2022a) propose a hierarchical conditional graph model (HQGA) to weave together visual facts from low-level entities to higher-level video elements through graph aggregation and pooling, to enable vision-text matching at multi-granularity levels. To leverage the semantics of the 3D scene, (Cherian et al., 2022) transfer the video frames to a 2.5D (pseudo-3D) scene graph and then split it into static and dynamic sub-graphs, allowing the pruning of redundant detections.

With a good ability for information communication, graph architectures have shown promising results on inference VideoQA. Nonetheless, the emphasis and difficulty lie in how to skillfully design the graph structure for video representation.

Modular Networks. (Le et al., 2020) find that most VideoQA models design tailor-made network architectures. They point out such hand-crafted architectures are inflexible in dealing with varied data modality, video length and question types. Therefore, they design a reusable neural unit - Conditional Relation Network (CRN), which captures the relations of input features given the global context and encapsulates them hierarchically to form networks. Such a constituted architecture has shown better generalization ability and flexibility in handling different types of questions. Following similar design philosophy, (Dang et al., 2021) and (Xiao et al., 2022a) design the spatio-temporal graph and conditional graph respectively as neural building blocks. The neural building blocks are hierarchically stacked to achieve good reasoning performances. While the above works aim for repeating a single module for videoQA. Recently, (Qian et al., 2022) design multiple modules tailored for compositional video question-answering (Grunde-McLaughlin et al., 2021), and has also demonstrated success. Overall, modular networks are of improved flexibility and transparency. Nonetheless, they either lack explicit logic for reasoning (Le et al., 2020; Dang et al., 2021; Xiao et al., 2022a), or can only handle questions that can be parsed into pre-defined subtasks of limited scope.

Neural-Symbolic. (Yi et al., 2020) point out two essential elements for causal reasoning in VideoQA are object-centric video representation that is aware of the temporal and causal relations between the objects and events, and a dynamics model that is able to predict the object dynamics under unobserved or counterfactual scenarios. Motivated by the neural-symbolic method in ImageQA (Yi et al., 2018), (Yi et al., 2020) propose the NS-DR model, which extracts object-level representation with a video parser, turns a question into a functional program, extracts and predicts the dynamic scene of the video with a dynamics predictor, and runs the program on the dynamic scene to obtain an answer. NS-DR aims to combine neural nets for pattern recognition and dynamics prediction, and symbolic logic for causal reasoning. It achieves significant gain on the explanatory, predictive, and counterfactual questions on the synthetic object dataset (Yi et al., 2020). (Chen et al., 2021) and (Ding et al., 2021) promote further progress.

Despite the good reasoning ability of Neural-Symbolic methods on synthetic datasets, they are currently hard to be applied in unconstrained video with open-form natural questions.

Others. There are also *flexibly designed networks* to address specific problems. For example, (Kim et al., 2020) propose a framework that first detects a specific temporal moment from moments of interest candidates for temporally-aligned video and subtitle using pre-defined sliding windows, and then fuses information based on the localized moment using intra-modal and cross-modal attention mechanisms. Due to their focuses on specific purposes, the question remains on whether these networks can be generalized to other VideoQA tasks.

Studies are also conducted in terms of *input information*. (Falcon et al., 2020) explore several *data augmentation* techniques to prevent overfitting with only small-scale datasets. (Kim et al., 2021a) point out existing works suffer from signifi-

Table 2: Performance on Factoid VideoQA tasks. (Att: Attention, MG: Multi-Granularity, HL: Hierarchical Learning, CM-PF: Cross-modal Pre-training and Fine-tuning, Mem: Memory, GNN: Graph Neural Networks, MN: Modular Networks, TF: Transformer. RN: ResNet at frame-level, RX(3D): 3D ResNeXt at clip-level, RoI: Region-of-interest features from Faster R-CNN, GV: GloVe, BT: BERT, VG: Visual Genome (Krishna et al., 2017) , YT-T: Youtube-Temporal-180M (Zellers et al., 2021) , Web: WebVid2M (Bain et al., 2021) , CC: Conceptual Captions-3M (Sharma et al., 2018) . ViT (Dosovitskiy et al., 2020) and VSwin (Liu et al., 2021b) are Transformer-style visual encoders. Attention is found in all methods, but we omit it for those methods that do not emphasize attention.)

