Unleashing the Power of LLMs for Medical Video Answer Localization

Junbin Xiao *[1][0000−0001−5573−6195], Qingyun Li *[1], Yusen Yang[2], Liang Qiu[3][0000−0002−6784−1627], and Angela Yao†[1][0000−0001−7418−6141]

[1] National University of Singapore, Sinagpore
[2] Waseda University, Japan [3] Standford University, USA
junbin@comp.nus.edu.sg, liqingyun@u.nus.edu,
yanagiumori@suou.waseda.jp, qiuliang@stanford.edu, ayao@comp.nus.edu.sg

Abstract. Given an untrimmed medical instructional video and a textual question, medical video answer localization is to locate the precise temporal span that visually answers the question. Existing methods primarily rely on supervised learning to tackle this problem. This requires massive annotated data for training and shows limited flexibility in generalizing across different datasets, especially in the medical domain. With the remarkable advancements of large language models (LLMs) and their multimodal variants (MLLMs), we explore a Socratic approach to compose LLMs and MLLMs to achieve zero-shot video answer localization. Our method effectively takes advantage of the rich subtitles and visual descriptions in instructional videos to prompt LLMs. We also develop a subtitle refinement and early fusion strategy for better performance. Experiments on MedVidQA and COIN-Med show that our method outperforms existing state-of-the-art (SOTA) zero-shot multimodal models significantly by 41.0% and 20.3% in mIoU, respectively. It even surpasses SOTA supervised methods, suggesting the strength of our approach.

Keywords: Medical VideoQA · LLMs · Temporal Localization.

1 Introduction

Medical video analysis is pivotal in modern healthcare, supporting clinical decision-making, medical education, and patient care. Medical video answer localization (MedVidQA) [4] is a key task that identifies precise temporal segments that contain relevant answers to clinical questions. The capability enhances the accessibility of critical medical knowledge and enables more efficient retrieval from vast amounts of recorded medical procedures, lectures, and diagnostic sessions.

Compared to answer localization in common non-medical domain videos [3, 7, 25], medical instructional videos introduce unique challenges due to the intricate, domain-specific content and terminology. These videos often feature detailed and highly technical actions, such as surgical procedures or rehabilitation exercises that require a deep understanding of the specific medical domain. Moreover, the textual questions often

* The first two authors contribute equally to this work. †: Corresponding Author.

involve specialized terminology and ambiguous concepts, making it harder to match them directly with visual content.

Existing methods for video moment localization typically rely on supervised learning [6, 7, 11, 13], which demands a large amount of annotated data to train. However, for medical instructional videos, obtaining large-scale labeled data is costly and time-consuming, as it requires expert annotations that are precise and professional. Therefore, zero-shot approaches, especially those focus on exploiting the rich pretrained knowledge in large language models (LLMs), are more favorable. However, while LLMs are equipped with rich knowledge, it is highly challenging to effectively invoke them for medical video analysis. In common (non-medical) video domains, typical methods achieve zero-shot moment localization by either cross-modal matching between moment proposals and queries [1, 12, 23, 26, 27, 30] or adopting Socratic approaches [28] that converting video into timestamp-aware visual descriptions to prompt LLMs [2, 14, 22, 24, 29]. Despite their success, it is challenging to directly apply them to the medical video domain. In the former, matching between question and video content is difficult because of specialized terminology and professional actions in medical video QA. The latter approaches fall short in generating domain-specific descriptions about fine-grained motions with specialized terminology.

Aside from the visual and description data, in this paper, we highlight the importance of subtitles in instructional medical videos, since subtitles often provide fine-grained and time-attached textual cues that reflect the key procedural transitions and also contain medical terminology critical for understanding medical workflows. Unlike the methods introduced in [5, 9] that extract subtitles for model training, we adopt a training-free approach that directly feed the subtitles along with tailored prompts for LLMs to achieve temporal answer localization. A key challenge is that subtitles and corresponding visual content are often not well aligned. Like other instructional videos, we find that the spoken information and corresponding visual demonstration in medical instructional videos are often out of sync. To tackle this challenge, we propose a subtitle refinement strategy that aligns subtitles with visual frames based on their cross-modal semantic similarities for better synchronization.

