TOGA: Temporally Grounded Open-Ended Video QA with Weak Supervision

Ayush Gupta[1,2], Anirban Roy[1], Rama Chellappa[2], Nathaniel D. Bastian[3], Alvaro Velasquez[4], Susmit Jha[1]
[1]SRI [2]Johns Hopkins University [3]United States Military Academy [4]University of Colorado Boulder

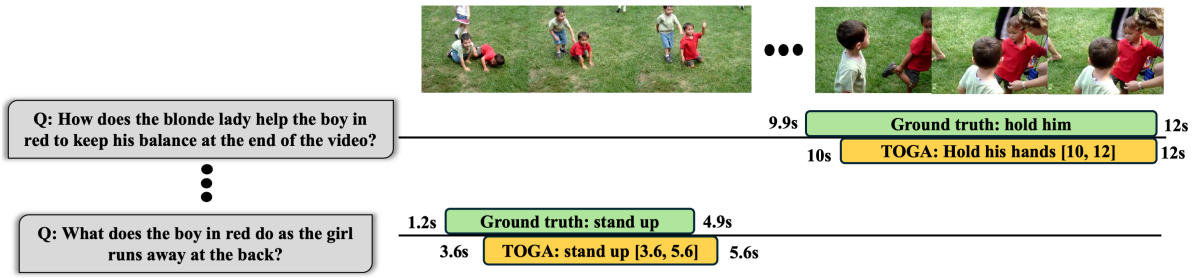

Figure 1. We present an approach for open-ended grounded video question answering. Given a video and open-ended questions, we generate open-ended responses with the grounding as `Answer [start time, end time]`. We consider long videos with multiple questions and answers per video. The questions refer to the interaction between multiple actors and the temporal ordering of events.

Abstract

We address the problem of video question answering (video QA) with temporal grounding in a weakly supervised setup, without any temporal annotations. Given a video and a question, we generate an open-ended answer grounded with the start and end time. For this task, we propose TOGA : a vision-language model for Temporally Grounded Open-Ended Video QA with Weak Supervision. We instruct-tune TOGA to jointly generate the answer and the temporal grounding. We operate in a weakly supervised setup where the temporal grounding annotations are not available. We generate pseudo labels for temporal grounding and ensure the validity of these labels by imposing a consistency constraint between the question of a grounding response and the response generated by a question referring to the same temporal segment. We notice that jointly generating the answers with the grounding improves performance on question answering as well as grounding. We evaluate TOGA on grounded QA and open-ended QA tasks. For grounded QA, we consider the NExT-GQA benchmark, which is designed to evaluate weakly supervised grounded question answering. For open-ended QA, we consider the MSVD-QA and ActivityNet-QA benchmarks. We achieve state-of-the-art performance for both tasks on these benchmarks.

Work partly done during Ayush Gupta's internship at SRI.

1. Introduction

We propose a weakly-supervised framework for grounded video question answering (videoQA). This framework generates open-ended, free-form sentence answers to video-based questions, providing temporal grounding with start and end times (as shown in Fig. 1). We achieve this without relying on expensive temporal annotations.

Grounded video question answering is challenging as, in addition to generating correct answers, it requires localizing evidence to support the answer [49]. The challenge is more prominent in a weakly supervised setup where grounding annotations are unavailable. Since each video can have multiple questions with overlapping temporal groundings, each answer needs to be grounded with a distinct temporal window. We consider relatively long videos with an average length of 40 seconds consisting of complex causal and temporal questions [49]. Generating correct answers to these questions requires understanding the temporal ordering of events and the spatiotemporal interaction between actors and objects. Further, we consider open-ended evaluation instead of a multiple-choice QA setup like previous works [54, 56], further enhancing the challenge.

We propose TOGA - a vision-language model (VLM) for the grounded videoQA task to address these challenges. TOGA builds a video processing framework and combines that with a large language model (LLM)-based text processing framework to process questions and generate

open-ended answers. We instruction-tune TOGA to jointly generate answers with temporal groundings in the format: `Answer [start time, end time]`. Given a video and a question, we sample video frames and compute visual features using a pre-trained vision transformer encoder [34]. The question is processed with the LLM tokenizer and embedding layer [16] to capture text features. Inspired by previous work [8, 12, 18, 46], we utilize a multi-scale vision-language connector (MS-VLC) to align the video and text features. MS-VLC processes the video at two granularities: one at a low frame rate to capture low-frequency temporal features and another at a high frame rate to capture high-frequency temporal features. Finally, we instruction-tune the LLM decoder [16] to jointly generate open-ended answers with temporal grounding by leveraging the cross-attention between text features and the multi-scale video features. Jointly generating answers with groundings allows capturing the dependency between an answer and the corresponding grounding duration, improving QA and grounding accuracy compared to the approaches that independently predict answers and their grounding [52, 54, 56].

To operate in a weakly supervised setup without grounding labels, we propose a multi-stage training approach. Firstly, we train TOGA to generate answers without grounding by leveraging the question-answer annotations and video descriptions. Then, this model is used to generate pseudo-labels for temporal grounding. We select temporal segments with specific starting and ending times and ask questions corresponding to each segment. The answers, along with the selected starting and ending times, are considered noisy grounding labels for the corresponding questions. Next, we instruction-tune the model to accept prompts with temporal references such as `What is the activity in [10, 20]?` and produce responses with temporal grounding `A boy in a red shirt is running [10, 20]`. Instruction tuning extends the model's abilities to accept temporal references in questions and produce grounding predictions in the answers. However, the grounding performance is limited due to noisy temporal labels. We improve the grounding performance by imposing a consistency constraint while generating the pseudo labels. We select the labels where the answer with a temporal grounding matches a question with the same temporal segment as the reference. For example, let's assume the temporally grounded answer to the question `What is the boy in a red shirt doing?` is `The boy is running [10, 20]`. Then the answer to a corresponding referring question `What is happening in [10, 20]?` is expected to match with `A boy in a red shirt is running`. Maintaining consistency between the answers to the referring and grounding questions is crucial in a weakly supervised setup, as shown in our ablation studies.

