Structuring Video Semantics with Temporal Triplets for Zero-Shot Video Question Answering

Linlin Zong
Dalian University of Technology
Dalian, China
llzong@dlut.edu.cn

Xinyu Zhai
Dalian University of Technology
Dalian, China
zhaixy@mail.dlut.edu.cn

Xinyue Liu
Dalian University of Technology
Dalian, China
xyliu@dlut.edu.cn

Wenxin Liang
Dalian University of Technology
Dalian, China
wxliang@dlut.edu.cn

Xianchao Zhang
Dalian University of Technology
Dalian, China
xczhang@dlut.edu.cn

Bo Xu*
Dalian University of Technology
Dalian, China
xubo@dlut.edu.cn

ABSTRACT

Current large vision-language models (VLMs) exhibit remarkable performance in basic video understanding tasks. However, existing VLMs are still limited to surface-level perception and lack fine-grained spatio-temporal understanding and combinatorial reasoning capabilities. Existing methods typically rely on expensive human annotations or subtitle extraction, yet they struggle to effectively model temporal relations between frames. This paper proposes a structured representation based on temporal triplets to address two major challenges in traditional approaches: temporal fragmentation and entity reference ambiguity. By modeling objects, attributes, and relationships within the video and incorporating temporal information, we convert semantic content from keyframes into a sequence of temporal triplets. This structured representation is then used as input for zero-shot video question answering (VideoQA). Experiments were conducted on four benchmark VideoQA datasets: NExT-QA, STAR, MSVD-QA, and MSRVTT-QA, showing that our method achieves competitive performance without requiring fine-tuning, validating its generality and effectiveness.

CCS CONCEPTS

• **Computing methodologies → Temporal reasoning**; **Information extraction**.

KEYWORDS

Zero-Shot Video Question Answering, Temporal Triplets, Large Language Model, Structured Video Representation

ACM Reference Format:
Linlin Zong, Xinyu Zhai, Xinyue Liu, Wenxin Liang, Xianchao Zhang, and Bo Xu. 2025. Structuring Video Semantics with Temporal Triplets for Zero-Shot Video Question Answering. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM '25), November 10–14, 2025, Seoul, Republic of Korea.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3746252.3760857

1 INTRODUCTION

Recently, vision-language models (VLMs) have demonstrated remarkable performance in Visual Question Answering (VQA) tasks across both image and video domains [6, 10, 14]. Video Question Answering (VideoQA), as a multimodal reasoning task integrating visual, textual, and temporal information, is inherently more challenging than image-based QA [12]. Videos exhibit temporal dynamics, with events involving multi-level semantics ranging from fine-grained object actions (e.g., "waving") to coarse-grained scene transitions (e.g., "entering a room"), and complex temporal continuity and causal relationships among actions [2]. Existing methods typically enhance video semantics by generating video captions, frame-level descriptions, or question-answer pairs [3, 13], but they suffer from two major issues: (1) Temporal fragmentation: Captions are linear textual representations, and independent frame-level descriptions struggle to explicitly model the temporal span of actions. (2) Entity reference ambiguity: The lack of object association across frames leads to inconsistent identification, where the same entity may be recognized as different objects at different time points.

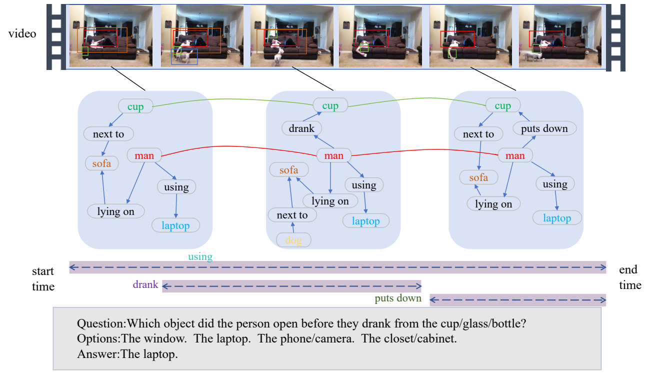
A visual scene is not only composed of objects, but also enriched with diverse attributes and relationships [11]. We believe that a fine-grained, structured representation of video is essential for comprehensive video understanding. Therefore, we propose a structured representation paradigm called Temporal Triplets. As shown in Figure 1, this approach decomposes the video into a sequence of temporal triplets, addressing the ambiguity of event temporal boundaries. In addition, by constructing global object representations, it enables consistent cross-frame entity alignment.