Methods	Techniques & Insights	Encoder		Pre-training Dataset	TGIF-QA (Frame-QA)	MSVD -QA	MSRVTT -QA
		Video	Text				
STVQA(Jang et al., 2019)	Att	RN, Flow	GV	/	52.0	/	/
PSAC(Li et al., 2019)	Att	RN	GV	/	55.7	/	/
QueST(Jiang et al., 2020)	Att	RN, C3D	GV	/	59.7	36.1	34.6
CoMem(Gao et al., 2018)	Mem	RN, Flow	GV	/	51.5	/	/
HME(Fan et al., 2019)	Mem	RN, VGG, C3D	GV	/	53.8	33.7	33.0
LGCN(Huang et al., 2020)	GNN	RN, RoI	GV	/	56.3	34.3	/
HGA(Jiang and Han, 2020)	GNN	RN, VGG, C3D	GV	/	55.1	34.7	35.5
B2A(Park et al., 2021)	GNN, MG	RN, RX(3D)	GV	/	57.5	37.2	36.9
HAIR(Liu et al., 2021a)	GNN, Mem, HL	RoI	GV	/	60.2	37.5	36.9
MASN(Seo et al., 2021a)	GNN	RN, I3D, RoI	GV		59.5	38.0	35.2
DualVGR(Wang et al., 2021)	GNN	RN, RX(3D)	GV	/	/	39.0	35.5
PGAT(Peng et al., 2021)	GNN, MG, HL	RN, RX(3D), RoI	GV	/	61.1	39.0	38.1
HCRN(Le et al., 2020)	MN, HL	RN, RX(3D)	GV	/	55.9	36.1	35.6
HOSTR(Dang et al., 2021)	MN, GNN, HL	RN, RX(3D), RoI	GV	/	58.2	39.4	35.9
HQGA(Xiao et al., 2022a)	MN, GNN, HL, MG	RN, RX(3D), RoI	BT	/	61.3	41.2	38.6
MHN(Peng et al., 2022)	TF, HL, MG	RN, RX(3D)	GV	/	58.1	40.4	38.6
VGT(Xiao et al., 2022b)	TF, GNN	RN, RoI	BT	/	61.6	/	39.7
ClipBERT(Lei et al., 2021)	TF, CM-PF	RN (E2E)	BT	VG&COCO	60.3	/	37.4
CoMVT(Seo et al., 2021b)	TF, CM-PF	S3D	BT	HowTo100M	/	42.6	39.5
VQA-T(Yang et al., 2021a)	TF, CM-PF	S3D	BT	H2VQA69M	/	46.3	41.5
SiaSRea(Yu et al., 2021)	TF, GNN, CM-PF	RN (E2E)	BT	VG&COCO	60.2	45.5	41.6
MERLOT(Zellers et al., 2021)	TF, CM-PF	ViT(E2E)	BT	YT-T & CC	**69.5**	/	43.1
VIOLET(Fu et al., 2021)	TF, CM-PF	VSwin (E2E)	BT	Web&YT-T&CC	68.9	**47.9**	**43.9**

Table 3: Performance on Inference VideoQA tasks. For the counting (Cnt) task in TGIF-QA, value of mean square error (MSE) is reported for evaluation.

Methods	Techniques & Insights	NExT-QA		TGIF-QA		
		Val.	Test	Act	Tran.	Cnt
STVQA(Jang et al., 2017)	Att	47.9	47.6	62.9	69.4	4.22
CoMem(Gao et al., 2018)	Mem	48.0	48.5	68.2	74.3	4.10
HME(Fan et al., 2019)	Mem	48.7	49.2	73.9	77.8	4.02
HCRN(Le et al., 2020)	MN, HL	48.2	48.9	75.0	81.4	3.82
HGA(Jiang and Han, 2020)	GNN, HL	49.7	50.0	75.4	81.0	4.09
MASN(Seo et al., 2021a)	GNN	/	/	84.4	87.4	3.75
MHN(Peng et al., 2022)	TF, HL, MG	/	/	83.5	90.2	**3.57**
IGV(Li et al., 2022d)	GNN, Causal	51.0	51.3	/	/	/
HQGA(Xiao et al., 2022a)	MN, GNN, HL, MG	51.4	51.8	76.9	85.6	/
P3D-G(Cherian et al., 2022)	GNN, TF, HL	53.4	/	/	/	/
ATP(Buch et al., 2022)	TF	54.3	/	/	/	/
VGT(Xiao et al., 2022b)	TF,GNN	**55.0**	**53.7**	**95.0**	**97.6**	/
ClipBERT(Lei et al., 2021)	TF, CM-PF	/	/	82.8	87.8	/
SiasRea(Yu et al., 2021)	TF, CM-PF, GNN	/	/	79.7	85.3	/
MERLOT(Zellers et al., 2021)	TF, CM-PF	/	/	94.0	96.2	/
VIOLET(Fu et al., 2021)	TF, CM-PF	/	/	92.5	95.7	/
Human	/	**88.4**	/	/	/	/

cant computational complexity and insufficient representation capability and they introduce VideoQA features obtained from *coded video bit-stream* to address the problem. To overcome spurious visual-linguistic correlations, (Li et al., 2022d,c) explore robust and trustworthy grounding framework from causal theory, which is promising to enhance the SOTA models' accuracy and trustability.

3.2 Performance Analysis

We analyze the advanced methods for Factoid VideoQA in Table 2 and Inference VideoQA in Table 3 based on the results reported on popular VideoQA benchmarks. Apart from normal VideoQA, advanced methods for MM VideoQA and KB VideoQA are also summarized in Table 4.

Table 2 reveals that the cross-modal pre-trained Transformer-style models can achieve superior performance for factiod QA than others. By focusing on methods without pre-training, graph-structured techniques are the most popular and have also shown great potential. It would be interesting to explore cross-modal pre-training of graph models for VideoQA. Besides, hierarchical learning and fine-grained object features usually help to improve performances. In addition to the datasets given in Table 2, the recent iVQA (Yang et al., 2021a) dataset has also received increasing attention, and we believe it could be a more effective dataset towards open-ended VideoQA for its high quality.

