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

Supplementary Material

In the supplemental material, we provide additional implementation details, experimental results with ablation studies, and qualitative results for video question answering and grounding.

1. Implementation Details

1.1. Training Details

The additional settings we use, including hyperparameters and implementation details, are shown in Tab. 1. We train on 8x A100 GPUs, and take 7 hours each for the instruction tuning stages and 54 hours for the pretraining stage.

1.2. Multi-scale vision language connector

Our MS-VLC module consists of two RegNet [36] stages along with a 3D convolution, as done in [5]. RegNet, or regular networks, are convolutional network architectures drawn from the regular design space identified by [36]. We take the architecture from [36], randomly initialize the weights, and train the RegNet and the 3D convolution from scratch over the training stages.

2. Additional experimental analysis

2.1. Multi-scale vision-language connector

In this section, we conduct ablations on the MS-VLC and analyze the performance under various conditions. Specifically, we evaluate the sparse-only and dense-only models on different setups and report the effect of the multi-scale connector on performance in these tasks.

Different question types Recall that the NExT-GQA [50] dataset consists of different question types, which are broadly classified into causal and temporal. In this section, we evaluate the performance of the sparse-only and dense-only connectors on the different question types. We observe that MS-VLC performs the best across all types of questions in the dataset, as shown in Tab. 2.

Different downsampling rate We sample 16 frames at the dense scale, and 4 frames at the sparse scale - a (16, 4) configuration. Thus, we use 4X downsampling from dense to sparse. In this section, we try a 2X downsampling rate - meaning we try a (16, 8) configuration. We observe that (16, 8) configuration achieves an mIoU of 21.7 on NExT-GQA, as compared to the mIoU of 24.4 achieved by the (16, 4) configuration. We think that 2X downsampling is not sufficient for the sparse branch, making it difficult to model long-term temporal relations due to the large amount of information even at the sparse scale.

Config	Stage		
	Pretraining	Instruction Tuning	Consistemcy
Vision encoder	clip-vit-large-patch14-336		
Vision select layer	-2		
Language Decoder	Mistral-7B-Instruct-v0.2		
Optimizer	AdamW		
Weight Decay	0		
Deepspeed	Zero3		
Epochs	1		
Warmup Ratio	0.03		
LR scheduler	cosine		
Decoder max length	2048		
Starting LR	1e-3	2e-5	
Batch size	256	128	

Table 1. Additional hyperparameter settings and other implementation details are used in our framework.

Connector type	Question type					All
	Causal		Temporal			
	Why	How	Present	Past	Future	
Sparse only	21.7	21.1	18	17.9	12.8	19.55
Dense only	21.8	22.2	18.8	18	13.6	20.12
MS-VLC	26.1	27.4	23.4	18	18.1	24.6

Table 2. Ablation on the MS-VLC on different question types in the NExT-GQA dataset. We report Acc@GQA for the specific question type. Recall that Acc@GQA considers the accuracy of both the answer and the temporal grounding. 'All' considers all the questions for calculating the metric.

	GT	Predictions
Mean centre position	49.9	50.1
Average length	21.6	21.5
% of timestamps starting at 0	14.40%	21.20%
% of timestamps ending at 100	10.10%	14.05%

Table 3. Statistics of the predicted and the ground truth (GT) timestamps. All timestamps are normalized in the range [0, 100] for this analysis. We notice that while the mean center position and average length are similar for GT and predictions, the predictions are more biased to start at 0 and end at 100 than the GT.

2.2. Novel open-Acc@GQA metric for NExT-GQA.

Recall that to compute the Acc@GQA metric for NExT-GQA, the answer and the grounding both need to be evaluated simultaneously. The grounding is deemed 'correct' if the IoP > 0.5. The answer is 'correct' if the correct option is chosen among the five pre-determined options for the question. An answer+grounding pair is correct for the Acc@GQA metric if both conditions are satisfied.

Error type		MS-VLC	Dense Only	Sparse only
Prediction mismatch %	Early start %	24.90%	29.48%	35.28%
	Late start %	26.80%	22.69%	25.10%
	Early end %	30.14%	37.08%	38.56%
	Late end %	28.60%	24.08%	28.74%
Mean Absolute Error	MAE in start time	19.0	19.1	22.2
	MAE in end time	21.3	22.4	25.6
	MAE in centre	18.6	19.2	22.3

Table 4. Error analysis of the predictions compared to the ground truth (GT). The top half of the table is the percentage of predictions starting/ending earlier/later than the GT. The margin is set to be 10% of the video length to be classified as early or late for this analysis. The bottom half of this table reports the Mean Absolute Error (MAE) in the start/end/center times. All error analysis is performed for the three different variations of the vision-language connector.