While video subtitles help to mitigate the domain gap between medical query and video content, they do not fully represent the video. Thus, we further design a fusion strategy for combining subtitles and visual information. Specifically, we first employ a multimodal large language model (MLLM) (*i.e.*, Qwen2-VL [21]) to describe the visual content into textual descriptions. The descriptions then represent the video visual information and are integrated with the refined subtitles to form a token sequence to be fed into LLM for answer span prediction.

Our experiments demonstrate the effectiveness of this approach on two medical VideoQA datasets: MedVidQA [4] and COIN-Med [20]. We show that such a training-free approach significantly outperforms the previous state-of-the-art (SOTA) zero-shot methods by a whopping 20% to 41% of mIoU. It even surpasses the recent fully-supervised method, highlighting the potential of our Socratic approach by composing MLLMs and LLMs to tackle this challenging task. Our contributions are as follows:

- For the first time, we unleash the strong power of LLMs for zero-shot medical instructional video answer localization, paving the way for more scalable and training-free instructional medical video analysis.
- We propose a subtitle refinement strategy to address misalignment between subtitles and visual contents and an early fusion strategy to jointly consider subtitle and visual information, improving the accuracy of temporal answer localization.

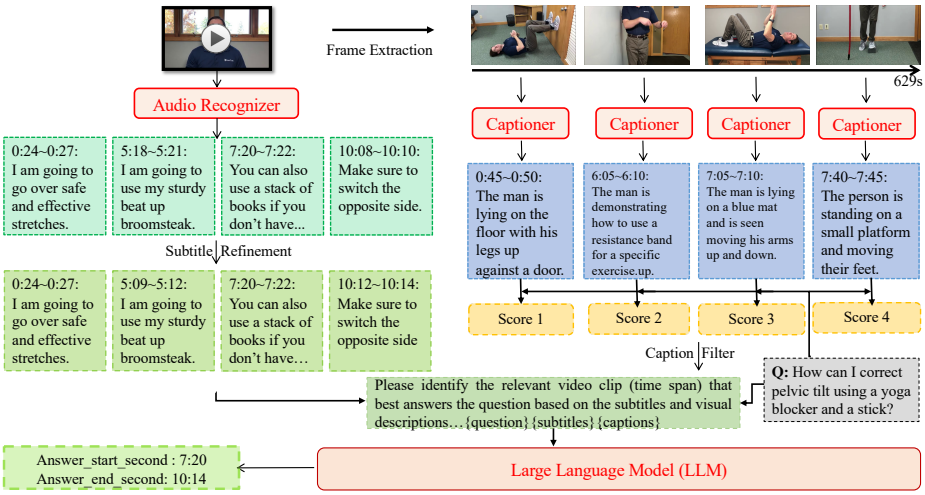


Fig. 1. Our approach for zero-shot instructional medical video answer localization.

– Our training-free method significantly surpasses previous zero-shot SOTA results, and even outperforms the recent supervised approaches.

2 Methodology

2.1 Overview

Given an untrimmed medical instructional video V of duration T seconds and a natural language question Q, the task is to locate in the video a relevant segment $V_{b,e}$ (where $0 \leq b \leq e \leq T$) that can visually answer the question. The problem can be formally defined as:

$$b, e = \phi(V, Q), \tag{1}$$

where ϕ denotes the answer localization method. To solve the problem, we propose SubGPT, a Socratic approach [28] that emphasizes the off-the-shelf use of LLMs and their multimodal variants MLLMs to accomplish answer localization. As illustrated in Fig. 1, our method first extracts from the video a set of subtitles and visual descriptions (accompanied with their corresponding time-stamps) using pretrained MLLMs (*e.g.*, Whisper [16] as audio recognizer and Qwen2-VL [21] as captioner). We then refine the subtitles to be more temporally aligned with the visual contents, and also filter the redundant captions under the instruction of the question. Finally, both the refined subtitles and the cleaned visual descriptions are early fused and fed into LLMs (*e.g.*, GPT-4) for answer localization. At the core of this process, we highlight the strength of a subtitle refinement mechanism and an early fusion strategy for better performance. We elaborate on the details below.

2.2 Subtitle Refinement

For medical instructional videos, discrepancies arise between the speech of the speakers and the corresponding visual frames. For example, an instructional step demonstrated

by the presenter may run ahead or lag behind the corresponding verbal explanation. Thus, to mitigate possible shift of localiation windows resulted from unaligned subtitles, synchronizing the subtitles with the visual contents is of crucial importance. In this section, we introduce a flexible subtitle refinement strategy catering to the difference of video subtitle density across different datasets.