Inference VideoQA is a nascent task that challenges mainly visual relation reasoning of video information. It also receives increasing attention. Graph-structured techniques, causal discovery, and hierarchical learning have shown promising performance (see Table 3). Notably, we find that cross-modal pre-training and fine-tuning not only achieves good performance on factoid VideoQA, but also significantly improves the results on inference VideoQA. Particularly, the accuracies of reasoning tasks on TGIF-QA reach unprecedentedly high. This dataset is likely not challenging enough and has serious language bias as revealed by recent studies (Peng et al., 2021; Piergiovanni et al., 2022; Xiao et al., 2022b). In contrast, NExT-QA is much more challenging; it emphasizes causal and temporal relation reasoning between multiple objects in real-world videos. Table 3 shows that SOTA methods still struggle on NExT-QA. As such,

Table 4: Performance on MM VideoQA and KB VideoQA tasks. For TVQA, we report results on test-public data split. (ts: Timestamp Annotation.)

Methods	Techniques & Insights	TVQA		TVQA+	KnowIT VQA
		w/o ts	w/ ts		
PAMN(Kim et al., 2019)	Mem	66.8	/	/	/
STAGE(Lei et al., 2020)	Att	**70.2**	/	74.8	/
HCRN(Le et al., 2021)	MN, HL	66.1	71.3	/	/
MSAN(Kim et al., 2020)	Att	/	71.1	/	/
BERT-VQA(Kenton and Toutanova, 2019)	TF	/	73.6	/	/
MMFT-BERT(Urooj et al., 2020)	TF	/	72.9	/	/
ROLL(Garcia and Nakashima, 2020)	TF	/	/	69.6	71.5
RHA(Li et al., 2021a)	GNN, HL	/	/	73.4	/
SPCR(Kim et al., 2021b)	Att	/	76.2	76.2	/
V2T(Engin et al., 2021)	TF	/	/	/	**78.1**
MERLOT(Zellers et al., 2021)	TF, CM-PF	/	**78.7**	**80.9**	/
Human	/	**89.4**	**91.5**	**90.5**	/

NExT-QA could be a more effective benchmark for visual reasoning of realistic video contents under natural language instructions. Additionally, NExT-QA also contains open-ended QA task that provide ample challenge for existing research.

MM and KB VideoQA require models to locate and perform reasoning in all heterogeneous modalities for answering the question. Similar to normal VideoQA, MM VideoQA also benefits from advanced networks and large-scale datasets. However, it is worth noting that modality shifting ability is essential (Kim et al., 2020; Engin et al., 2021).

4 Future Direction

From Recognition to Reasoning. Advanced neural network models excel at recognizing objects, attributes and even actions in visual data. Thus, answering the questions like "what is" is no longer the core of VideoQA. To enable more meaningful and in-depth human-machine interaction, it is urgent to study the casual and temporal relations between objects, actions, and events (Xiao et al., 2021). Such problems feature *video*-level understanding and demand inference ability for question answering. The focus on inference questions promotes research towards the core of human intelligence, which could be one of the "north stars" towards groundbreaking works (Fei-Fei and Krishna, 2022).

Knowledge VideoQA. To answer the questions that are beyond the visual scene, it is of crucial importance to inject knowledge into the reasoning stage (Jin et al., 2019; Garcia et al., 2020; Zhuang et al., 2020). Knowledge incorporation can not only greatly extend the scope of questions that can be asked about videos, but also enable the exploration of more human-like inference. Because we humans are natural to answer questions that may involve commonsense (Fang et al., 2020) or domain-specific knowledge (Xu et al., 2021; Gao et al., 2021). Reasoning with knowledge and diagnosing the retrieved knowledge for a specific question will help to enhance the model's interpretability and trustability. It will also serve as important groundwork for the future multi-modal conversation systems (Nie et al., 2019; Li et al., 2022e).

Cross-modal Pre-training and Fine-tuning. Cross-modal pre-trained representations (Zellers et al., 2021; Fu et al., 2021) have shown great benefit for VideoQA (see Table 2 and 3). However, most models only demonstrate their good performance on VideoQA tasks that challenge the recognition or shallow description of the video contents. Also, it demands a lot of computation and other resources to handle large-scale video-text data. Therefore, how to pre-train vision-language models more efficiently and how to adapt them to reasoning type of VideoQA tasks deserve more attention.

Interpretability, Robustness and Generalization. Despite the strong power of the advanced pre-training models, it is still unknown how they work, to what extent they can generalize, when they will fail, and how to gain further technical improvement. Recent works towards interpretability and logical robustness (Li et al., 2021b; Sheng et al., 2021) have achieved initial success. (Gandhi et al., 2022) design a benchmark to diagnose whether models can gain true understanding by examining compositional consistency. However, there is a still long way to go towards model interpretability, robustness and generalization. We believe this is of great significance towards practical QA systems.

5 Conclusion

This paper gives a quick overview to the broad aspect of video question answering. We mainly categorized the related datasets and techniques. Also, we discussed some meaningful insights and analyzed the performances of different techniques on different type of datasets. We finally concluded several promising future directions. With these efforts, we hope this survey can shed light and attract more research to VideoQA, and eventually, foster more efforts towards strong AI systems that can demonstrate their understanding of the dynamic visual world by making meaningful responses to our natural language instructions or queries.