Method	MSVD-QA	ActivityNet-QA
Video-LLaMA2[5]	82	80
Video-LLaVA[26]	81	74
Ours	**88**	**83**

Table 5. Comparison with other methods on the open-ended QA task, evaluated using Llama 3.1. The metric reported here is the QA accuracy as judged by Llama. We observe that Llama is more lenient than GPT in terms of evaluating responses. However, the trend is similar and our approach outperforms other methods on the open-ended video QA task.

However, our approach generated free-form answers in an open-ended setup. To compute the Acc@GQA metric and compare our method to other methods, we have to 'choose' an option after generating open-ended answers. Recall that we perform GPT-assisted retrieval for this, retrieving the most similar option to the open-ended answer generated by our method.

In this section, we propose a new metric more suitable for evaluating the open-ended nature of our approach, called *Open-Acc@GQA*. This is based on the GPT-assisted open-ended QA evaluation performed in other works [5, 26, 29]. Specifically, we query GPT to compare our response and the ground truth answer, given the question. We ask GPT to come up with a 'yes' or 'no' response to whether the prediction is similar in meaning to the ground truth. Note that as opposed to the GPT-assisted retrieval performed earlier, the options are not passed to GPT.

We impose similar conditions on the Open-Acc@GQA metric, as imposed by Acc@GQA. Specifically, the predictions are to be similar in meaning to the ground truth (as described earlier), and the IoP of the grounding must be > 0.5. In this way, Open-Acc@GQA becomes more suited to evaluate the open-ended grounded video QA task. We achieve **21.4%** on the Open-Acc@GQA metric. Note that we can not directly compare this to existing approaches for grounded video QA since they do not generate open-ended answers. It is interesting to compare this number to the Acc@GQA metric, on which the same model achieves 24.6%. Open-Acc@GQA is a more difficult metric than Acc@GQA due to the lack of options in the former.

2.3. Prediction statistics and error analysis

We report statistics of the predicted and ground truth timestamps in Tab. 3. We notice that the average length and mean center position are similar for both, but predictions are slightly biased toward the early start and late end.

We also perform some additional error analysis on the predicted timestamps to get insights into where the model is going wrong. We compute the percentage of times the model makes an early/late prediction for start/end times. We also compute the mean absolute error (MAE) in the start, end, and center of the predicted timestamp compared to the GT. This analysis is presented in Tab. 4. These observations entail that errors are mostly uniform in either direction for both start and end times.

2.4. Open-ended Llama-based evaluation

We have evaluated the open-ended video QA task using GPT-assisted evaluation, to be consistent with previous work [5, 26, 29]. However, GPT is a closed-source and is expensive to use. Thus, to improve accessibility and reproducibility, we also evaluate our approach using the open-source pretrained LLM Llama 3.1. Similar to the GPT evaluation, we pass the question, prediction, and ground truth to Llama. We ask it to come up with a 'yes' or 'no' response to whether the prediction is similar to the ground truth answer. The percentage of 'yes' responses is the accuracy of the model. To compare with other approaches for open-ended video QA, we also reproduce other approaches and evaluate them using the same Llama-based evaluation.

As shown in Tab. 5, Llama is generally more lenient in evaluation compared to GPT, since the numbers for all methods are higher than the corresponding GPT-assisted evaluation. Nonetheless, we observe a similar trend as the GPT evaluation, and our approach outperforms others in this evaluation.

2.5. Effect of chat template

Recall that we apply a chat template to every question before passing it to the model. In the template, we include the desired output format, such as `answer` or `answer [<start>, <end>]`. While evaluating on NExT-GQA, we generate the answers and the grounding jointly, so we include the latter format in the queries in NExT-GQA. In this section, we try an experiment where we do not apply any template to the question. We observe that the model fails to predict the grounding for any query in the NExT-GQA test set. Thus, we conclude that the template is essential for the model to generate temporal grounding along with the answers.