TOGA has several advantages over existing approaches. We jointly generate answers with the groundings. Thus, the model can adjust the temporal duration based on the answers by capturing the correlation between the answer and the grounding. Approaches making independent predictions [52, 54, 56] may lack this ability. We instruction-tune the language decoder to generate open-ended answers. Thus, we are not limited to restrictive formats of answers, such as choosing an answer from multiple choices or answering with single-word responses. Finally, we use consistency constraints to generate pseudo-labels for training with weak supervision. TOGA does not rely on temporal annotations or external models to generate annotations [41].

Our main contributions include:

- We propose TOGA, a large vision-language model for open-ended grounded videoQA. TOGA jointly generates open-ended answers with temporal grounding.
- TOGA operates in a weakly supervised setup where groundings annotations are unavailable. We train the model with reliable pseudo labels by imposing consistency between the answers to temporal grounding and temporal referring questions.
- We evaluate our approach on weakly supervised temporal grounding and video QA tasks and achieve state-of-the-art performance.

2. Related Work

Video question answering. Video QA aims to answer questions related to the visual content in videos. Compared to image visual question answering [1, 10], videoQA requires capturing the temporal evolution of scenes. Several approaches are proposed to address this task [7, 9, 14, 15, 19, 21, 22, 38, 47, 59]. Some approaches only utilize visual cues for answering [14, 15, 47]. However, others utilize additional modalities, like transcripts or subtitles of videos, or the movie plot [21, 22, 38]. External knowledge bases [31, 36] can be effective for the task [7, 9]. However, most videoQA approaches commonly consider a multiple-choice setup, where the task is to select a candidate answer from a set of predefined options. We consider an open-ended setup and generate free-form answers.

Open-ended video QA More recently, with the advancement of LLMs [32, 39, 40], videoQA approaches aim to generate open-ended answers [5, 23, 25, 28, 37, 42, 43, 48, 52, 57]. These methods combine vision module and language models to develop general-purpose VLMs [5, 23, 25, 28, 57]. These models are typically trained on large-scale datasets with various multimodal tasks. Yang et al., [52] develop specialized models that are fine-tuned for the target video QA task. There are parameter-free methods that use the collaboration between multiple VLM agents [37, 42, 43] for the video QA task. However, many current approaches are not able to provide evidence for the generated answers

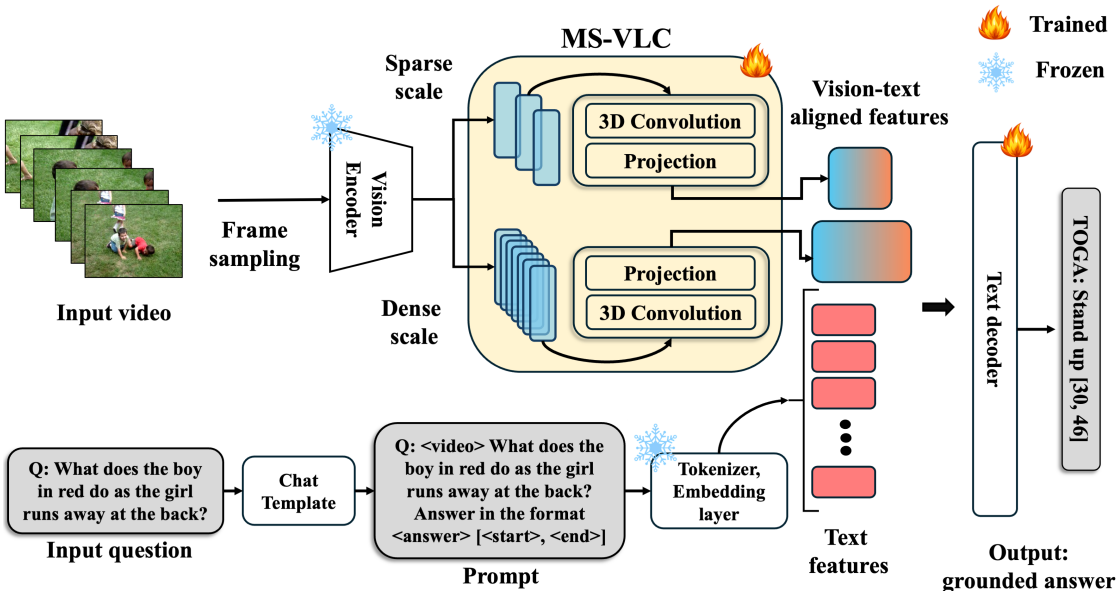

Figure 2. We propose a VLM framework for grounded videoQA. Given an input video, we sample frames and compute framewise features using a vision encoder. A multi-scale vision language connector (MS-VLC) is employed to process the framewise features at two temporal resolutions: a sparse and a dense scale. The input question is processed with a chat template according to the desired prompt format. The prompt is processed through a tokenizer and a frozen language encoder to generate tokenwise text features. The MS-VLC module is trained to align the multiscale vision and text features. Finally, the language decoder is trained to generate answers with temporal grounding.

in the video in the form of grounding and may be relying on language priors instead of true visual reasoning [48]. We focus on addressing this issue by grounding the answers in videos with temporal segments.

Grounded video QA. Grounded video QA aims to generate answers and provide evidence as a timestamp. Some works address this by exploring the similarity between visual and textual content [21, 22, 53]. However, these approaches tend to be biased towards localizing subtitles in TV shows [21] or only being able to deal with a few objects [53]. Some recent works leverage the power of VLMs for this task [12, 33, 41, 45], but these need expensive temporal annotations for training, or generate grounding annotations externally. To overcome this, some works use a weakly supervised setup where temporal annotations are not available [49, 54, 56]. However, these approaches consider a multiple-choice setup, with potential answers provided to the model during evaluation.