Acknowledgements

The research is supported by the Sea-NExT Joint Lab. The research is also supported by the National Natural Science Foundation of China (No.62236003 and No.62276030), and China Scholarships Council (No.202106470037).

Limitations

Although we have tried to comprehensively analyze the literature of VideoQA research, we realize that we fail to cover and detail all the datasets and algorithms due to the thriving VideoQA research and the limited space. Hence, we complement the survey by maintaining a repository https://github.com/VRU-NExT/VideoQA. The repository contains the latest VideoQA papers, datasets, and their open-source implementations. We will periodly update the repository to trace the progress of the latest research.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

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

Shyamal Buch, Cristóbal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. 2022. Revisiting the" video" in video-language understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2917–2927.

Santiago Castro, Mahmoud Azab, Jonathan Stroud, Cristina Noujaim, Ruoyao Wang, Jia Deng, and Rada Mihalcea. 2020. Lifeqa: A real-life dataset for video question answering. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4352–4358.

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

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

Feilong Chen, Duzhen Zhang, Minglun Han, Xiuyi Chen, Jing Shi, Shuang Xu, and Bo Xu. 2022. Vlp: A survey on vision-language pre-training. *arXiv preprint arXiv:2202.09061*.

Jingyuan Chen, Xinpeng Chen, Lin Ma, Zequn Jie, and Tat-Seng Chua. 2018. Temporally grounding natural sentence in video. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 162–171.

Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. 2017. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5659–5667.

Zhenfang Chen, Jiayuan Mao, Jiajun Wu, Kwan-Yee Kenneth Wong, et al. 2021. Grounding physical concepts of objects and events through dynamic visual reasoning. *International Conference on Learning Representations*.

Anoop Cherian, Chiori Hori, Tim K Marks, and Jonathan Le Roux. 2022. (2.5+ 1) d spatio-temporal scene graphs for video question answering. In *AAAI*.

Seongho Choi, Kyoung-Woon On, Yu-Jung Heo, Ahjeong Seo, Youwon Jang, Minsu Lee, and Byoung-Tak Zhang. 2021. Dramaqa: Character-centered video story understanding with hierarchical qa. In *AAAI*, volume 35.

Anthony Colas, Seokhwan Kim, Franck Dernoncourt, Siddhesh Gupte, Daisy Zhe Wang, and Doo Soon Kim. 2020. Tutorialvqa: Question answering dataset for tutorial videos. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5450–5455.

Long Hoang Dang, Thao Minh Le, Vuong Le, and Truyen Tran. 2021. Hierarchical object-oriented spatio-temporal reasoning for video question answering. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 636–642.

Mingyu Ding, Zhenfang Chen, Tao Du, Ping Luo, Josh Tenenbaum, and Chuang Gan. 2021. Dynamic visual reasoning by learning differentiable physics models from video and language. *Advances in Neural Information Processing Systems*, 34.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Deniz Engin, François Schnitzler, Ngoc QK Duong, and Yannis Avrithis. 2021. On the hidden treasure of dialog in video question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2064–2073.

Alex Falcon, Oswald Lanz, and Giuseppe Serra. 2020. Data augmentation techniques for the video question answering task. In *European Conference on Computer Vision*, pages 511–525. Springer.

Chenyou Fan. 2019. Egovqa-an egocentric video question answering benchmark dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0.

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

Zhiyuan Fang, Tejas Gokhale, Pratyay Banerjee, et al. 2020. Video2commonsense: Generating commonsense descriptions to enrich video captioning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 840–860.

Li Fei-Fei and Ranjay Krishna. 2022. Searching for computer vision north stars. *Journal of the American Academy of Arts & Sciences*, page 85.

Christiane Fellbaum. 1998. Wordnet. In *Wiley Online Library*.

Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, et al. 2021. Violet: End-to-end video-language transformers with masked visual-token modeling. *arXiv:2111.12681*.

Mona Gandhi, Mustafa Omer Gul, Eva Prakash, Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. 2022. Measuring compositional consistency for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Difei Gao, Ruiping Wang, Ziyi Bai, and Xilin Chen. 2021. Env-qa: A video question answering benchmark for comprehensive understanding of dynamic environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1675–1685.

Jiyang Gao, Runzhou Ge, Kan Chen, and Ram Nevatia. 2018. Motion-appearance co-memory networks for video question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6576–6585.

Noa Garcia and Yuta Nakashima. 2020. Knowledge-based video question answering with unsupervised scene descriptions. In *European Conference on Computer Vision*, pages 581–598. Springer.

Noa Garcia, Mayu Otani, Chenhui Chu, and Yuta Nakashima. 2020. Knowit vqa: Answering knowledge-based questions about videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10826–10834.

Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. 2021. Agqa: A benchmark for compositional spatio-temporal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11287–11297.

Pranay Gupta and Manish Gupta. 2022. Newskvqa: Knowledge-aware news video question answering. *arXiv preprint arXiv:2202.04015*.

Vivek Gupta, Badri N Patro, Hemant Parihar, and Vinay P Namboodiri. 2022. Vquad: Video question answering diagnostic dataset. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 282–291.

Deng Huang, Peihao Chen, Runhao Zeng, Qing Du, Mingkui Tan, and Chuang Gan. 2020. Location-aware graph convolutional networks for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11021–11028.