Dense Subtitle Refinement For the video with rich subtitles, we synchronize the subtitles with the corresponding visual contents locally, based on our observation that the subtitles are often close though not well-aligned with the visual contents. Formally, let $S = \{s_1, s_2, \ldots, s_T\}$ be a sequence of subtitles, each associated with a timestamp t_i. For each subtitle s_i at timestamp t_i, we define an adjacent temporal window $[t_i - N, t_i + N]$ (where N is a predefined duration). Within this window, we compute the EMScore [19] $E(s_i, v_j)$ between the subtitles and the video segment v_j at time t_j. Note that EMScore is originally proposed to evaluate the quality of video descriptions by calculating both coarse-grained (sentence-video) and fine-gained (word-frame) cross-modal similarity between textual captions and visual content. Here we repurpose it to obtain the cross-modal similarity between subtitles and video segment. The refined timestamp t_i^* for subtitle s_i is thus determined via

$$t_i^* = \arg \max_{t_j \in [t_i - N, t_i + N]} E(s_i, v_j). \tag{2}$$

This process maximally ensures that each subtitle is aligned with its most relevant visual content, thereby facilitating the subsequent video answer localization via subtitles.

Sparse Subtitle Refinement For the video with sparse subtitles, we synchronize the subtitles with the corresponding visual contents globally, since we find that the problem of temporal misalignment between the visual contents and subtitles is more serious. Formally, let $S = \{s_1, s_2, \ldots, s_T\}$ denote the sequence of subtitles, and $V = \{v_1, v_2, \ldots, v_M\}$ represent a sequence of video frames extracted from the whole video. We compute the EMScore between each subtitle and the temporally continuous video frames to construct a cost matrix D, where each element $D(i, j : j + m)$ signifies the distance between subtitle s_i and a segment of m consecutive video frames starting from v_j. we identify the optimal alignment path $\pi = \{(i_k, j_k)\}_{k=1}^{K}$ (where i_k and j_k denote the respective index of a subtitle and video frame at alignment step k) that minimizes the cumulative distance by applying the DTW algorithm to this cost matrix:

$$\mathrm{DTW}(S, V) = \min_{\pi} \left(\sum_{k=1}^{K} D(i_k, j_k : j_k + l) \right), \tag{3}$$

subject to boundary and monotonicity constraints. This approach ensures precise alignment of subtitles with video content while maximally preserving the global temporal order. The output of this step is an optimized alignment path π that assigns each subtitle s_{i_k} to the most semantically relevant video segment $(j_k, j_k + l)$, ensuring precise synchronization between textual and visual information for downstream tasks.

2.3 Early Fusion of Subtitles and Visual Descriptions

As aforementioned, subtitles alone cannot cater to the queries that necessitate fine-grained motion and visual appearance information. Therefore, we further extract visual

Table 1. Dataset statistics.

Datasets		# Vid.	# Ques.	Video Len. (s)	Seg. Len. (s)	Avg. Subtitles
MedVidQA [4]	Val	47	145	333.9	66.8	86.6
	Test	48	155	345.8	56.9	94.7
COIN-Med [20]	Val	101	173	138.3	31.0	45.1
	Test	110	173	140.9	37.8	43.5

appearance descriptions from the videos, which complement the subtitles for better answer localization. As illustrated in the right part of Fig. 1, we evenly partition the video into non-overlapping clips of length L. Each video clip is then input into pretrained MLLM (captioner) to generate corresponding textual descriptions. Before sending the captions to LLM, we further remove some captions of lower relevance (with a threshold τ) to the question to limit the length of prompt for efficiency. This is achieved by calculating the cosine similarity score between the query and each caption based on their CLIP representations. Noteworthy, the similarity scores of the retained captions are attached along with the captions to serve as additional indications, since we find that this can slightly boost the performance.

Finally, the subtitles and the cleaned captions are fed into LLM along with the question for answer localization. More concretely, given the filtered caption set C', the refined subtitles S', and the query Q. We prompt the LLM using the a instruction like "`You are an expert specializing in analyzing medical instructional videos. Please identify the timestamp that best answers the question based on the video captions and subtitles. Question:` $\{Q\}$. `Subtitles:` $\{S'\}$. `Captions:` $\{C'\}$." The full prompt will be released along with the code. This methodology effectively integrates visual and textual clues for LLM decision making, effectively boosting the performance of medical video question-answering tasks.

