Supplementary Material of *TimeCraft*: Navigate Weakly-Supervised Temporal Grounded Video Question Answering via Bi-directional Reasoning

Huabin Liu[1], Xiao Ma[2], Cheng Zhong[2], Yang Zhang[2], and Weiyao Lin[1]

[1] Shanghai Jiao Tong University, Shanghai, China
{huabinliu, wylin}@sjtu.edu.cn
[2] AI Lab, Lenovo Research, Beijing, China,
{maxiao3, zhongcheng3, zhangyang20}@lenovo.com

1 More details about dual Q&A generation

1.1 Prompt engineering

To streamline the process of deriving dual Q&A from the original Q&A pair, we utilize LLMs like Llama2 and Gemini to generate prompts. Initially, our requests for prompts, such as "*Task: given the following original Q&A and a dual Q&A, design a prompt for me to send to LLMs to complete the task of transferring the original one to the dual one. [Original Q&A: ..., Dual Q&A: ...]*', enable the LLMs to provide some reasonable but basic prompts, which are able to prompt LLMs to complete this dual Q&A generation task.

Subsequently, we manually refine these prompts to enhance the quality of the generated dual Q&A through a human-in-the-loop methodology, incorporating feedback from LLM outputs. Initially, we noted instances where LLMs produced incomplete pairs containing only the question text. To address this, we augmented our requirements to ensure that outputs consistently feature both

Table 1: Examples of enhanced prompts for generating dual Q&A.

Index	Prompt Description
1	Task: Create a question that mirrors the given answer in causal or temporal structure to generate a dual Q&A pair.
2	Task: Create a dual Q&A pair, which revser the given Q&A in causal or temporal structure.
3	Task: Create a question that mirrors the given answer in causal or temporal structure to generate a dual Q&A pair. Ensure the generated Q&A consists of both question and answer.
4	Task: Create a question that mirrors the given answer in causal or temporal structure to generate a dual Q&A pair. Ensure the generated Q&A consists of both question and answer. For some Q&A pairs that you cannot well transfer to a complete dual Q&A, just maintain the original Q&A.
5	Task: Create a question that mirrors the given answer in causal or temporal structure to generate a dual Q&A pair. Ensure the generated Q&A consists of both question and answer. For some Q&A pairs that you cannot well transfer to a complete dual Q&A, just maintain the original Q&A. Here are some examples: <Example-1> # Given Q&A: Why did the boy touch the front wheel after looking up? Prevent himself from falling. # Generated Q&A: How did the boy prevent himself from falling? Touch the front wheel after looking up. <Example-2> # Given Q&A: What did the man do before he put the baby on the floor? Carry baby and chase dog. # Generated Q&A: What did the man do after he carried the baby and chased the dog? Put the baby on the floor.

Fig. 1: Results of dual Q&A generation for different types of question using Llama2.

question and answer texts. Moreover, we encountered challenges with descriptive Q&A pairs, where generating a meaningful dual question proved difficult. For instance, in cases like "`How many people are in this video? Three.`", generating a dual question is less practical. To tackle such scenarios, we instruct LLMs to bypass these descriptive Q&A pairs by echoing the original text as the dual counterpart. Crucially, we have integrated the widely adopted few-shot prompting strategy, supplementing prompts with a few real examples to elucidate the task further. This strategy has proven effective in enhancing the outputs of LLMs, serving as a vital component in improving the overall quality of the generated dual Q&A pairs. Some potential prompt candidates are shown in Tab. 1.

1.2 More results of generated dual Q&A

To underscore the effectiveness of employing LLMs for the dual Q&A generation task, we present additional output results produced by Llama2 in Fig. 1. These results demonstrate the capability of advanced LLMs in effectively executing this task.

Table 2: The results under different values of hyper-parameters in the training loss.