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

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

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

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

Weike Jin, Zhou Zhao, Yimeng Li, Jie Li, Jun Xiao, and Yueting Zhuang. 2019. Video question answering via knowledge-based progressive spatial-temporal attention network. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 15(2s):1–22.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Khushboo Khurana and Umesh Deshpande. 2021. Video question-answering techniques, benchmark datasets and evaluation metrics leveraging video captioning: A comprehensive survey. *IEEE Access*, 9:43799–43823.

Junyeong Kim, Minuk Ma, Kyungsu Kim, Sungjin Kim, et al. 2019. Progressive attention memory network for movie story question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8337–8346.

Junyeong Kim, Minuk Ma, Trung Pham, Kyungsu Kim, and Chang D Yoo. 2020. Modality shifting attention network for multi-modal video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10106–10115.

Kyung-Min Kim, Min-Oh Heo, Seong-Ho Choi, and Byoung-Tak Zhang. 2017. Deepstory: video story qa by deep embedded memory networks. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 2016–2022.

Nayoung Kim, Seong Jong Ha, and Je-Won Kang. 2021a. Video question answering using language-guided deep compressed-domain video feature. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1708–1717.

Seonhoon Kim, Seohyeong Jeong, Eunbyul Kim, Inho Kang, and Nojun Kwak. 2021b. Self-supervised pre-training and contrastive representation learning for multiple-choice video qa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13171–13179.

Thomas N Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. *International Conference on Learning Representations*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.

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

Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. 2021. Hierarchical conditional relation networks for multimodal video question answering. *International Journal of Computer Vision*, 129(11):3027–3050.

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

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

Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal. 2020. Tvqa+: Spatio-temporal grounding for video question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8211–8225.

Fangtao Li, Ting Bai, Chenyu Cao, Zihe Liu, Chenghao Yan, and Bin Wu. 2021a. Relation-aware hierarchical attention framework for video question answering. In *Proceedings of the 2021 International Conference on Multimedia Retrieval*, pages 164–172.

Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. 2022a. Learning to answer questions in dynamic audio-visual scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19108–19118.

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

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. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2046–2065.

Linjie Li, Jie Lei, Zhe Gan, and Jingjing Liu. 2021b. Adversarial vqa: A new benchmark for evaluating the robustness of vqa models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2042–2051.

Xiangpeng Li, Jingkuan Song, Lianli Gao, Xianglong Liu, et al. 2019. Beyond rnns: Positional self-attention with co-attention for video question answering. In *AAAI*, pages 8658–8665.

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

Yicong Li, Xiang Wang, Junbin Xiao, Wei Ji, and Tat-Seng Chua. 2022d. Invariant grounding for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2928–2937.

Yongqi Li, Wenjie Li, and Liqiang Nie. 2022e. Mmcoqa: Conversational question answering over text, tables, and images. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4220–4231.

Fei Liu, Jing Liu, Weining Wang, and Hanqing Lu. 2021a. Hair: Hierarchical visual-semantic relational reasoning for video question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1698–1707.

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

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. *Advances in neural information processing systems*, 29.

Tegan Maharaj, Nicolas Ballas, Anna Rohrbach, Aaron Courville, and Christopher Pal. 2017. A dataset and exploration of models for understanding video data through fill-in-the-blank question-answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6884–6893.

Jianguo Mao, Wenbin Jiang, Xiangdong Wang, Zhifan Feng, Yajuan Lyu, Hong Liu, and Yong Zhu. 2022. Dynamic multistep reasoning based on video scene graph for video question answering. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3894–3904.

Jonghwan Mun, Paul Hongsuck Seo, Ilchae Jung, and Bohyung Han. 2017. Marioqa: Answering questions by watching gameplay videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2867–2875.

Seil Na, Sangho Lee, Jisung Kim, and Gunhee Kim. 2017. A read-write memory network for movie story understanding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 677–685.

Liqiang Nie, Wenjie Wang, Richang Hong, Meng Wang, and Qi Tian. 2019. Multimodal dialog system: Generating responses via adaptive decoders. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1098–1106.

Jungin Park, Jiyoung Lee, and Kwanghoon Sohn. 2021. Bridge to answer: Structure-aware graph interaction network for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15526–15535.

Devshree Patel, Ratnam Parikh, and Yesha Shastri. 2021. Recent advances in video question answering: A review of datasets and methods. In *International Conference on Pattern Recognition*, pages 339–356. Springer.

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

Min Peng, Chongyang Wang, Yuan Gao, Yu Shi, and Xiang-Dong Zhou. 2022. Multilevel hierarchical network with multiscale sampling for video question answering. *IJCAI*.

AJ Piergiovanni, Kairo Morton, Weicheng Kuo, Michael S Ryoo, and Anelia Angelova. 2022. Video question answering with iterative video-text co-tokenization. *European Conference on Computer Vision*.

Zi Qian, Xin Wang, Xuguang Duan, Hongyang Chen, and Wenwu Zhu. 2022. Dynamic spatio-temporal modular network for video question answering. In *Proceedings of the 30th ACM International Conference on Multimedia*, page 4466–4477.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763.