Specifically, our method consists of the following key steps: First, we leverage LLM to extract objects and attributes from the video, thereby constructing global object representations. Based on predefined templates, we leverage LLM to extract objects, attributes, and their relationships from key frames, generating temporal triplets that precisely localize key events and align entities across frames. Finally, the original video content is transformed into a structured sequence of temporal triplets, which is then converted into textual form and fed into the LLM to facilitate zero-shot VideoQA and

*Corresponding author.

CIKM '25, November 10–14, 2025, Seoul, Republic of Korea

ACM ISBN 979-8-4007-2040-6/2025/11
https://doi.org/10.1145/3746252.3760857



Question:Which object did the person open before they drank from the cup/glass/bottle?
Options:The window. The laptop. The phone/camera. The closet/cabinet.
Answer:The laptop.

Figure 1: Using temporal triplets to represent videos enables structured modeling of objects, attributes, and relationships (with the same color indicating semantically consistent entities), capturing their temporal semantic evolution.

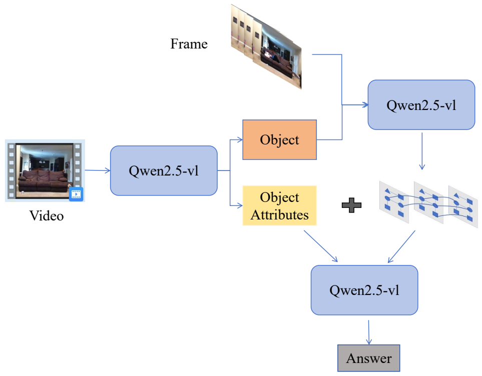


Figure 2: Our pipeline for Zero-shot VideoQA.

reasoning. In summary, our main contributions are as follows: (1) We propose a zero-shot VideoQA framework based on temporal triplets, which organizes visual information with explicit temporal modeling to enhance reasoning. (2) Experimental results on four VideoQA benchmark datasets demonstrate that our method achieves competitive performance without requiring additional annotations or fine-tuning.

2 METHOD

In this section, we present a model-agnostic, training-free, zero-shot VideoQA approach based on temporal triplets. As shown in Figure 2. Given an input video, we first extract global object representations. Then, we perform uniform sampling to select keyframes and generate detailed frame-level descriptions. Next, we construct temporal triplets to abstract the visual content into structured semantic representations. Finally, the temporal triplets and their attribute information are used to perform reasoning for VideoQA.

2.1 Video Processing

A video is composed of a series of consecutive frames, but there is often redundancy in the temporal dimension, as adjacent frames tend to share highly similar semantic content. Keyframe sampling is a common video sampling method, but it can easily miss important frames that represent scene or event changes. Therefore, we propose a semantic-aware keyframe clustering strategy that minimizes redundant features while preserving fine-grained semantics.

We employ a vision-language model (Qwen2.5-VL [1]) to generate image-text descriptions for each frame and use the BGE-M3 model to compute semantic similarity between adjacent frames. If two consecutive frames exhibit high similarity, it indicates they likely represent the same underlying event and can be grouped into a coherent temporal unit. Specifically, we initialize the keyframe set V' with the first frame. For each frame in V, we calculate its similarity d with the latest frame in the current keyframe set V'. If d is less than a predefined threshold, it indicates the current frame is semantically different from the most recent keyframe and is added to V'; otherwise, the frame is discarded.