Unlike existing approaches, TOGA aims to generate temporally grounded, open-ended answers in a weakly supervised setup without ground truth temporal annotations.

3. Approach

Our VLM framework has four main modules: 1) a vision encoder to compute frame-wise features from videos, 2) a large-language text encoder to compute text features from questions, 3) a multi-scale vision-language connector (MS-VLC) to align the vision and text features facilitating answer generation, and 4) a large-language text decoder to generate open-ended answers with grounding as `Answer [start time, end time]`. Our approach is trained using our proposed multi-stage training framework leveraging consistency between predictions to learn the grounding ability without temporal labels. Among the four components, vision and text encoders are kept frozen while the MS-VLC and text decoder are trained, like previous works [26]. The framework is shown in Fig. 2. We describe the components below.

Vision encoder. Given a video, we uniformly sample frames and compute frame-wise features. We choose the CLIP [34] vision encoder that is trained with large-scale vision-language datasets. We keep this encoder frozen as it already generates the text-aligned features.

Text encoder. We process the question with an LLM text encoder. The text encoder combines the tokenizer and the embedding layer of an LLM. This module generates text features that feed to the text decoder. Following the common practice [5, 25, 26], we keep this module frozen as it is pre-trained on large and diverse text datasets.

Multi-scale vision-language connector (MS-VLC). Given the framewise features, we aim to capture the temporal cues to generate answers with temporal grounding. This module is trained to generate video features that are aligned with the text features derived from the question. The MS-VLC module processes the framewise features at two temporal resolutions: a sparse scale with a low frame rate to

<table>
<tr><td>Goal:</td><td>Vision-text alignment</td><td>Instruction tuning for grounding</td><td>Instruction tuning and consistent grounding</td></tr>
<tr><td>Trained modules:</td><td>MS-VLC</td><td>MS-VLC & language decoder</td><td>MS-VLC & language decoder</td></tr>
<tr><td rowspan="2">Prompts:</td><td rowspan="2">User prompts:

Q: Provide a brief summary of the video.

Q: Give a concise description of the events in the video.

Q: Relay a brief, clear account of the clip.</td><td colspan="2">System prompt:
<<SYS>> The assistant gives helpful, detailed, and polite answers to the human's questions after analyzing the input image or video. The assistant is able to ground objects and events in videos by providing timestamps. The user may refer to certain sections of the video by similar timestamps. The user requests an answer using <answer>, and a timestamp using [<normalized start>, <normalized end>]. The assistant provides only what the user asks for.
<</SYS>></td></tr>
<tr><td>User prompts:

Q_g: Point out the scene where a person is holding a glass. Answer in the format [<normalized start>, <normalized end>].
A: [0, 29]

Q_r: Describe the scene in [45, 72]. Answer in the format <answer>.
A: A person opens a refrigerator.</td><td>User prompts:

Q_g: What is the boy doing? Answer in the format <answer> [<normalized start>, <normalized end>].
A: play with toy [25, 73]

Q_r: What is happening in [25, 73]? Answer in the format <answer>.
A: play with toy

GT: play with toy</td></tr>
</table>

Figure 3. We train TOGA in three stages. Each stage focuses on a specific task with prompts designed for the task. Q_g is a grounding question where a temporal grounding is expected in the response, such as `What is the boy doing?`. Q_r is a referring question focusing on a temporal segment in the video, such as `What is happening in [25, 73]`. We impose consistency on the answers to Q_g and Q_r. These two queries are expected to have the same answer `play with toy`. Further, we ensure the answer matches the ground truth (GT) answer. The multistage training gradually enables question answering, temporal referring, and temporal grounding. Temporal referring and temporal grounding enable grounding with weak supervision.

capture low-frequency temporal cues and a dense scale with a high frame rate to capture high-frequency temporal cues. We sample 4 frames at the sparse scale and 16 frames at the dense scale. The sparse module captures the long-term temporal features suitable for grounding longer segments, while the dense module captures the short-term temporal features suitable for grounding shorter segments. Each VLC block is implemented with RegNet [35] and 3D convolutions [5]. The parameters are shared between the two VLC blocks. Our experiments show that the MS-VLC is crucial for both accurate answering and grounding. Processing time series at multiple scales is shown to be effective for activity recognition in videos [8, 46] and event detection in audio [18].

Text decoder. We train an LLM text decoder to generate answers from multi-scale video features and token-wise text features. We consider Mistral-7B Instruct [16] as the language decoder. This module learns the cross-attention between the multi-scale video features and the text features to generate answers in the desired format. This allows us to jointly generate open-ended answers with temporal grounding, unlike the approaches with frozen text decoder [52, 54, 56] that generate the answer and grounding separately. We train the model on the next token prediction task as commonly used in language models [30, 32, 40].

Multi-stage training. We propose a multi-stage training strategy for temporally grounded Video QA under weak supervision. **First**, we train only the MS-VLC module to align multiscale video and text features similar to [26]. This alignment is crucial for the downstream tasks. Given a question input, the goal is to generate aligned visual features enabling the text decoder to produce corresponding answers. We consider diverse prompt and response pairs for training at this stage, which include video captioning, sentence completion, and question answering. **Second**, we instruction-tune the MS-VLC and the language decoder modules for the grounding task. The main goal of this stage is to train the model to understand prompts with temporal references such as `What does the boy do at [10, 20]?` and produce responses with temporal grounding `A boy in a red shirt is running [10, 20]`. As we do not have temporal annotations in the weakly supervised setup, we generate pseudo-temporal labels for training. We crop temporal segments in videos with known start and end times. We generate descriptions for these segments by considering them as full videos by using the model trained in the previous stage. These descriptions with the selected start and end times are considered as pseudo labels for answers and temporal grounding, respectively. **Finally,** we train for accurate grounding by imposing a consistency constraint between the grounding response and the response generated by a question with the same grounding as input. For example, let's consider a response `Stands up [5, 10]` corresponding to a query `What does the boy in red do after the girl left?` Then we generate a paired question

as `What does the boy do in [5, 10]?` with the same start and end times. We train the model to produce a consistent response `Stands up`. Furthermore, based on the available question-answer annotations, we ensure the answer `Stands up` is accurate. These self-consistent question-answers with temporal labels enable TOGA to improve both answering and grounding accuracy.