Setting (α,β)	(1,1)	(0.5,1)	(1,0.5)	(0.2,1)	(1,0.2)	(1,0)	(0,1)
Acc@GQA	18.2	**18.5**	17.7	17.0	17.2	16.7	15.6
mIoP	27.6	**28.1**	27.2	26.8	27.4	26.8	25.5
mIoU	**15.8**	15.6	15.5	13.5	12.9	12.1	8.4

Table 3: The results using the different number of examples that showing to LLMs as part of the prompt (denoted by the number of shot).

Prompt Setting	zero-shot	1-shot	3-shot	5-shot	7-shot	10-shot
Acc@GQA	17.1	17.6	18.5	18.5	**18.7**	18.6
mIoP	26.8	27.7	28.1	28.0	28.1	**28.4**
mIoU	14.7	15.0	15.6	**15.8**	15.7	15.8

2 Additional ablation studies

2.1 Study on hyper-parameters of loss

As we have described in our main manuscript, the overall loss function of our framework is:

$$\mathcal{L} = \mathcal{L}_{QA} + \alpha\mathcal{L}_{recon} + \beta\mathcal{L}_{ground}$$

In our default setting, we set the hyper-parameters of α and β as 1 and 0.5, respectively. Here, we study how the value of these hyper-parameters impacts the final GQA performance. Results are summarized in Tab. 2.

2.2 Impact of Few-shot prompting strategy

As discussed earlier, the few-shot prompting strategy plays a pivotal role in prompts. To this end, we conducted further analysis to examine how the number of examples presented to LLMs impacts the final results. Tab. 3 summarizes the outcomes using varying numbers of examples, ranging from zero-shot to 10-shot. Notably, the importance of few-shot prompting becomes evident, as the zero-shot setting exhibits a rapid decrease in performance across all metrics.

However, we also observed that increasing the number of examples provided to LLMs does not always guarantee improved performance. Consequently, we have chosen the 3-shot configuration as our default setting to expedite the inference process of LLMs.

2.3 Potential of *TimeCraft* to ground both question and answer

As discussed in Section 4.3.2, various grounded results (question-critical, answer-critical, or their union) can be obtained during inference. To delve deeper into

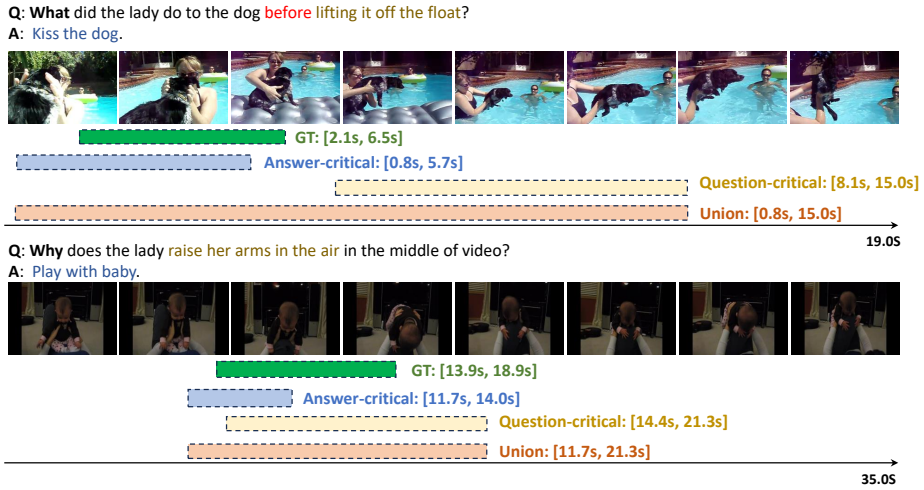

Fig. 2: Visualization results of (i) Answer-critical (ii) Question-critical (iii) Union grounded moment for the same video (Definition of these three outputs can be referred to Section 4.3.2 in our main manuscript.).

the different grounding styles, we visualize all the varied grounded results for the same video, as depicted in Fig. 5. In the context of causal QA, the predicted answer-critical and question-critical moments exhibit some overlap. However, in temporal QA, they may demonstrate distinct temporal distributions. This observation aligns with our expectations. The causal QA tasks in the NeXT-GQA dataset typically necessitate visual evidence relevant to information present in both question and answer texts. Furthermore, the actions described in the questions and answers frequently occur either simultaneously or in close succession. Conversely, in temporal QA, greater emphasis is placed on the answer-critical moment for prediction, as actions described in the question and answer texts occur sequentially (either after or before one another). These findings underscore the potential of our proposed TimeCraft in providing grounded moments tailored to the differing preferences of question and answer information.