2.2 Semantic Modeling

When constructing the temporal triplet representation of a video, the input consists of three components: the original video, sampled frames, and textual prompts for generation guidance. During the question-answering phase, reasoning is performed solely based on the triplet sequence, without using raw video or frame data. Therefore, we need to generate high-quality triplets that effectively capture multi-level and fine-grained spatiotemporal semantics.

For each video, we design prompts to extract triplets from keyframes. The temporal triplet structure consists of frame-level sets, each containing subject-predicate-object triplets. To model object interactions, we divide the construction into the following steps:

Object and Attribute Modeling. The generation of triplet encompasses objects, their attributes, and the relationships between them. We aim not only to have a global view of the objects but also to achieve global semantic modeling of their attributes and interactions. Therefore, we propose a top-down triplet generation strategy, where we first extract a unified set of object nodes at the video level as semantic anchors, enabling consistent entity alignment across frames. Specifically, we extract a unified set of objects at the video level and generate fine-grained attribute representations for each object. We require the model to refer to each object using a distinctive and unique attribute, thereby ensuring entity consistency in subsequent subtitle generation and semantic alignment.

This global object set approach avoids issues such as attribute drift and resolves the challenge of diverse expressions for the same entity in question-answering scenarios. For example, questions such as “What is the man in white doing?” and “What is the man wearing a hat doing?” may refer to the same person, though described differently. Traditional frame-level independent object detection methods struggle to bridge such identity mappings. Our method allows the use of unified descriptions such as “the man in white” instead of vague expressions like “a man,” enabling consistent references to the same entity across different frames.

Relation Modeling and Triplet Extraction. For each pair of object nodes (o_i, o_j), we model their interaction by constructing subject-predicate-object triplets, expressed as a relation edge:

$$r_{i,j} = (o_i, p_{i,j}, o_j) \tag{1}$$

where $p_{i,j}$ represents the predicate relation between o_i and o_j.

Concretely, for a uniformly sampled set of T frames $\{f_1, f_2, \ldots, f_T\}$, we employ the Qwen2.5-VL-7B-Instruct [1] model to generate fine-grained semantic descriptions d_i for each frame, guided by carefully designed prompts. This process is formalized as:

$$d_i = I_b(f_i, P), \quad i = 1, 2, \ldots, n_d \tag{2}$$

where P is an instruction-style prompt guiding the model to produce detailed natural language descriptions for frame f_i. We then design a structured extractor that uses LLM to extract structured semantic units in the form of triplets $(\text{subject}, \text{predicate}, \text{object})$ from each description d_i, denoted as:

$$T_i = \text{ExtractTriplets}(d_i), \quad T_i = \{(s_k, p_k, o_k)\}_{k=1}^{m_i} \tag{3}$$

where T_i is the set of triplets extracted from the i-th frame and m_i is the number of triplets in that frame.

Finally, we represent the original video V as a sequence of structured frame-level representations:

$$\mathcal{T}_V = \{T_1, T_2, \ldots, T_n\} \tag{4}$$

2.3 Cross-frame Alignment and Temporal Fusion Modeling

Since we performed redundancy removal on the video frames, when extracting frame-level structured triples, we need to construct their temporal spans in the video, i.e., the time range during which each triple's semantics are maintained. To this end, we use frame indices to simulate the temporal intervals of the video. Each structured triple is assigned a frame interval from the current frame to the next key frame in chronological order, represented as:

$$m_k = (o_i, p_k, o_j; [t_1, t_2]) \tag{5}$$

where $[t_1, t_2]$ indicates the time period during which the action occurs. We then convert the set of frame indices for each triple into a continuous time interval representation, thereby obtaining a structured sequence of temporal triples.

The structured triples generated by LLM may have inconsistent expressions, and semantic redundancy can introduce noise into reasoning. On the other hand, in videos, a behavior may span multiple time segments and is not always continuously present. For example, a person may "raise and wave their hand multiple times" throughout the video. If we model this based on independently extracted triples from each frame, such behaviors would be fragmented into multiple segments, making it difficult to capture their complete semantic scope. Therefore, we further design a fusion mechanism based on semantic embeddings. Specifically, we use the BGE-M3 model to encode all triples and construct a cosine similarity matrix to aggregate semantically similar triples. For any two triples, if their similarity exceeds a threshold, they are considered to form a semantic group. We then merge the similar triples and combine their corresponding time intervals to produce a unified temporal sequence. Through this method, we can both distinguish object behaviors and capture their temporal characteristics.