Prompt design. We design prompts suitable for the tasks at various stages of training as shown in Fig. 3. We define a special `<video>` token to include visual features with text tokens, like [5, 25]. For vision-text alignment, we consider multiple user prompts to train the MS-VLC module. For instruction tuning, we design prompts to include temporal references. Since we ask the model to output the grounding as text tokens, defining a specific format for the grounding outputs is necessary. This format is provided as part of the prompt in the individual stages of training. Specifically, we include the text 'Answer in the format `<format>`' as part of the user prompt, specifying the format based on the task. We notice that including the output format in the user prompt is important to generate grounding responses. We use the following formats: 1) `answer` when we only generate the answer, 2) `[<start>, <end>]` when we only generate temporal grounding, or 3) `answer [<start>, <end>]` when we generate an answer with grounding.

Inference. We use the same system prompt during inference. We add instructions specifying the format of the response as described in the previous section. For grounded videoQA, we prompt the model to perform both grounding and answering. For open-ended videoQA on MSVD-QA and ActivityNet-QA, we prompt the model to generate only answers. Inference takes 0.6 seconds on average to generate the grounded answers on an A100 GPU.

Implementation details. We use `CLIP-ViT-Large` [34] as the vision encoder and Mistral-7B Instruct [16] as the LLM. MS-VLC comprises two RegNet stages, separated by a 3D conv layer. We sample 16 frames at the dense scale and 4 frames at the sparse scale. In the first vision-text alignment stage, we train for one epoch with a batch size of 256 using the AdamW optimizer [27]. The learning rate is set to 1e-3 for the alignment stage and lowered to 2e-5 for the instruction tuning stages. We use video-text pairs from Video-ChatGPT [28] to train MS-VLC at the first stage of training. The first stage of training takes 54 hours, and the latter two stages take 7 hours each with 8 A100 GPUs.

4. Experiments

We first describe the datasets and metrics, compare TOGA with the state-of-the-art, and perform analysis and ablation studies. Then, we present failure cases and discuss limitations and future directions.

4.1. Datasets and metrics

We evaluate TOGA on four videoQA benchmarks: NExT-GQA [49] and RexTime [3] for grounded QA, MSVD-QA [50] and ActivityNet-QA [55] for open-ended QA.

NExT-GQA [49]. This is designed to evaluate weakly supervised grounded videoQA. Unlike other videoQA benchmarks [13, 44, 51] consisting of short 3-15 second-long videos, NExT-GQA selects long videos with an average length of 40 seconds. Each video has multiple questions, and the answers' grounding can overlap. The videos include a sequence of atomic events, and the questions involve interaction between multiple actors and objects. Questions are of two types: causal why/how questions and temporal questions with when/before/after clauses. Causal questions, such as `Why are there two men standing at the center island and holding their camera?`, require localizing the evidence of the answer, which is `Recording` for this question. Temporal questions, such as `What did the boy do after the green man walked past him?`, require understanding the temporal evolution of the events to generate an answer such as `Look at the man in green`. The train set consists of 3,870 videos and 34,132 QA pairs without the grounding annotations. The test set consists of 990 videos with 5,553 QA pairs. Though the dataset provides multiple-choice answers, we generate open-ended responses without observing the options.

ReXTime [3]. ReXTime is a benchmark designed to evaluate temporal reasoning abilities within video events. It specifically focuses on the challenging scenarios where questions and answers occur in different video segments, necessitating an understanding of cause-and-effect relationships across time. The benchmark comprises 921 validation samples and 2,143 test samples. We utilize this dataset for zero-shot evaluation, and directly evaluate our model on the test samples without any fine tuning on this dataset.

Metrics for grounded videoQA. We consider five metrics: intersection over union (IoU), intersection over prediction (IoP), IoU@0.5, IoP@0.5, and Acc@GQA proposed in NExT-GQA [49]. For ReXTime [3], we use IoU, IoU@0.3 and IoU@0.5. IoU measures the overlap between the ground truth and predicted groundings. IoP measures the portion of predicted grounding containing the ground truth, similar to precision. IoU@0.5 and IoP@0.5 refer to the cases where IoU and IoP ≥ 0.5. While these metrics focus on grounding, Acc@GQA focuses on both the correctness of the answers and correct grounding with IoP ≥ 0.5. Apart from these metrics, we consider the Acc@QA metric for QA performance in the supplementary section.

MSVD-QA [50]. This dataset considers open-ended video QA where the videos are selected from the video-description pairs of the MSVD dataset [2]. QA pairs for a video are generated from the associated descriptions [11].