Arka Sadhu, Kan Chen, and Ram Nevatia. 2021. Video question answering with phrases via semantic roles. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2460–2478.

Ahjeong Seo, Gi-Cheon Kang, Joonhan Park, and Byoung-Tak Zhang. 2021a. Attend what you need: Motion-appearance synergistic networks for video question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 6167–6177.

Paul Hongsuck Seo, Arsha Nagrani, and Cordelia Schmid. 2021b. Look before you speak: Visually contextualized utterances. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16877–16887.

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

Xindi Shang, Yicong Li, Junbin Xiao, Wei Ji, and Tat-Seng Chua. 2021. Video visual relation detection via iterative inference. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3654–3663.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.

Sasha Sheng, Amanpreet Singh, Vedanuj Goswami, Jose Magana, Tristan Thrush, Wojciech Galuba, Devi Parikh, and Douwe Kiela. 2021. Human-adversarial visual question answering. *Advances in Neural Information Processing Systems*, 34:20346–20359.

Xiaomeng Song, Yucheng Shi, Xin Chen, and Yahong Han. 2018. Explore multi-step reasoning in video question answering. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 239–247.

Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. 2015. End-to-end memory networks. *Advances in neural information processing systems*, 28.

Guanglu Sun, Lili Liang, Tianlin Li, Bo Yu, et al. 2021. Video question answering: a survey of models and datasets. *Mobile Networks and Applications*, 26(5):1904–1937.

Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4631–4640.

Aisha Urooj, Amir Mazaheri, Mubarak Shah, et al. 2020. Mmft-bert: Multimodal fusion transformer with bert encodings for visual question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4648–4660.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Jianyu Wang, Bingkun Bao, and Changsheng Xu. 2021. Dualvgr: A dual-visual graph reasoning unit for video question answering. *IEEE Transactions on Multimedia*.

Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. 2021a. Star: A benchmark for situated reasoning in real-world videos. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Tianran Wu, Noa Garcia, Mayu Otani, Chenhui Chu, Yuta Nakashima, and Haruo Takemura. 2021b. Transferring domain-agnostic knowledge in video question answering. *arXiv:2110.13395*.

Zhibiao Wu and Martha Stone Palmer. 1994. Verb semantics and lexical selection. In *32nd Annual Meeting of the Association for Computational Linguistics, 27-30 June 1994, New Mexico State University, Las Cruces, New Mexico, USA, Proceedings*, pages 133–138.

Junbin Xiao, Xindi Shang, Xun Yang, Sheng Tang, and Tat-Seng Chua. 2020. Visual relation grounding in videos. In *European conference on computer vision*, pages 447–464. Springer.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9777–9786.

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

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

Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. 2017. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653.

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

Li Xu, He Huang, and Jun Liu. 2021. Sutd-trafficqa: A question answering benchmark and an efficient network for video reasoning over traffic events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9878–9888.

Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. 2021a. Just ask: Learning to answer questions from millions of narrated videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1686–1697.

Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. 2022a. Learning to answer visual questions from web videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Pinci Yang, Xin Wang, Xuguang Duan, Hong Chen, Runze Hou, Cong Jin, and Wenwu Zhu. 2022b. Avqa: A dataset for audio-visual question answering on videos. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3480–3491.

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

Zekun Yang, Noa Garcia, Chenhui Chu, Mayu Otani, Yuta Nakashima, and Haruo Takemura. 2020. Bert representations for video question answering. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1556–1565.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, et al. 2020. Clevrer: Collision events for video representation and reasoning. In *International Conference on Learning Representations*.

Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. 2018. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. *Advances in neural information processing systems*, 31.

Weijiang Yu, Haoteng Zheng, Mengfei Li, Lei Ji, Lijun Wu, Nong Xiao, and Nan Duan. 2021. Learning from inside: Self-driven siamese sampling and reasoning for video question answering. *Advances in Neural Information Processing Systems*, 34.

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

Heeseung Yun, Youngjae Yu, Wonsuk Yang, Kangil Lee, and Gunhee Kim. 2021. Pano-avqa: Grounded audio-visual question answering on 360deg videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2031–2041.

Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. 2019. Social-iq: A question answering benchmark for artificial social intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8807–8817.

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

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

Ao Zhang, Yuan Yao, Qianyu Chen, Wei Ji, Zhiyuan Liu, Maosong Sun, and Tat-Seng Chua. 2022. Fine-grained scene graph generation with data transfer. In *European conference on computer vision*.

Wentian Zhao, Seokhwan Kim, Ning Xu, and Hailin Jin. 2020. Video question answering on screencast tutorials. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pages 1061–1068.

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

Zhou Zhao, Qifan Yang, Deng Cai, Xiaofei He, and Yueting Zhuang. 2017b. Video question answering via hierarchical spatio-temporal attention networks. In *IJCAI*, volume 2, page 8.

Zhou Zhao, Zhu Zhang, Shuwen Xiao, Zhou Yu, Jun Yu, Deng Cai, Fei Wu, and Yueting Zhuang. 2018. Open-ended long-form video question answering via adaptive hierarchical reinforced networks. In *IJCAI*, volume 2, page 8.

Linchao Zhu, Zhongwen Xu, Yi Yang, and Alexander G Hauptmann. 2017. Uncovering the temporal context for video question answering. *International Journal of Computer Vision*, 124(3):409–421.