2.4 Reasoning Module

We adopt the Qwen2.5 [1] model as the reasoning module in this work. To enable the model to reason more effectively for video-based question answering, we input the temporal triplets S in JSON format, sorted by time. We then combine the global object information O, temporal triplets S, prompt P, and question Q as input to the large model, forming a QA instruction M as follows:

$$M = \text{Concat}(S, O, P, Q, A) \tag{6}$$

which guides the model to select the correct answer from the options. For generative QA, A is empty, and the model is required to generate the predicted answer based on the input.

3 EXPERIMENTS

3.1 Experimental Setup

Dataset. To test our approach, we conducted experiments on several mainstream VideoQA benchmark datasets, including multiple-choice (NExT-QA [18], STAR [17]) and open-ended question formats (MSVD-QA [19], and MSRVTT-QA [20]). We follow the original split settings of the datasets without any modifications.

Baselines. We compare our method with existing baselines, covering LLM-based methods (e.g., VideoChat [6], Video-LLaVA [23], Q-ViD [13]) and non-LLM methods (e.g., FrozenBiLM [21], InternVideo [16]), to comprehensively evaluate the performance of different modeling paradigms on VideoQA tasks.

Settings. For each video, we perform uniform frame sampling. We sample T frames from each video. Considering the differences in scene change rates and durations across videos, we choose different sampling intervals based on the dataset type: for the NExT-QA [18] and STAR [17] datasets, we set $T = 20$; for the MSVD-QA [19] and MSRVTT-QA [20] datasets, we set $T = 12$. This setup effectively covers the event evolution within the video. In our experiments, we adopt Qwen2.5-VL-7B-Instruct [1] as our backbone model to perform object extraction, temporal triplet generation, and video question answering (VideoQA) reasoning tasks.

3.2 Overall Performance

Evaluation Details. We evaluate our method using the following benchmarks: (1) We adopt Accuracy as the primary evaluation metric, defined as the ratio of correctly answered questions to the total number of questions, to measure the overall correctness of the model's answers. (2) For open-ended question answering tasks, we follow the evaluation protocol of VideoChatGPT [6] and use GPT-3.5-Turbo as an automatic evaluator, reporting the following two metrics: Accuracy and average score (where ChatGPT rates each response on a scale of 0-5, with the mean score calculated).

As shown in Table 1, we evaluated our proposed method on the MVBench [7] benchmark (excluding the FP sub-task) and compared it with several existing approaches. Notably, compared to previous methods, our approach also achieved excellent performance on the comprehensive evaluation benchmarks.

The results are shown in Table 2 and Table 3. In the table, we present the performance of our method on four benchmark datasets: NExT-QA [18], STAR [17], MSVD-QA [19], and MSRVTT-QA [20], to evaluate the effectiveness of our proposed framework. Our method

Table 1: Comparison of MVBench benchmark. We bold the best results, and underline the second-best results. Ours shows to be competitive and even outperform some more complex frameworks for zero-shot video QA.