3 Experiment

3.1 Datasets

MedVidQA [4] contains 3,010 curated healthcare questions, each linked to a visual answer segment from 899 medical instructional videos sourced from reputable institutions, including accredited medical schools, health organizations and medical professionals. **COIN-Med** is a medical subset of the COIN dataset [20]. It contains 566 curated healthcare questions, each linked to a visual answer segment from 393 medical instructional videos. The medical subset is selected by filtering the videos about "Nursing" (such as "PerformCPR", "UseEpinephrineAuto-Injector", "BandageHead" and etc.), based on the original class labels provided in COIN. For both datasets, we only use the validation and test sets for hyperparameter (and prompt) tuning and model evaluation respectively. Other related details are presented in Table 1. It is worth mentioning that the video subtitles in MedVidQA are denser than that of COIN-Med.

3.2 Implementation Details

For subtitle extraction (or audio recognition), we use Whisper [16] and the YouTube Transcript API to obtain subtitles for videos in the MedVidQA and COIN-Med re-

Table 2. Comparison of different VQA models on the MedVidQA and COIN-Med.

Models	MedVidQA				COIN-Med			
	mIoU	IoU@0.3	IoU@0.5	IoU@0.7	mIoU	IoU@0.3	IoU@0.5	IoU@0.7
Supervised Methods								
VPTSL [9]	57.8	77.4	61.9	44.5	22.1	30.1	18.7	9.4
Zero-shot Methods								
TimeChat [18]	3.8	4.2	2.1	0.7	3.1	4.9	3.3	0.8
VTG-GPT [26]	12.1	16.8	8.3	3.2	18.1	25.9	16.1	4.3
TFVTG [30]	17.0	19.9	11.2	9.1	16.2	21.3	10.3	7.7
SubGPT-mini (Ours)	53.8	70.6	54.6	39.2	35.9	47.4	33.0	18.5
SubGPT (Ours)	58.0	76.9	63.6	44.8	38.4	50.9	36.4	21.4

spectively, which we find is the best practice for each dataset. For subtitle refinement, the temporal window N is set to 5s, which is searched on MedVidQA validation set. Moreover, we sample the video at 6 fps within the window and utilize CLIP [15] to extract frame and text embeddings to calculate the EMscore. Additionally, we apply dense subtitle refinements to MedVidQA and sparse subtitle refienment to COIN-Med according to their dense and sparse subtitle densities respectively. For visual captioning, a video is partitioned into non-overlapping clips of length $L = 5$s. To caption each clip, we extract at 6 fps and assess multiple MLLMs, such as Video-LLaVA [10], LLaVA-OV [8] and Qwen2-VL-Instruct [21], and find that Qwen2-VL-7B-Instruct [21] to be the best for this task. The threshold τ for caption filtering is set to 0.25. For LLM reasoning, we use both GPT-4o-mini and GPT-4o for their well-known powerful reasoning and instruction following capabilities. Unless otherwise specified, all results are averaged over three runs.

3.3 Comparison with the State-of-the-Arts

TimeChat [18] is an end-to-end general-purpose Video-LLM developed for long video understanding. In this paper, we adapt it for zero-shot temporal grounding in medical domain. Like us, VTG-GPT [26] and TFVTG [30] are two zero-shot video temporal grounding methods based on large foundation models. VTG-GPT adopts a cross-modal matching solution to find the best temporal proposal that matches the query in feature space (via sentence-BERT [17]). TFVGT uses a similar idea but improves over VTG-GPT by introducing LLM (GPT-4 Turbo) to parse the query into dynamic and static sub events for fine-grained cross-modal matching. Additionally, VPTSL[9] is a supervised model which also highlights the importance of subtitles for instructional video analysis. Since these methods (except for VPTSL) do not report results on MedVidQA and COIN-Med, we reproduce the results with their official code.