3 Visualization results

In Fig. 3, we provide more visualized results of our TimeCraft on the test set of the NeXT-GQA dataset. As we can see, in most cases, our TimeCraft can provide correct answers with precise and relevant temporal moments. Compared to Temp[CLIP], FrozenBiLM could provide the right answer even though it didn't ground the relevant key moment. This is attributed to its prior knowledge within the pre-trained language model, which further emphasizes the importance of achieving reliable QA by correctly grounding temporal locations.

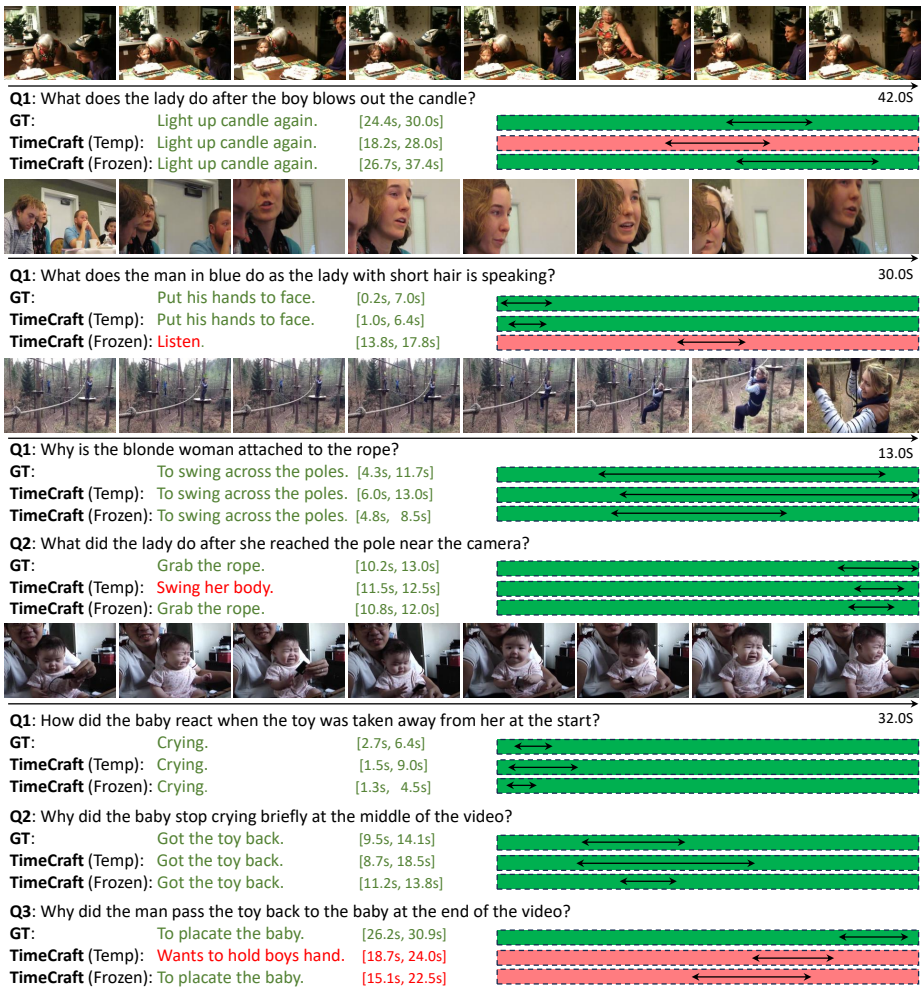

Fig. 3: Visualization results of grounded VQA in the test set of NeXT-GQA. Correct predictions are highlighted in green, while wrong predictions are in red.