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

Yueting Zhuang, Dejing Xu, Xin Yan, Wenzhuo Cheng, Zhou Zhao, Shiliang Pu, and Jun Xiao. 2020. Multichannel attention refinement for video question answering. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 16(1s):1–23.

A Appendix: Details of VideoQA datasets in the literature

Due to limited space, details of VideoQA datasets are listed in Table A1.

B Appendix: Timeline of VideoQA techniques

In the literature, the VideoQA datasets and techniques jointly evolve in time (as shown in Figure A1). Some of the datasets and techniques influence each other. As the cross-modal pre-training and fine-tuning technique develops, the performance of early-stage datasets like TGIF-QA (Jang et al., 2017) and TVQA+ (Lei et al., 2020) reaches unprecedentedly high (close to human performance, refer to Table 3 and Table 4). The new research focus turns to the more challenging VideoQA datasets like NExT-QA (Xiao et al., 2021), which invokes complicated inference among multiple objects and relations. In turn, the inference QA datasets motivate new research interests in new techniques. CLEVRER (Yi et al., 2020) has inspired new works using neuro-symbolic learning (Yi et al., 2020; Chen et al., 2021; Ding et al., 2021), and NExT-QA has promoted a lot of recent works on graph models (Xiao et al., 2022a,b). Diagnostic datasets like AGQA (Grunde-McLaughlin et al., 2021) and AGQA 2.0 (Gandhi et al., 2022) analyze existing methods by checking compositional consistency to examine whether they gain true understanding. These diagnostic datasets are promising to find the existing defects and motivate new methods (Qian et al., 2022).

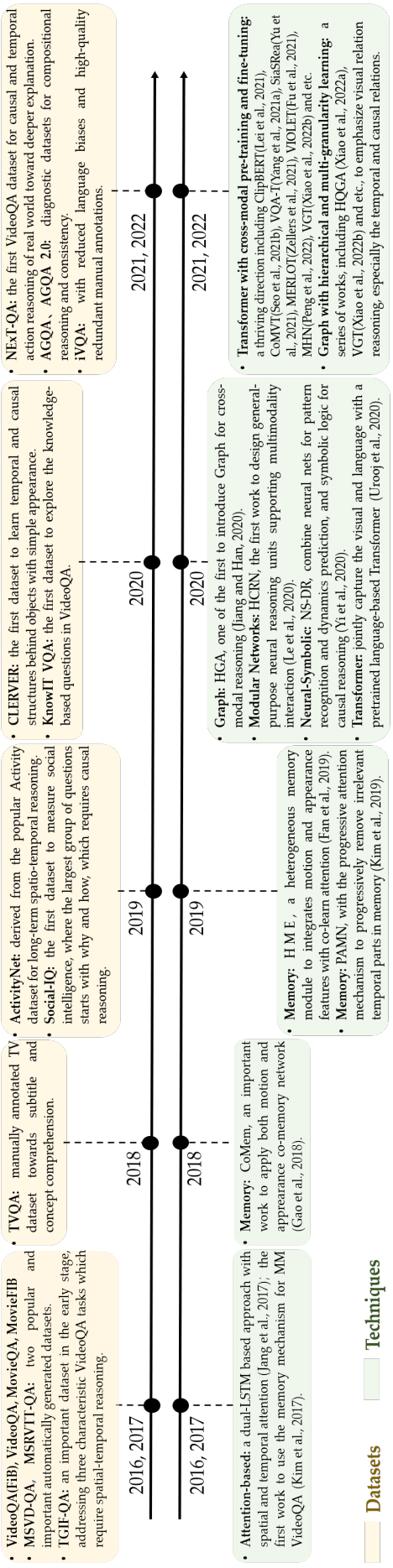

Figure A1: Timeline of VideoQA datesets and techniques. In the literature of VideoQA, datasets and techniques jointly evolve over time, and some of them influence each other.

Table A1: VideoQA datasets in the literature. (MTax: Modality-based Taxonomy, QTax: Question-based Taxonomy, Vid: VideoQA, MM: Multi-modal VideoQA, KB: Knowledge-based VideoQA, F: Factoid VideoQA, I: Inference VideoQA, Auto: automatic generation, Man: manual annotation, MC: multi-choice QA, OE: open-ended QA.)