	Open-ended evaluation	mIoU	mIoP	IoU@0.5	IoP@0.5	Acc@GQA
IGV [24] *(MM '22)*	✗	14	21.4	9.6	18.9	10.2
Temp[CLIP](NG+) [34, 49] *(CVPR '24)*	✗	12.1	25.7	8.9	25.5	16
FrozenBiLM (NG+) [49, 52] *(CVPR '24)*	✗	9.6	24.2	6.1	23.7	17.5
SeViLA [54] *(NeurIPS '23)*	✗	21.7	29.5	13.8	22.9	16.6
LLoVi [56] *(arXiv '24)*	✗	20	37.3	15.3	36.9	24.3
Grounded-VideoLLM [41] *(arXiv '24)*	✗	21.1	34.5	18	34.4	**26.7**
VideoStreaming [33] *(NeurIPS '24)*	✗	19.3	32.2	13.3	31	17.8
TOGA (Ours)	✓	**24.4**	**40.5**	**21.1**	**40.6**	24.6

Table 1. Comparison with the state of the art on NExT-GQA [49]. TOGA improves the state of the art on the grounding metrics. Other approaches select an answer from a fixed set of options, while TOGA generates open-ended answers. Note that [41] uses ground truth labels for temporal grounding from other datasets (e.g., ActivityNet-Captions [20]) and creates grounding labels with GPT-4 guidance. [33] utilizes ground truth temporal labels from Panda-70M [4] to curate grounding labels. TOGA does not use explicit grounding labels.

It includes 1,970 video clips and 50K+ QA pairs.

ActivityNet-QA [55]. This dataset considers open-ended video QA with videos selected from ActivityNet [6]. The videos are collected from YouTube. The dataset consists of 5,800 annotated videos and 58K QA pairs. The QA pairs are crowd-sourced by human annotators.

Metrics for open-ended QA. We consider two metrics: accuracy and score. An LLM generates a yes/no response by comparing the ground truth and predicted answers. The percentage of 'yes' responses is the accuracy metric. The LLM also provides a score between 1 to 5 for the comparisons. We report the average score.

LLM-based evaluation. Commonly, grounded videoQA approaches consider a closed-set setup where answer options are available during inference, and the goal is to choose the correct option [49, 54, 56]. To compare our open-ended approach with other closed-set methods, we need to choose an option among the available choices and calculate the metrics. For this, we use another pretrained LLM to select an option with the highest similarity to our prediction. Such LLM-assisted evaluation is commonly used in open-ended videoQA [5, 25, 28]. We use GPT-3.5-turbo [32] to be consistent with these. We also experiment with the openly available LLama 3.1 [40] for evaluation and present the results in the supplemental material.

4.2. Comparison with state of the art

Grounded videoQA. We compare TOGA with the state of the art on NExT-GQA [49] and show the result in Tab. 1. TOGA outperforms the existing approaches on grounding and on generating correct answers. LLoVi [56] is a zero-shot approach, while other methods are weakly supervised. It should be noted that our open-ended setup is more challenging than the closed-set setup followed in the existing approaches since the model is not able to view the options while generating the answer. Additionally, other approaches generate the grounding output separately - either using a post-hoc approach, or using separate modules for answering and grounding. However, we generate the answer and the grounding jointly. We believe that capturing the multi-scale temporal features and instruction-tuning the model to jointly predict answers with the grounding helps us achieve this performance.

	Model	mIoU	R@1 (IoU=0.3)	R@1 (IoU=0.5)
Non Generative	UniVTG	28.17	41.34	26.88
	CG-DETR	23.87	31.31	16.67
LLM based	VTimeLLM	20.14	28.84	17.41
	TimeChat	11.65	14.42	7.61
	LITA	21.49	29.49	16.29
	Ours	**25.53**	**29.91**	**19.79**

Table 2. We compare our generative LLM-based grounding method against other state-of-the-art methods on the ReXTime [3] zero-shot grounding task. The best generative scores are shown in bold, and the best non-generative model scores are underlined.

Method	MSVD-QA		ActivityNet-QA	
	Accuracy	Score	Accuracy	Score
FrozenBiLM [52] *NeurIPS '22*	32.2	-	24.7	-
VideoChat [23] *(arXiv '23)*	56.3	2.8	-	2.2
LLaMA-Adapter [58] *(ICLR '24)*	54.9	3.1	34.2	2.7
Video-LLaMA [57] *(EMNLP '23)*	51.6	2.5	12.4	1.1
Video-ChatGPT [28] *(ACL '24)*	64.9	3.3	35.2	2.7
Chat-UniVi [17] *(CVPR '24)*	65	3.6	45.8	3.2
Video-LLaVA [25] *(EMNLP '24)*	70.7	3.9	45.3	3.3
Video-LLaMA2 [5] *(arXiv '24)*	70.9	3.8	50.2	3.3
Ours	**73.8**	**3.9**	**52.0**	**3.4**

Table 3. Comparison with the state of the art for open-ended videoQA on MSVD-QA and ActivityNet-QA. We outperform existing approaches against both metrics.

Zero-shot grounding. We evaluate our approach on the ReXTime benchmark [3] on the task of zero-shot query grounding. We compare our method with other generative LLM-based methods in Tab. 2. We also include non-generative grounding models for reference, but it is important to note that these non-generative methods can not answer open-ended queries like the generative methods.

Open ended videoQA. We consider MSVD-QA and ActivityNet-QA for this task. As shown in Tab. 3, TOGA outperforms the state of the art on both datasets. We believe that capturing the multi-scale features with the MS-VLC enables TOGA to accurately answer questions.

Qualitative results. We include some qualitative examples from the NExT-GQA dataset in Fig. 4. We can observe that due to the open-ended nature of our setup and because of the fact that the model is not seeing the options to the question, the model can generate responses similar in mean-