Model	AS	AP	AA	FA	UA	OE	OI	OS	MD	AL	ST	AC	MC	MA	SC	CO	EN	ER	CI	Avg.
Video-ChatGPT[10]	23.5	26.0	62.0	22.5	26.5	54.0	28.0	40.0	23.0	20.0	31.0	30.5	25.5	39.5	48.5	33.0	29.5	26.0	35.5	32.8
Video-LLaMA[23]	27.5	25.5	51.0	29.0	39.0	48.0	40.5	38.0	22.5	22.5	43.0	34.0	22.5	32.5	45.5	40.0	30.0	21.0	37.0	34.1
VideoChat[6]	33.5	26.5	56.0	33.5	40.5	53.0	40.5	30.0	25.5	27.0	48.5	35.0	20.5	42.5	46.0	41.0	23.5	23.5	36.0	35.9
Video-LLaVA[8]	46.0	42.5	56.5	39.0	53.5	53.0	48.0	**41.0**	29.0	31.5	82.5	**45.0**	26.0	53.0	41.5	41.5	27.5	38.5	31.5	43.5
ShareGPT4Video[3]	49.5	39.5	**79.5**	40.0	54.5	**82.5**	54.5	32.5	**50.5**	**41.5**	**84.5**	35.5	**62.5**	**75.0**	**51.0**	46.5	28.5	**39.0**	**51.5**	**52.5**
Otter[5]	23.0	23.0	27.5	27.0	29.5	53.0	28.0	33.0	24.5	23.5	27.5	26.0	28.5	18.0	38.5	22.0	23.5	19.0	19.5	27.1
Ours	**55.5**	**49.5**	68.5	**41.0**	**58.5**	66.3	**57.4**	36.0	43.0	37.0	74.5	40.0	58.0	66.0	42.5	**69.0**	**35.5**	32.5	40.0	51.1

consistently outperforms existing approaches. Notably, on the STAR [17] dataset, it achieves a 5.3% improvement over previous methods. This significant gain may be attributed to the precise modeling of event temporal dynamics, which enables our framework to effectively capture key events and temporal information in videos.

Table 2: Zero-shot results on multiple-choice VideoQA.

Model	**NExT-QA**				**STAR**
	Tem.	**Cau.**	**Des.**	**Avg.**	
InternVideo[16]	-	-	-	49.1	41.6
VideoChat[6]	-	-	-	52.8	45.0
Video-ChatGPT[10]	-	-	-	53.0	48.7
SeViLa[22]	61.3	61.5	75.6	63.6	44.6
OneLLM-7b[4]	61.3	61.5	75.6	63.6	44.6
Video-LLaVA[8]	-	-	-	57.3	50.6
UIO-2[9]	-	-	-	-	52.2
Q-ViD[13]	**61.6**	67.6	72.2	66.3	45.7
Ours	61.4	**68.8**	**76.2**	**67.5**	**57.5**

Table 3: Zero-shot results on open-ended VideoQA.

Model	**MSVD-QA**		**MSRVTT-QA**	
	Acc.	**Score**	**Acc.**	**Score**
FrozenBiLM[21]	54.8	-	47.0	-
InternVideo[16]	55.5	-	47.1	-
LLaMA Adapter[24]	54.9	3.1	43.8	2.7
VideoChat[6]	56.3	2.8	45.0	2.5
Video-LLaMA[23]	51.6	-	29.6	-
Video-ChatGPT[10]	64.9	3.3	49.3	**2.8**
Emu2-Chat[15]	49.0	-	31.4	-
OneLLM-7b[4]	56.5	-	43.8	-
ShareGPT4Video[3]	45.6	-	43.0	-
Otter-7B[5]	55.0	-	47.0	-
Ours	**65.0**	**3.4**	**51.6**	**2.8**

3.3 Ablation Study

To further analyze the contribution of each component in our video reasoning framework, we perform ablation studies on the MSVD [19] dataset, focusing on different visual-semantic inputs: (1) **VideoDesc**: A single holistic description for the video, (2) **FrameDesc**: Per-frame descriptions with temporal indices, and (3) **VideoTriplets**: Directly extract triplets from the video. The experimental results are shown in table 4. We observe that our method is capable of extracting more fine-grained information and provides better interpretability. Although FrameDesc yields slightly higher scores, it relies on verbose frame-level descriptions, whereas our method provides more structured, compact, and interpretable representations suitable for generalizable reasoning.

Table 4: Ablation study of individual components.