Dataset	MTax	QTax	Data Source	Goal	#Video/#QA	Annotation	Task
VideoQA(FiB) (Zhu et al., 2017)	Vid	F	Multiple source	Temporal reasoning	109K/390K	Auto	MC
VideoQA (Zeng et al., 2017)	Vid	F	Web videos	Description	18K/174K	Auto, Man	OE
MSVD-QA (Xu et al., 2017)	Vid	F	Web videos	Description	1.9K/50K	Auto	OE
MSRVTT-QA (Xu et al., 2017)	Vid	F	Web videos	Description	10K/243K	Auto	OE
YouTube2Text-QA (Zhao et al., 2017a)	Vid	F	Web videos	Description	1.9K/48K	Auto	MC, OE
MarioQA (Mun et al., 2017)	Vid	F	Game	Temporal reasoning	92K/92K	Auto	OE
ActivityNet-QA (Yu et al., 2019)	Vid	F	Web videos	Description	5.8K/58K	Man	OE
EgoVQA (Fan, 2019)	Vid	F	Egocentric videos	First-person VideoQA	520/580	Man	MC
HowToVQA69M (Yang et al., 2021a)	Vid	F	Web videos	Pre-training for downstream tasks	69M/69M	Auto	OE
iVQA (Yang et al., 2021a)	Vid	F	Web videos	Removing language bias	10K/10K	Man	OE
ASRL-QA (Sadhu et al., 2021)	Vid	F	Internet videos	VideoQA with phrases	35K/162K	Auto	OE
Charades-SRL-QA (Sadhu et al., 2021)	Vid	F	Crowd-Sourced	VideoQA with phrases	9.5K/71K	Auto	OE
WebVidVQA3M (Yang et al., 2022a)	Vid	F	Web videos	Pre-training for downstream tasks	2M/3M	Auto	OE
FIBER (Castro et al., 2022b)	Vid	F	Web videos	Fill-in-the-blanks task with diverse answers	28K/28K	Man	OE
WildQA (Castro et al., 2022a)	Vid	F	In-the-wild videos	In-the-wild videos with evidence selection	369/916	Man	OE
MovieQA (Tapaswi et al., 2016)	MM	F	Movies	Text & Visual story comprehension	6.7K/6.4K	Man	MC
MovieFIB (Maharaj et al., 2017)	MM	F	Movies	Description	118K/348K	Auto	OE
PororoQA (Kim et al., 2017)	MM	F	Cartoon	Story comprehension	171/8.9K	Man	MC
TVQA (Lei et al., 2018)	MM	F	TV shows	Subtitle & Concept comprehension	21K/152K	Man	MC
TVQA+ (Lei et al., 2020)	MM	F	TV shows	Spatio-temporal VideoQA	4.1K/29K	Man	MC
LifeQA (Castro et al., 2020)	MM	F	Web videos	Real-life understanding	275/2.3K	Man	MC
How2QA (Li et al., 2020)	MM	F	Web videos	Multimodal challenges	22K/44K	Man	MC
Env-QA (Gao et al., 2021)	MM	F	Egocentric videos	Exploring & interacting with environments	23K/85K	Auto, Man	OE
Pano-AVQA (Yun et al., 2021)	MM	F	360° videos	Spherical spatial & audio-visual relation	5.4K/51.7K	Man	OE
DramaQA (Choi et al., 2021)	MM	F	TV shows	Story comprehension	23K/17K	Man	MC
MUSIC-AVQA (Li et al., 2022a)	MM	F	Musical videos	Audio-Visual VideoQA	9.3K/45K	Man	OE
TGIF-QA (Jang et al., 2017)	Vid	I	Animated GIF	Spatio-temporal reasoning	71K/165K	Auto, Man	MC, OE
SVQA (Song et al., 2018)	Vid	I	Synthetic videos	Logical compositional questions	12K/118K	Auto	OE
Social-IQ (Zadeh et al., 2019)	MM	I	Web videos	Measuring social intelligence	1.2K/7.5K	Man	MC
PsTuts-VQA (Zhao et al., 2020)	KB	I	Tutorial videos	Narrated instructional videos	76/17K	Man	MC
KnowIT VQA (Garcia et al., 2020)	KB	I	TV shows	Knowledge in VideoQA	12K/24K	Man	MC
KnowIT-X VQA (Wu et al., 2021b)	KB	I	TV shows	Transfer learning	12K/21K	Man	MC
NEWSKVQA (Gupta and Gupta, 2022)	KB	I	News videos	Knowledge-based QA of news videos	12K/1M	Auto	MC
V2C-QA (Fang et al., 2020)	Vid	I	Web videos	Commonsense reasoning	1.5K/37K	Auto	OE
TutorialVQA (Colas et al., 2020)	Vid	I	Tutorial videos	Multi-step & non-factoid VideoQA	408/6.1K	Man	OE
CLEVRER (Yi et al., 2020)	Vid	I	Synthetic videos	Temporal & causal structures	10K/305K	Auto	MC, OE
TGIF-QA-R (Peng et al., 2021)	Vid	I	Animated GIF	Overcoming answer biases	71K/165K	Auto	MC
SUTD-TrafficQA (Xu et al., 2021)	Vid	I	Traffic scenes	Understanding & inference in traffic	10K/62K	Man	MC
AGQA(Grunde-McLaughlin et al., 2021)	Vid	I	Homemade videos	Compositional reasoning	9.6K/192M	Auto	OE
AGQA 2.0(Gandhi et al., 2022)	Vid	I	Homemade videos	Compositional consistency	9.6K/4.55M	Auto	OE
NExT-QA (Xiao et al., 2021)	Vid	I	Web videos	Causal & temporal action interactions	5.4K/52K	Man	MC, OE
STAR (Wu et al., 2021a)	Vid	I	Homemade videos	Situated reasoning in real-world videos	22K/60K	Auto	MC
Causal-VidQA (Li et al., 2022b)	Vid	I	Web videos	Evidence & commonsense reasoning	26K/107K	Man	MC
VQuAD (Gupta et al., 2022)	Vid	I	Synthetic videos	Spatio & temporal reasoning	7K/1.3M	Auto	OE