2.6. Acc@QA metric

The Acc@GQA metric evaluates the grounding and QA performance simultaneously. However, another metric Acc@QA is also used in [50] to evaluate just the QA ability separately. For this, we discard the temporal groundings in the generated answer and just evaluate the answers using our GPT-assisted retrieval method. We achieve an accuracy of 67.0%, which is close to the SOTA [57], which achieves 67.7%. Note that our model does not consider options while generating the answer, as opposed to [57].

2.7. Frame numbers in prompt

Taking inspiration from [12, 30], we try to input the frame indices between the visual features in the prompt, to see if it further improves the grounding performance. However, we observe that this decreases performance - achieving a mIoU/mIoP of 21.2/36.4, respectively, as opposed to 24.4/40.5 of our best model. We observe that these approaches work with image-level features and interleave the frame indices between features of individual frames. However, our multi-scale connector comprises 3D convolutions that learn the correlations between frame features within the connector itself. Hence, explicit frame indices are less important in the prompt, since the temporal information is already present in 3D convolutional features.

2.8. Effect of language decoder

We have used the Mistral-7B [16] language decoder in our experiments. We also experimented with another popular decoder, namely Vicuna-7B. We find that using the vicuna-7B language decoder in our approach achieves a mIoU/mIoP of 19.0/33.9, respectively, which is lower than our main model (24.4/40.5).

2.9. Traditional metrics for QA evaluation

Most of the Video QA works use GPT for evaluation [5, 26]. However, it may be seen as inconsistent over time due to API version changes in the closed-source GPT model. Hence, we also evaluate using open-source LLaMA in Sec. 2.4. Further, we utilize the traditional METEOR metric [21] for a more robust evaluation of our model. We achieve a METEOR score of 0.498 on the ActivityNet-QA dataset.

2.10. Evaluation on additional datasets

We evaluate our approach on the ActivityNet-RTL test dataset [12]. ActivityNet-RTL is another popular temporal reasoning and localization dataset involving temporal grounding. It contains QA pairs which require reasoning over complex events and actions. We perform the evaluation in a zero-shot setting - we do not train on ActivityNet-RTL. We obtain 22.9 mIOU and 18.2 P@0.5 on this dataset. We compare it to LITA 7B model's performance of 24.1 mIOU and 21.2 P@0.5, but also note that LITA has been trained on this dataset.

We additionally report results on the ReXTime [3] validation set in a zero-shot setting, without finetuning on ReXTime videos too. We get a mIoU of 27.4 and mIoP of 41.9 on the validation set of ReXTime.

3. Qualitative examples

We include additional qualitative examples for the NExT-GQA dataset in Fig. 1 and the ActivityNet dataset in Fig. 2. We include three questions for each video in Fig. 1. We can observe that the model can detect events that happen in a relatively short temporal window compared to the length of the video. We can also observe that the predicted answers may not exactly match the ground truth in both of these figures. This is due to the open-ended nature of our approach.

Even though ActivityNet is not a grounding dataset, we still try to ground the questions in the dataset to observe the performance. We show examples in Fig. 3. TOGA generates some groundings, but we can not evaluate the quality without the ground truth annotations. Manually looking at these samples, the predictions seem reasonable based on the correlation between answers and the frames.