Table 2 presents the performance comparison on the MedVidQA and COIN-Med datasets. From the experimental results, we can observe the followings:

(1) Our method (SubGPT and SubGPT-mini) significantly outperforms the recent SOTA zero-shot methods on both datasets, and this strength is stable across different metrics. It is worth highlighting that both VTG-GPT and TFVTG use powerful LLMs (GPT-4 turbo and Baichuan2), but seem in untended ways for medical video analysis given their poor behavior. Additionally, the end-to-end model TimeChat [18] shows the worst adapting performance in the medical domain, indicating the severe overfitting problem for general-domain video temporal grounding. In fact, the case study

Table 3. Model ablation on MedVidQA and COIN-Med.

Configuration	Modality		LLM	MedVidQA				COIN-Med			
	Cap.	Sub.		mIoU	IoU@0.3	IoU@0.5	IoU@0.7	mIoU	IoU@0.3	IoU@0.5	IoU@0.7
SubGPT	✓	✓	4o	58.0	76.9	63.6	44.8	38.4	50.9	36.4	21.4
SubGPT-mini	✓	✓	4o-mini	53.8	70.6	54.6	39.2	35.9	47.4	33.0	18.5
w/o Sub. Refine	✓	✓	4o-mini	52.9	69.2	53.8	38.3	34.5	46.2	32.4	17.9
w/o Caption		✓	4o-mini	52.5	69.9	52.5	37.8	34.4	45.1	29.5	17.3
Post-Fusion	✓	✓	4o-mini	51.7	68.5	52.6	37.7	33.9	43.9	30.1	15.6
Raw subtitle alone		✓	4o-mini	51.0	67.7	51.6	33.6	32.7	42.2	28.9	15.0
Raw caption alone	✓		4o-mini	19.8	28.0	16.1	9.1	19.4	26.5	16.9	9.6

in Fig. 2(c) suggests that the end-to-end model TimeChat tends to localize totally irrelevant video segments in zero-shot cross-domain generalization, while other cross-modal matching based methods can obtain coarse localization results, and our Socratic approach achieves the best result.

(2) Our method even surpasses the supervised SOTA method VPTSL on the COIN-Med dataset. However, we find that our model variant SubGPT-mini (using GPT-4o mini) underperforms VPTSL on MedVidQA, reflecting the importance of a strong LLM for our success. Since VPTSL also emphasizes the use of subtitle information, we attribute our strength of SubGPT to the additional subtitle refinement and captioning modules, along with GPT-4o level intelligence for question answering. The above findings demonstrate the remarkable superiority of our solution, which suggests a promising direction for medical video analysis in the LLM era.