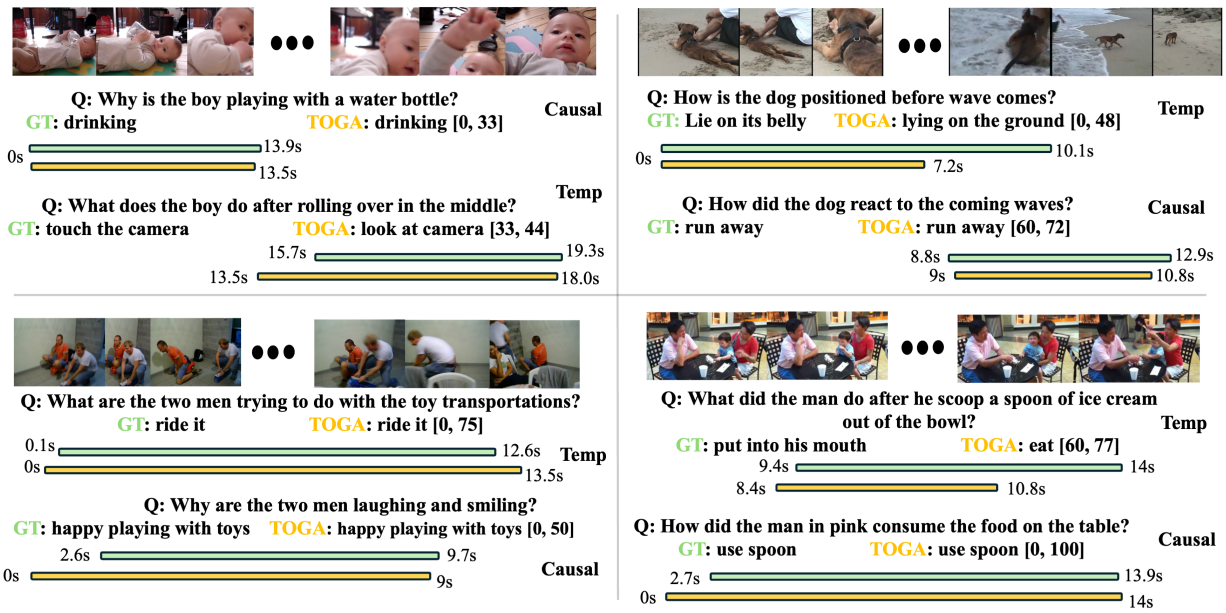

Figure 4. Qualitative results on NExT-GQA. We present the ground truth and predicted answers with groundings. Ground truth segments are marked in green, and predicted segments are marked in yellow. A causal and a temporal question are selected from a set of questions for each video. TOGA mostly generates correct answers and grounding. In some cases, predictions are different from the ground truth (GT) answers. This can be attributed to the open-ended nature of our approach.

ing to, but not matching, the ground truth exactly. Along with generating the answer, our method is able to simultaneously ground the answer. Each video in NExT-GQA has multiple questions, and we show examples of one temporal and one causal query for each video shown here. We also include qualitative examples taken from the MSVD-QA dataset in Fig. 5. Here as well, the model may generate answers that do not match the ground truth exactly but might convey the same meaning.

4.3. Ablation and analysis

We perform ablations on major components of our approach and analyze the results on various experimental setups.

Ablation on multi-scale vision-language connector. Our MS-VLC module processes videos at two resolutions (16 dense, 4 sparse frames). To demonstrate its effectiveness, we evaluated frameworks with single temporal resolutions on NExT-GQA [49]. We categorize answers by true grounding length (short: $< 30\%$ video, medium: $30 - 70\%$, long: $> 70\%$) and report IoU in Tab. 4. The multi-scale model consistently outperforms single-scale variants, with a more pronounced advantage for short and long-duration events.

Ablation on the consistency constraint. For weakly supervised grounding, at the final stage of training, we impose a consistency constraint between the grounding response and the response generated by a question with the same grounding. To justify this, we train a variant without this stage of training where we instruction tune the model only with pseudo-temporal groundings. This results in an mIoU of 12.1, which is significantly lower compared to the mIoU of 24.4 on NExT-GQA with the final stage of training.

Model type	Query Type			
	All	short	medium	long
Sparse only	20.0	16.2	28.9	47.5
Dense Only	22.1	18.3	32.2	32.1
Multi-Scale (MS-VLC)	**24.4**	**20.5**	**34.7**	**49.3**

Table 4. The MS-VLC improves grounding performance across a range of temporal durations of queries.

Analysis of the type of questions. NExT-GQA consists of two primary types of questions: causal and temporal queries. Causal questions include why and how questions. Temporal questions are divided into past, present, and future types based on the timing of the answer. For example, `What did the boy do after standing up` is an example of a future temporal question. Both of these question types involve different levels of reasoning. We present the results corresponding to the Acc@GQA metric in Tab. 5. We notice that the temporal questions are more difficult than causal questions, as they require understanding the sequence of events. Temporal questions, especially those referring to the past or future, e.g. `What did the boy do` *`after`* `he stood up from the ground?` require more long-term reasoning to generate correct answers with grounding.

Analysis of the number of frames. We explore the ef-

Question types	Causal		Temporal		
	Why	How	Present	Past	Future
Acc@GQA	26.1	27.4	23.4	18	18.1

Table 5. Acc@GQA for causal and temporal question types on NeXT-GQA. We observe that temporal questions are more difficult, specifically questions referring to past and future events.

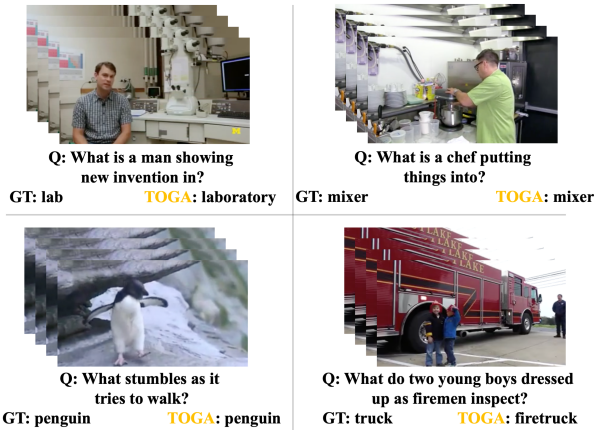

Figure 5. Qualitative results for open-ended videoQA on MSVD. TOGA can generate correct but slightly different answers from the ground truth due to its open-ended nature.

fect of the frame count on the grounding performance on NExT-GQA. We choose a setup (16, 4), i.e., 16 frames for the dense and 4 for the sparse scale. We experiment with other setups with lower frames (8, 2) and higher frames (32, 8). The (8, 2) setup achieves a mIoU of 20.8. The (32, 8) setup achieves a better mIoU of 21.5 with a three times higher training time. Both are lower than the mIoU of 24.4 with (16, 4). We believe more frames better encode the video features, but may introduce spurious features and make it harder to train the model. More combinations, along with an interleaved frame version, are in the appendix.