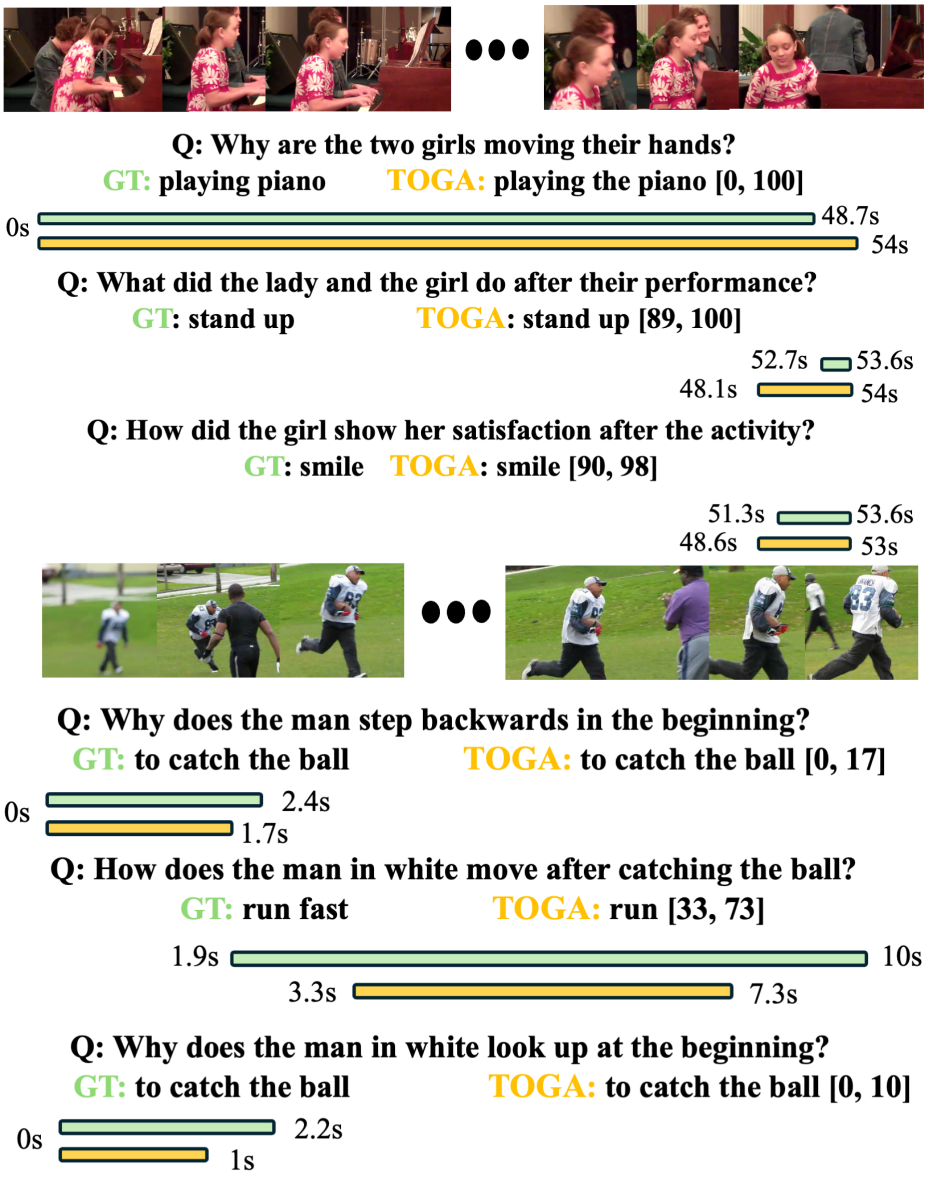
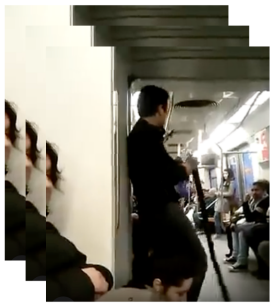


Figure 1. Additional qualitative examples from the NExT-GQA dataset, with longer videos and multiple questions per video. Compared to the length of the video, the grounding evidence for the answer may be small. However, our model can capture events in small temporal windows as well.



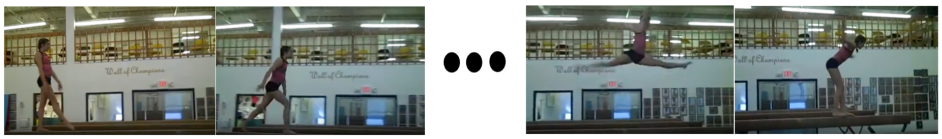
Q: What is the man standing in the video?
GT: subway **TOGA: subway**

Q: Where is the person in the blue coat?
GT: road **TOGA: street**

Q: What is the guy in blue swimming trunks doing?
GT: swimming **TOGA: diving**

Q: Where is the place in the video?
GT: snow land **TOGA: ski field**

Figure 2. Additional qualitative examples from the ActivityNet dataset. Our model may generate words that are similar in meaning but not the same as the ground truth, due to the open-ended nature of our approach.



Q: What is the athlete doing in the video?
GT: exercise gymnastics **TOGA: performing gymnastics [21, 100]**

2.0s 9.5s

Q: What is the name of the ball in the video?
GT: volleyball **TOGA: volleyball [1, 100]**

0s 59s

Figure 3. Examples of grounding on ActivityNet videos. TOGA can perform grounding even on this dataset, but the questions may be relevant to the whole video rather than a part of it.