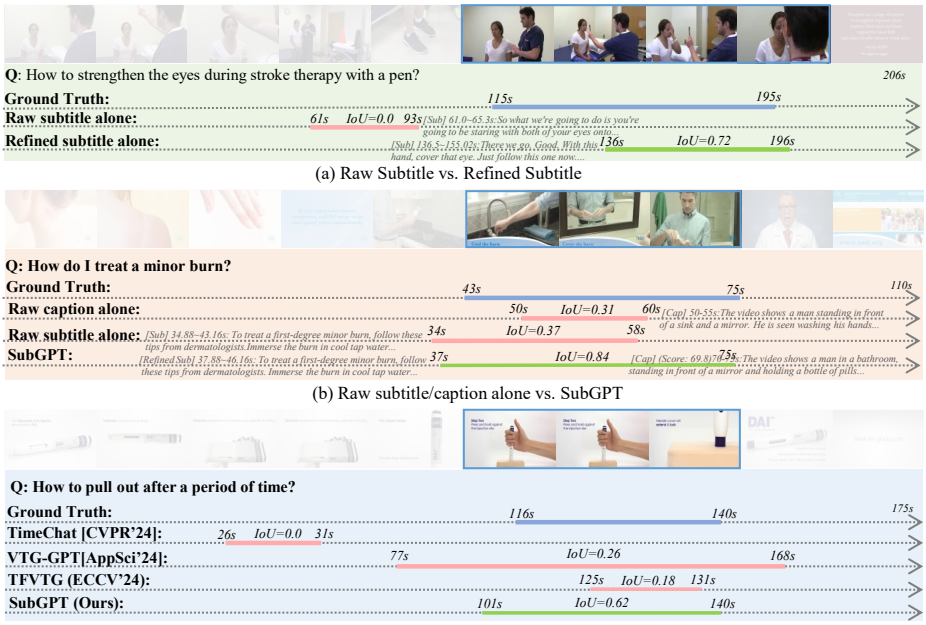
3.4 Ablation Studies

In this section, we comprehensively analyze the impact of different configurations of our method. We first substitute GPT-4o with a faster and more efficient version GPT-4o mini, and observe a sharp performance drop. For example, the mIoU drops by 4.1% and 3.5% on MedVidQA and COIN-Med respectively. This indicates the importance of a powerful LLM for precise long-context reasoning. We then remove the subtitle refinement and captioning module, and find that the performances under different metrics decline by ∼1% and ∼1.4% respectively, suggesting the effectiveness of our subtitle refinement strategy and the complementary effect of subtitles and captions (examples shown in Fig. 2(a) and (b)). We also investigate a post-fusion method to combine the independent answer localization results of subtitles and captions, and find that the performance is even worse than that of using the refined subtitles alone. This reflects the importance of our early fusion approach. Finally, we study two naive baselines by separately feeding the raw subtitles and captions into LLM and find that the mIoUs decline by ∼3% on both datasets for using raw subtitles alone. Meanwhile, the performances are extremely worse if we use only the captions, likely because of a domain gap between common visual descriptions and professional queries.

Given the differences in subtitle density and video length between the MedVidQA and COIN-Med datasets, we apply dense and sparse subtitle refinement methods for different datasets. For the MedVidQA, which features higher subtitle density and longer videos, we used the dense refinement method. Because High-density subtitles typically contain more detailed textual information, resulting in shorter time discrepancies between subtitles and visual content. On the other hand, for the COIN-Med dataset, which has lower subtitle density and shorter videos, we employ the sparse refinement

Table 4. Results of different subtitle refinement strategies.

Methods	MedVidQA				COIN-Med			
	mIoU	IoU@0.3	IoU@0.5	IoU@0.7	mIoU	IoU@0.3	IoU@0.5	IoU@0.7
Baseline	51.0	67.7	51.6	33.6	32.7	42.2	28.9	15.0
Dense Refine	**52.5**	**69.9**	**52.5**	**37.8**	33.5	44.5	28.9	16.8
Sparse Refine	51.9	68.5	52.4	35.9	**34.4**	**45.1**	**29.5**	**17.3**



(a) Raw Subtitle vs. Refined Subtitle

(b) Raw subtitle/caption alone vs. SubGPT

(c) Previous SOTA methods vs. SubGPT (Ours)

Fig. 2. Visualization of answer localization.

method. Table 4 shows that this adaptive strategy effectively brings the optimal results on both datasets.

4 Conclusion

This paper presents a novel training-free approach for question-answer localization in medical instructional videos. Our method highlights the critical role of video subtitles and visual descriptions in enabling large language models (LLMs) to perform effective temporal localization. To this end, we introduce a subtitle refinement strategy and an early fusion mechanism that integrates subtitles and visual captions for improved performance. Experiments on two standard Medical VideoQA datasets show that our approach achieves new state-of-the-art results. Extensive ablation studies further validate the importance of powerful LLMs in temporal reasoning, the effectiveness of our subtitle alignment strategy, and the complementary benefits of incorporating visual

captions. With these efforts, we hope this work offers new insights into harnessing LLMs for instructional medical video analysis.