Analysis of temporal representation. We consider a temporal representation of [<start>, <end>] where start and end are integers $\in [0, 100]$, with 0 marking the start of the video and 100 marking the end. We chose this range based on the length of the videos in seconds. We scale the $[0, 100]$ range to video start and end times in seconds. We also considered a scenario where the start and end are floats $\in [0, 1]$. We achieve a lower mIoU of 19.0 on NeXT-GQA with the [0, 1] range, compared to 24.4 with [0,100]. We notice it is harder to learn the floating point representation with language tokens. A similar behavior is shown in [29].

Failure cases. We present the failure cases on grounding and QA tasks in Fig. 6. On top, we present a case from NExT-GQA where TOGA 's response is somewhat relevant but does not match the ground truth. This is due to the open-ended answer generation. In another instance, we generate a closely matched answer `smile` vs. `laugh`, but the grounding is inaccurate. This is due to the weakly supervised learning, where we train with noisy grounding labels. At the bottom, we present two examples from MSVD. TOGA generates a suitable answer of `towel` in response to `what does a person hold up`, which is different from the ground truth. In the other case, the response `interviewer` is a special type of `person`.

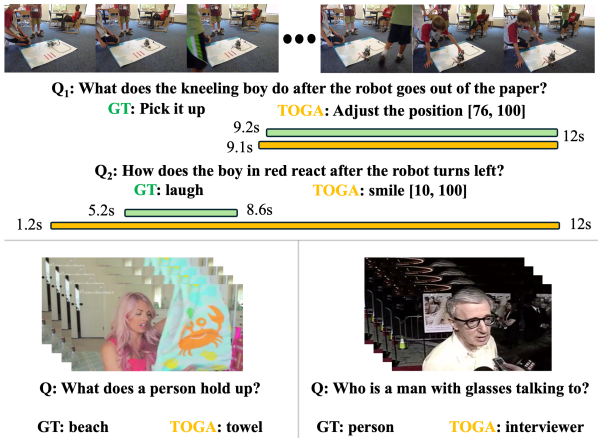

Figure 6. The top part presents an example from NExT-GQA where TOGA fails to ground the answer. The bottom part presents two examples from MSVD. TOGA generates answers that may be relevant but do not exactly match the GT.

Limitations. We generate pseudo labels to train TOGA to perform grounding in a weakly supervised setup. We consider one grounding interval for each answer. Thus, we cannot handle cases where the evidence for an answer is distributed across multiple intervals. We also answer all questions for a video independently. However, capturing the temporal dependencies between the answers could help generate more accurate answers with groundings. This is particularly helpful for the temporal questions with before/after clauses on NExT-GQA.

5. Conclusion

We have presented TOGA , a vision-language model for open-ended video QA with temporal grounding. We operate in a weakly supervised setup and do not rely on temporal annotations. TOGA consists of an MS-VLC module to capture both high-frequency and low-frequency temporal features. We have instruction tuned the MS-VLC and language decoder to jointly generate open-ended answers with temporal grounding. Unlike existing approaches, we do not require the options to generate an answer. Our experiments show that jointly generating grounded answers improves the accuracy of both answers and temporal grounding. We have evaluated TOGA on NExT-GQA for grounded QA and MSVD-QA and ActivityNet-QA for open-ended QA. We achieve state-of-the-art performance on these benchmarks.

Acknowledgment

This work was supported in part by the U.S. Air Force and DARPA under Contract No. FA8750-23-C-0519, the U.S. Army Research Laboratory Cooperative Research Agreement W911NF-17-2-0196, the U.S. Army Combat Capabilities Development Command (DEVCOM) Army Research Laboratory under Support Agreement No. USMA 21050, and the DARPA under Support Agreement No. USMA 23004. The opinions, findings, and conclusions expressed in this paper are those of the authors and do not reflect the position of the United States Department of Defense or the United States Government.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015. 2

[2] David Chen and William Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 190–200, Portland, Oregon, USA, 2011. Association for Computational Linguistics. 5

[3] Jr-Jen Chen, Yu-Chien Liao, Hsi-Che Lin, Yu-Chu Yu, Yen-Chun Chen, and Yu-Chiang Frank Wang. Rextime: A benchmark suite for reasoning-across-time in videos. *arXiv preprint arXiv:2406.19392*, 2024. 5, 6

[4] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, and Sergey Tulyakov. Panda-70m: Captioning 70m videos with multiple cross-modality teachers, 2024. 6

[5] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. 2, 3, 4, 5, 6

[6] Bernard Ghanem Fabian Caba Heilbron, Victor Escorcia and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970, 2015. 6

[7] Zhiyuan Fang, Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. Video2commonsense: Generating commonsense descriptions to enrich video captioning. *arXiv preprint arXiv:2003.05162*, 2020. 2

[8] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 2, 4

[9] Noa Garcia, Mayu Otani, Chenhui Chu, and Yuta Nakashima. Knowit vqa: Answering knowledge-based questions about videos. In *Proceedings of the AAAI conference on artificial intelligence*, pages 10826–10834, 2020. 2

[10] Yash Goyal, Tejas Khot, Aishwarya Agrawal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. *Int. J. Comput. Vision*, 127(4): 398–414, 2019. 2

[11] Michael Heilman and Noah A Smith. Question generation via overgenerating transformations and ranking. *DTIC Document*, 2009. 5