	Accuracy	**Score**
Ours	65.0	3.32
VideoDesc	62.0	3.36
FrameDesc	65.0	3.48
VideoTriplets	60.7	3.32

In addition, we conducted experiments on key parameters in the Semantic Modeling stage. We set the similarity threshold to 0.9. As shown in Figure 3, this strategy successfully removes nearly half of the frames while maintaining model accuracy, significantly reduces the number of input frames, merges events across different time segments, and enhances the interpretability of the model.

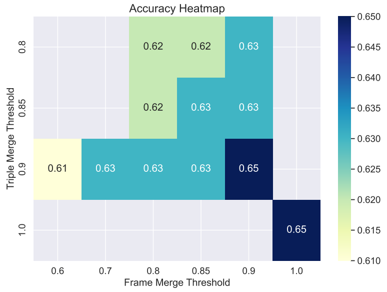


Figure 3: Impact of Frame Description Threshold and Triplet Merging Threshold Combinations on Accuracy.

4 CONCLUSION

In this work, we propose a zero-shot VideoQA framework based on temporal triplets, which converts videos into structured, fine-grained semantic sequences without training. A top-down object construction approach aligns entities across frames, and carefully designed prompts extract rich semantics for temporally fused triplets. Our method achieves strong results across multiple evaluation tasks, demonstrating broad applicability in video question answering and reasoning.

ACKNOWLEDGMENTS

This work was supported in part by the Liaoning Provincial Natural Science Foundation Joint Fund Program (2023-MSBA-003) and the Fundamental Research Funds for the Central Universities (DUT24MS003).

GENAI USAGE DISCLOSURE

This study employs the multimodal large language model Qwen2.5-VL-7B-Instruct for key tasks such as video scene understanding, structured semantic representation, and semantic reasoning. Additionally, the GPT-3.5-Turbo model was used to evaluate the correctness of the model's question-answering results. All task design, prompt engineering, experimental implementation, and result analysis were independently conducted by the authors. This paper used ChatGPT to correct grammar errors and polish it, but no large chunks of content were generated.

REFERENCES

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923* (2025).

[2] Ziyi Bai, Ruiping Wang, Difei Gao, and Xilin Chen. 2024. Event Graph Guided Compositional Spatial–Temporal Reasoning for Video Question Answering. *IEEE Transactions on Image Processing* 33 (2024), 1109–1121.

[3] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Zhenyu Tang, Li Yuan, et al. 2024. Sharegpt4video: Improving video understanding and generation with better captions. *Advances in Neural Information Processing Systems* 37 (2024), 19472–19495.

[4] Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. 2024. Onellm: One framework to align all modalities with language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 26584–26595.

[5] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Joshua Adrian Cahyono, Jingkang Yang, Chunyuan Li, and Ziwei Liu. 2025. Otter: A multi-modal model with in-context instruction tuning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2025).

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

[7] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. 2024. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22195–22206.

[8] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2023. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122* (2023).

[9] Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. 2024. Unified-io 2: Scaling autoregressive multimodal models with vision language audio and action. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 26439–26455.

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

[11] Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. 2024. Compositional chain-of-thought prompting for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14420–14431.

[12] Tianwen Qian, Jingjing Chen, Shaoxiang Chen, Bo Wu, and Yu-Gang Jiang. 2022. Scene graph refinement network for visual question answering. *IEEE Transactions on Multimedia* 25 (2022), 3950–3961.

[13] David Romero and Thamar Solorio. 2024. Question-instructed visual descriptions for zero-shot video question answering. *arXiv preprint arXiv:2402.10698* (2024).

[14] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*. Springer, 146–162.

[15] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. 2024. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14398–14409.

[16] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. 2022. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191* (2022).

[17] Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. 2024. Star: A benchmark for situated reasoning in real-world videos. *arXiv preprint arXiv:2405.09711* (2024).

[18] 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*. 9777–9786.

[19] 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*. 1645–1653.

[20] 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*. 5288–5296.

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

[22] Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. 2023. Self-chained image-language model for video localization and question answering. *Advances in Neural Information Processing Systems* 36 (2023), 76749–76771.

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

[24] Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. 2023. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199* (2023).