Disclosure of Interests

The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Diwan, A., Peng, P., Mooney, R.: Zero-shot video moment retrieval with off-the-shelf models. In: Transfer Learning for Natural Language Processing Workshop. pp. 10–21. PMLR (2023)
2. Fan, Y., Ma, X., Wu, R., Du, Y., Li, J., Gao, Z., Li, Q.: Videoagent: A memory-augmented multimodal agent for video understanding. In: European Conference on Computer Vision. pp. 75–92. Springer (2024)
3. Gao, J., Sun, C., Yang, Z., Nevatia, R.: Tall: Temporal activity localization via language query. In: Proceedings of the IEEE international conference on computer vision. pp. 5267–5275 (2017)
4. Gupta, D., Attal, K., Demner-Fushman, D.: A dataset for medical instructional video classification and question answering. Scientific Data **10**(1), 158 (2023)
5. Gupta, D., Demner-Fushman, D.: Overview of the medvidqa 2022 shared task on medical video question-answering. In: Proceedings of the 21st Workshop on Biomedical Language Processing. pp. 264–274 (2022)
6. Jung, M., Xiao, J., Zhang, B.T., Yao, A.: On the consistency of video large language models in temporal comprehension. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 13713–13722 (2025)
7. Lei, J., Berg, T.L., Bansal, M.: Detecting moments and highlights in videos via natural language queries. Advances in Neural Information Processing Systems **34**, 11846–11858 (2021)
8. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al.: Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326 (2024)
9. Li, S., Li, B., Sun, B., Weng, Y.: Towards visual-prompt temporal answer grounding in instructional video. IEEE Transactions on Pattern Analysis & Machine Intelligence pp. 1–18 (2024)
10. Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P., Yuan, L.: Video-llava: Learning united visual representation by alignment before projection. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 5971–5984 (2024)
11. Lin, K.Q., Zhang, P., Chen, J., Pramanick, S., Gao, D., Wang, A.J., Yan, R., Shou, M.Z.: Univtg: Towards unified video-language temporal grounding. In: IEEE/CVF International Conference on Computer Vision. pp. 2794–2804 (2023)
12. Luo, D., Huang, J., Gong, S., Jin, H., Liu, Y.: Zero-shot video moment retrieval from frozen vision-language models. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5464–5473 (2024)
13. Meinardus, B., Batra, A., Rohrbach, A., Rohrbach, M.: The surprising effectiveness of multimodal large language models for video moment retrieval. arXiv preprint arXiv:2406.18113 (2024)

14. Qin, H., Xiao, J., Yao, A.: Question-answering dense video events. arXiv preprint arXiv:2409.04388 (2024)
15. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
16. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. In: International conference on machine learning. pp. 28492–28518. PMLR (2023)
17. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 3982–3992 (2019)
18. Ren, S., Yao, L., Li, S., Sun, X., Hou, L.: Timechat: A time-sensitive multimodal large language model for long video understanding. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14313–14323 (2024)
19. Shi, Y., Yang, X., Xu, H., Yuan, C., Li, B., Hu, W., Zha, Z.J.: Emscore: Evaluating video captioning via coarse-grained and fine-grained embedding matching. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 17929–17938 (2022)
20. Tang, Y., Ding, D., Rao, Y., Zheng, Y., Zhang, D., Zhao, L., Lu, J., Zhou, J.: Coin: A large-scale dataset for comprehensive instructional video analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1207–1216 (2019)
21. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
22. Wang, X., Zhang, Y., Zohar, O., Yeung-Levy, S.: Videoagent: Long-form video understanding with large language model as agent. In: European Conference on Computer Vision. pp. 58–76. Springer (2024)
23. Wattasseril, J.I., Shekhar, S., Döllner, J., Trapp, M.: Zero-shot video moment retrieval using blip-based models. In: International Symposium on Visual Computing. pp. 160–171. Springer (2023)
24. Xiao, J., Huang, N., Qin, H., Li, D., Li, Y., Zhu, F., Tao, Z., Yu, J., Lin, L., Chua, T.S., et al.: Videoqa in the era of llms: An empirical study. International Journal of Computer Vision pp. 3970–3993 (2025)
25. Xiao, J., Yao, A., Li, Y., Chua, T.S.: Can i trust your answer? visually grounded video question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13204–13214 (2024)
26. Xu, Y., Sun, Y., Xie, Z., Zhai, B., Du, S.: Vtg-gpt: Tuning-free zero-shot video temporal grounding with gpt. Applied Sciences **14**(5), 1894 (2024)
27. Xu, Y., Sun, Y., Zhai, B., Li, M., Liang, W., Li, Y., Du, S.: Zero-shot video moment retrieval via off-the-shelf multimodal large language models. arXiv preprint arXiv:2501.07972 (2025)
28. Zeng, A., Attarian, M., Choromanski, K.M., Wong, A., Welker, S., Tombari, F., Purohit, A., Ryoo, M.S., Sindhwani, V., Lee, J., et al.: Socratic models: Composing zero-shot multimodal reasoning with language. In: The Eleventh International Conference on Learning Representations (2023)
29. Zhang, C., Lu, T., Islam, M.M., Wang, Z., Yu, S., Bansal, M., Bertasius, G.: A simple llm framework for long-range video question-answering. In: Proceedings of

the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 21715–21737 (2024)
30. Zheng, M., Cai, X., Chen, Q., Peng, Y., Liu, Y.: Training-free video temporal grounding using large-scale pre-trained models. In: European Conference on Computer Vision. pp. 20–37. Springer (2024)