[12] De-An Huang, Shijia Liao, Subhashree Radhakrishnan, Hongxu Yin, Pavlo Molchanov, Zhiding Yu, and Jan Kautz. Lita: Language instructed temporal-localization assistant. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXIV*, page 202–218, Berlin, Heidelberg, 2024. Springer-Verlag. 2, 3

[13] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 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, 2017. 5

[14] Yunseok Jang, Yale Song, Chris Dongjoo Kim, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Video question answering with spatio-temporal reasoning. *IJCV*, 2019. 2

[15] Yunseok Jang, Yale Song, Chris Dongjoo Kim, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Video question answering with spatio-temporal reasoning. *Int. J. Comput. Vision*, 127 (10):1385–1412, 2019. 2

[16] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. 2, 4, 5

[17] Peng Jin, Ryuichi Takanobu, Caiwan Zhang, Xiaochun Cao, and Li Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding. *arXiv preprint arXiv:2311.08046*, 2023. 6

[18] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Slow-fast auditory streams for audio recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 855–859. IEEE, 2021. 2, 4

[19] Dohwan Ko, Ji Soo Lee, Miso Choi, Jaewon Chu, Jihwan Park, and Hyunwoo J Kim. Open-vocabulary video question answering: A new benchmark for evaluating the generalizability of video question answering models. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2023. 2

[20] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos, 2017. 6

[21] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L. Berg. Tvqa: Localized, compositional video question answering, 2019. 2, 3

[22] Jie Lei, Licheng Yu, Tamara L. Berg, and Mohit Bansal. Tvqa+: Spatio-temporal grounding for video question answering, 2020. 2, 3

[23] Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023. 2, 6

[24] Yicong Li, Xiang Wang, Junbin Xiao, and Tat-Seng Chua. Equivariant and invariant grounding for video question answering. In *Proceedings of the 30th ACM International Conference on Multimedia*, page 4714–4722, New York, NY, USA, 2022. Association for Computing Machinery. 6

[25] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023. 2, 3, 5, 6

[26] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 3, 4

[27] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[28] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models, 2024. 2, 5, 6

[29] Boris Meinardus, Anil Batra, Anna Rohrbach, and Marcus Rohrbach. The surprising effectiveness of multimodal large language models for video moment retrieval, 2024. 8

[30] Nicolo Micheletti, Samuel Belkadi, Lifeng Han, and Goran Nenadic. Exploration of masked and causal language modelling for text generation, 2024. 4

[31] George A. Miller. Wordnet: a lexical database for english. *Commun. ACM*, 38(11):39–41, 1995. 2

[32] OpenAI. Gpt-4 technical report, 2024. 2, 4, 6

[33] Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Shuangrui Ding, Dahua Lin, and Jiaqi Wang. Streaming long video understanding with large language models. *Advances in Neural Information Processing Systems*, 37:119336–119360, 2025. 3, 6

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 2, 3, 5, 6

[35] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollar. Designing network design spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 4

[36] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: an open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, page 4444–4451. AAAI Press, 2017. 2

[37] Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11888–11898, 2023. 2

[38] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4631–4640, 2016. 2

[39] Gemini Team. Gemini: A family of highly capable multimodal models, 2024. 2

[40] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. 2, 4, 6

[41] Haibo Wang, Zhiyang Xu, Yu Cheng, Shizhe Diao, Yufan Zhou, Yixin Cao, Qifan Wang, Weifeng Ge, and Lifu Huang. Grounded-videollm: Sharpening fine-grained temporal grounding in video large language models, 2024. 2, 3, 6

[42] Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. *arXiv preprint arXiv:2403.10517*, 2024. 2

[43] Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. Videotree: Adaptive tree-based video representation for llm reasoning on long videos. *arXiv preprint arXiv:2405.19209*, 2024. 2

[44] Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. Star: A benchmark for situated reasoning in real-world videos. *arXiv preprint arXiv:2405.09711*, 2024. 5

[45] Yongliang Wu, Xinting Hu, Yuyang Sun, Yizhou Zhou, Wenbo Zhu, Fengyun Rao, Bernt Schiele, and Xu Yang. Number it: Temporal grounding videos like flipping manga, 2024. 3

[46] Fanyi Xiao, Yong Jae Lee, Kristen Grauman, Jitendra Malik, and Christoph Feichtenhofer. Audiovisual slowfast networks for video recognition. *arXiv preprint arXiv:2001.08740*, 2020. 2, 4

[47] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa:next phase of question-answering to explaining temporal actions, 2021. 2

[48] Junbin Xiao, Nanxin Huang, Hangyu Qin, Dongyang Li, Yicong Li, Fengbin Zhu, Zhulin Tao, Jianxing Yu, Liang Lin, Tat-Seng Chua, and Angela Yao. Videoqa in the era of llms: An empirical study, 2024. 2, 3

[49] Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. Can i trust your answer? visually grounded video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13204–13214, 2024. 1, 3, 5, 6, 7

[50] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *ACM Multimedia*, 2017. 5

[51] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653, 2017. 5

[52] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Zero-shot video question answering via

frozen bidirectional language models. In *NeurIPS*, 2022. 2, 4, 6

[53] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Tubedetr: Spatio-temporal video grounding with transformers. In *CVPR*, 2022. 3

[54] Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model for video localization and question answering. In *NeurIPS*, 2023. 1, 2, 3, 4, 6

[55] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *AAAI*, pages 9127–9134, 2019. 5, 6

[56] Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. A simple llm framework for long-range video question-answering, 2023. 1, 2, 3, 4, 6

[57] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. 2, 6

[58] Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023. 6

[59] Yaoyao Zhong, Junbin Xiao, Wei Ji, Yicong Li, Weihong Deng, and Tat-Seng Chua. Video question answering: Datasets, algorithms and challenges, 2022. 2